35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Shift Invariant Feature Extraction for sEMG-Based Speech Recognition with Electrode Grid*

Takatomi Kubo[1], Masaki Yoshida[2], Takumu Hattori[1], and Kazushi Ikeda[1]

Abstract— **For Japanese vowel recognition based on surface electromyography (sEMG), an electrode grid has been shown to be effective in our previous studies. In this study, we aim to leverage potential of the electrode grid further by using with a spatial shift invariant feature extraction method that can compensate deviation of the attached site of the electrode grid. We verified efficiency of the shift invariant feature extraction method in improving the recognition accuracy. 2-D dual tree complex wavelet transform was employed as such a shift invariant feature extraction method. Our result shows that shift invariant feature can provide additional information that cannot be provided when the channel signals are utilized independently.**

I. INTRODUCTION

Speech is a unique, complex, and dynamic motor activity through which individuals express thoughts and emotions. It is one of the most powerful tools of the human species, and it contributes greatly to the quality of life. Dysarthria deprives people of such an invaluable tool. In order to support their communication, there are some researches where automatic speech recognition (ASR) was applied to dysarthric patients to estimate what they want to say, since ASR technology has advanced to the point of being utilized in our daily lives. However, users' speech impairments have caused low recognition accuracy. To overcome this problem, surface electromyography based ASR (sEMG-ASR) has been investigated as an augmentative or alternative information source [1], [2]. sEMG is a procedure that measures muscle electrical activity associated with muscle fiber contraction by using electrodes attached on the skin. Not only in cases when a user makes usual voiced speech, but also when voiceless mouthed speech is made, sEMG-ASR can support communication.

Over the last decade, there has been significant progress in the research on sEMG-ASR [1]–[14]. Previous studies have indicated the potential effectiveness of sEMG-ASR, not only for healthy people, but also for dysarthric patients. Deng et al. [1] proposed an ASR system based on sEMG, with and without acoustic signal, wherein they showed that a high word recognition accuracy (over 95%) could be achieved for dysarthric patients. Their result indicates that sEMG-ASR has the potential to be a novel type of speech prosthesis. To achieve a high recognition accuracy in sEMG-ASR, it is necessary to decide the appropriate location of the electrodes. However, in previous studies, disc electrodes or parallel bar electrodes were used and located empirically, based on anatomical knowledge. Because there exist relatively small muscles in proximity to each other in the face or neck region, it is difficult to avoid the influence of cross talks and innervation zones when conventional measurement methods are applied.

In order to improve signal-to-noise ratio of sEMG signals recorded from the lower facial muscles, Lapatki et al. [15]–[17] proposed an sEMG system using electrode grid. We also used an electrode grid which consists of densely-spaced multielectrodes in our previous experiments [18], [19] to avoid missing out information about speech in the measurement step. The sEMG signals were measured from the submental region with the electrode grid during the production of five vowel sounds. As a preliminary study, we conducted vowel recognition experiments by applying linear discriminant analysis (LDA) and hidden Markov model (HMM) to the data obtained as described above. We achieved approximately 80% to 85% recognition accuracies, which outperformed those of the results of virtually reconstructed single bipolar signals. In addition, by using sparse discriminant analysis (SDA) proposed by Clemmensen et al. [20], [21], we evaluated the redundancies in the features resulting from redundancy of the channels in the electrode grid [19].

In this study, we aim to leverage potential of the electrode grid further by using with a spatial shift invariant feature. The shift invariant feature extraction method can be expected to have a possibility to compensate deviation of the attached site of the electrode grid, because it can be regarded as relative spatial shift of signal sources to the sensors. Therefore, we verify the efficiency of the shift invariant feature extraction method in improving the recognition accuracy. Specifically, we employ the 2-D dual-tree complex wavelet transform (DT ℂWT) proposed by Kingsbury [22]–[25] as such a feature extraction method.

II. MATERIAL AND METHODS

A. sEMG System

For our experimental setup, we used an sEMG system developed by Hattori et al. with few modifications made on the electrode grid [18], [26]. The electrodes which consisted of silver bars in Hattori's study were substituted with spring connector pins, with each pin having a diameter of 0.8 mm, to absorb any vertical displacement of the attached site (Fig.

*This study was supported in part by a research grant from the Foundation for the Nara Institute of Science and Technology

[1] T. Kubo, T. Hattori, and K. Ikeda are with the Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan takatomi-k at is.naist.jp

[2] M. Yoshida is with the Faculty of Biomedical Engineering, Osaka Electro-Communication University, 1130-70 Kiyotaki, Shijonawate, Osaka 575-0063, Japan yoshida at isc.osakac.ac.jp

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

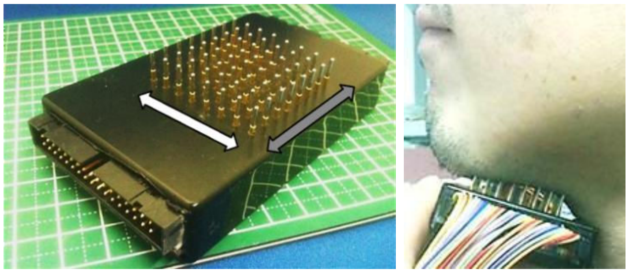

Fig. 1. (*left*) The electrode grid. The white double-headed arrow indicates the row direction, and the gray double-headed arrow indicates the column direction.
(*right*) The location of the electrode grid on the submental region in lateral view.

1). The set of electrodes were arranged in an array of 8 rows by 8 columns, with the interelectrode distance set to 5.08 mm, from center to center, in both directions. To reduce skin impedance, a voltage follower circuit was built with each electrode. The electric potential differences between each pair of electrodes neighboring in column direction were amplified up to 66 dB with band-pass filtering between 10 to 1500 Hz. Subsequently, the electric potential differences (56 channels) were digitized with a 16-bit analog-to-digital converter (National Instruments, NI USB-6255) and a laptop computer running MATLAB with its Data Acquisition Toolbox (MathWorks, 2010a). A microphone (KNOWLES, SP0103NC3-3) was also attached in front of the electrode grid.

B. Data Collection

We obtained data from six Japanese native speakers (two female and four male with mean age of 26.2 years), who had no known speech impairments. In each trial, the subjects were asked to produce each of the five Japanese vowels (/a/, /i/, /u/, /e/, and /o/) once in random order. The task vowels were presented on a screen for 1 second with an interval of 2 seconds between each of vowels, and the subjects were instructed to start vowel production at the onset of a visual presentation and stop at the offset. A total of 50 trials were conducted by each subject. During vowel production, the sEMG signals were recorded with the electrode grid attached on the submental region as shown in Fig. 1. The grid's centerline in the column direction and the last row were aligned with the center of the mandible and the posterior edge of the submental triangle, respectively. As preparation, the skin on the submental region was cleaned with an alcohol swab prior to attaching the electrode grid. Both the sEMG and acoustic signals were then captured and digitized at 16 kHz with an analog-to-digital converter. Written informed consents were obtained from the all subjects prior to the experiment. This study was approved by the institutional ethics committee.

C. Data Preprocessing

The sEMG signals were filtered with an 8th order low-pass Butterworth filter having a cut-off frequency of 500 Hz, and then downsampled to 2 kHz. The onsets and offsets of the acoustic signals were used as reference to determine those of the sEMG signals. The criteria applied in detecting the onsets and offsets of the acoustic signals were based on a set of amplitude thresholds. Also, to consider the delay between the sEMG signals and the acoustic signals [7], the onset of the sEMG signals were set to precede that of the acoustic signals by 150 msec. As for the offsets of sEMG signals, these were set to 150 msec after the offsets of the acoustic signals. These onsets and offsets of the sEMG signals were used to extract data for the following feature extraction process.

D. Feature Extraction

We employed 2-D dual-tree complex wavelet transform (DT $\mathbb{C}$WT) [22]–[25] in combination with cepstral coefficients as features for this study. For comparison, we prepared two other feature sets, (i) the cepstral coefficients calculated directly from each channel, and (ii) the cepstral coefficients calculated from both channel signals and wavelet coefficients.

1) Dual-Tree Complex Wavelet Transform: DT $\mathbb{C}$WT was proposed by Kingsbury [22]–[24] and it employs two real discrete wavelet transforms (DWT). One DWT gives the real part of the transform while the other gives the imaginary part. The two DWTs use different sets of filters that make an approximately analytic transform possible. DT $\mathbb{C}$WT has the following properties:

- Approximate shift invariance
- Good directional selectivity (for multidimensional signals)
- Perfect reconstruction
- Limited redundancy (independent of the number of scales, 2^m for m-dimensional signals)

Extention to 2-D is achieved by separable filtering along columns and then rows. The 2-D DT $\mathbb{C}$WT produces six bandpass subimages of complex coefficients at each level, which are oriented at angles of ±15°, ±45°, ±75°. For more details, see [22]–[25].

We used MATLAB codes that are available from [27] to implement the 2-D DT $\mathbb{C}$WT. After interpolation from 7 by 8 data array of each time point to 24 by 28, we applied the 2-D DT $\mathbb{C}$WT to it. The decomposition level was set to two for preventing increase of computational time, although higher decomposition level should be preferable for utilizing spatial information sufficiently.

2) Cepstral Coefficients: In our preliminary study, the cepstral coefficients indicated higher recognition accuracies than time domain features [18]. Therefore, the cepstral coefficients were also utilized in this study and extracted from the windowed signals of each channel and each coefficient of the 2-D DT $\mathbb{C}$WT. The window length was set to 25 msec, while the window period was set to 12.5 msec. The real parts of the lower 15 cepstral coefficients (including the 0th coefficients), Δ features, and $\Delta\Delta$ features were used as features. The cepstral coefficients calculated from the all 56 channels and/or the all 1344 coefficients of the 2-D DT $\mathbb{C}$WT were concatenated as features for each feature sets. Because this concatenation makes feature dimension so high, feature selection, which is described in the next section, is necessary to avoid the curse of dimensionality.

E. Feature Selection

In this study, we used SDA [20], [21] proposed by Clemmensen et al. for feature selection as was used in our previous study [19]. SDA can perform feature selection simultaneously with dimension reduction by imposing sparseness constraint. SDA software in MATLAB is available from [21].

Let $\mathbf{X}$ denote an $n \times p$ data matrix with observations down the rows and features in the columns, and let $\mathbf{Y}$ denote an $n \times K$ (classes) matrix of dummy variables which indicate belonging classes. Clemmensen et al. defined the sparse optimal scoring criterion as

$$\arg\min_{\boldsymbol{\theta},\boldsymbol{\beta}} n^{-1}(\|\mathbf{Y}\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\Omega}^{\frac{1}{2}}\boldsymbol{\beta}\|_2^2 + \gamma\|\boldsymbol{\beta}\|_1) \ , \quad (1)$$

$$\text{subject to } n^{-1}\|\mathbf{Y}\boldsymbol{\theta}\|_2^2 = 1 \ . \quad (2)$$

where $\boldsymbol{\beta}$ is a $p \times q$ matrix of parameters which leads to q components of directions, $\boldsymbol{\theta}$ is $K \times q$ matrix of scores, λ and γ are nonnegative tuning parameters, and $\boldsymbol{\Omega}$ is a symmetric positive definite matrix. This method involves recasting the classification problem as a regression problem by turning categorical variables into quantitative variables, via $\boldsymbol{\theta}$. Iterative algorithm is used for finding a local minimum of the criterion (1) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. For fixed $\boldsymbol{\theta}$, $\boldsymbol{\beta}_j, j = 1, \ldots, q$, is obtained by solving the modified elastic net problem [28]:

$$\boldsymbol{\beta}_j = \arg\min_{\boldsymbol{\beta}_j} n^{-1}(\|\mathbf{Y}\boldsymbol{\theta}_j - \mathbf{X}\boldsymbol{\beta}_j\|_2^2 + \lambda\boldsymbol{\beta}_j^T\boldsymbol{\Omega}\boldsymbol{\beta}_j + \gamma\|\boldsymbol{\beta}_j\|_1) \ . \quad (3)$$

When γ is large, the L_1 penalty on $\boldsymbol{\beta}_j$ results in sparseness. For fixed $\boldsymbol{\beta}$, the criterion becomes

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} n^{-1}\|\mathbf{Y}\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \ , \quad (4)$$

$$\text{subject to } n^{-1}\|\mathbf{Y}\boldsymbol{\theta}\|_2^2 = 1 \ . \quad (5)$$

Steps related to the equations (3) and (4) are iterated until convergence or until a maximum number of iterations is reached.

In this study, we set the number of selected features per component to 100, and λ to 0.01, which were based on findings from our preliminary study [19]. The number of components q was set to 4. We applied the SDA to each of the feature sets.

F. Vowel Recognition

Continuous HMM was adopted for vowel modeling, since it has been shown that the HMM is effective for sEMG-ASR as well as for acoustic ASR. An HMM represents a stochastic process that takes sequential data as the inputs, and outputs the probabilities that the data are generated by the model. For each vowel, we used a 9 state left-to-right HMM with 3 Gaussian mixtures, whose covariance matrices in each state are diagonal. Expectation maximization (EM) algorithm [29] was utilized in parameter estimation, and the vowel with the maximum likelihood was adopted as the recognition result. Hidden Markov Model Toolbox [30] was used to implement the HMMs in this experiment. 5-fold cross-validations were conducted to calculate the recognition accuracies.

III. RESULT AND DISCUSSION

Fig. 2 shows recognition accuracies obtained with different feature sets calculated from the channel signals, the coefficients of the 2-D DT ℂWT, and both of them. The recognition accuracies with the coefficients of the 2-D DT ℂWT outperformed those with the channel signals only with respect to the subject 2. For subjects 3, 4, and 5, the recognition accuracies obtained with the coefficients of the 2-D DT ℂWT indicated the highest values among the three conditions. Although, on average, the results with the channel signals showed the best performance, including the coefficients of the 2-D DT ℂWT could improve recognition accuracies for the four of six subjects. Note that the parameters used in feature selection were optimized for the condition with the channel signals of subject 1 in our previous study [19] and there is no guarantee that these parameters are also optimal for the other feature sets that include features based on the coefficients of the 2-D DT ℂWT. This result suggests that shift invariance is important property that should be taken into account in order to compensate the deviation of attached site of the electrode grid. Also, this result is based on the decomposition level set to two that can be insufficient to extract spatial shift invariant information. For the future work, we are planning to deal with higher decomposition level. In addition, the 3-D DT ℂWT will be included in our future works, because sMEG signals spread spatio-temporally.

IV. CONCLUSIONS

We verified the efficiency of the shift invariant feature extraction method in improving the vowel recognition accuracy based on sEMG obtained with the electrode grid. The 2-D dual tree complex wavelet transform was employed as the shift invariant feature extraction method in this study. We conclude that shift invariant feature can provide additional information that cannot be provided by the channel signals used independently. Our approach has a possibility to overcome the deviation of the attached site of the electrode grid and realize a robust sEMG-ASR system against it.

References

[1] Y. Deng, R. Patel, J. T. Heaton, G. Colby, L. D. Gilmore, J. Cabrera, S. H. Roy, C. J. De Luca, and G. S. Meltzner, "Disordered speech recognition using acoustic and sEMG signals," in *INTERSPEECH*, Brighton, UK, 2009, pp. 644–647.

[2] O. Fukuda, S. Fujita, and T. Tsuji, "A substitute vocalization system based on emg signals," *IEICE Transactions on Information and Systems*, vol. J88-D-2, no. 1, pp. 105–112, 2005.

[3] 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.

[4] A. D. C. Chan, K. Englehart, B. Hudgins, and D. Lovely, "Myo-electric signals to augment speech recognition," *Medical and Biological Engineering and Computing*, vol. 39, no. 4, pp. 500–504, 2001.

[5] A. D. C. Chan, K. B. Englehart, B. Hudgins, and D. Lovely, "Multiexpert automatic speech recognition using acoustic and myoelectric signals," *IEEE Trans Biomed Eng*, vol. 53, no. 4, pp. 676–85, 2006.

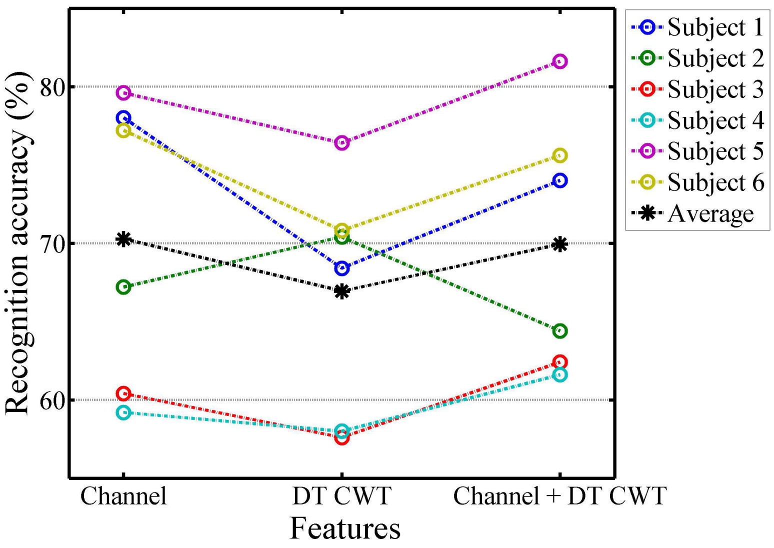

Fig. 2. Comparison of recognition accuracies between the conditions with the different sets of the features. “*Channel*”, “*DT CWT*”, and “*Channel + DT CWT*” denote the recognition accuracies obtained by using features based on the channel signals, the coefficients of the 2-D DT ℂWT, and both of them, respectively.

[6] S. C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, “Towards continuous speech recognition using surface electromyography,” in *INTERSPEECH*, Pittsburgh, PA, Sep 2006, pp. 573–576.

[7] S. C. Jou, L. Maier-Hein, T. Schultz, and A. Waibel, “Articulatory feature classification using surface electromyography,” in *ICASSP*, Toulouse, France, May 2006, pp. 605–608.

[8] K. S. Lee, “SNR-adaptive stream weighting for audio-MES ASR,” *IEEE Trans Biomed Eng*, vol. 55, no. 8, pp. 2001–2010, 2008.

[9] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, “Session independent non-audible speech recognition using surface electromyography,” in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Cancun, Mexico, 2005, pp. 331–336.

[10] H. Manabe and Z. Zhang, “Multi-stream hmm for emg-based speech recognition,” in *IEMBS’04. 26th Annual International Conference of the IEEE*, vol. 2, San Francisco, CA, 2004, pp. 4389–4392.

[11] H. Manabe, A. Hiraiwa, and T. Sugimura, “Unvoiced speech recognition using EMG - mime speech recognition,” in *CHI ’03 Extended Abstracts on Human Factors in Computing Systems*, Ft. Lauderdale, FL, 2003, pp. 794–795.

[12] G. S. Meltzner, J. Sroka, J. T. Heaton, L. D. Gilmore, G. Colby, S. H. Roy, N. Chen, and C. De Luca, “Speech recognition for vocalized and subvocal modes of production using surface EMG signals from the neck and face,” in *INTERSPECH*, Brisbane, Australia, 2008, pp. 2667–2670.

[13] T. Schultz and M. Wand, “Modeling coarticulation in EMG-based continuous speech recognition,” *Speech Communication*, vol. 52, no. 4, pp. 341–353, 2010.

[14] Q. Zhou, N. Jiang, K. Englehart, and B. Hudgins, “Improved phoneme-based myoelectric speech recognition,” *IEEE Trans Biomed Eng*, vol. 56, no. 8, pp. 2016–23, 2009.

[15] B. G. Lapatki, J. P. van Dijk, I. E. Jonas, M. J. Zwarts, and D. F. Stegeman, “A thin, flexible multielectrode grid for high-density surface EMG,” *Journal of Applied Physiology*, vol. 96, no. 1, pp. 327–336, 2004.

[16] B. G. Lapatki, R. Oostenveld, J. P. Van Dijk, I. E. Jonas, M. J. Zwarts, and D. F. Stegeman, “Topographical characteristics of motor units of the lower facial musculature revealed by means of high-density surface EMG,” *Journal of Neurophysiology*, vol. 95, no. 1, pp. 342–354, 2006.

[17] ——, “Optimal placement of bipolar surface EMG electrodes in the face based on single motor unit analysis,” *Psychophysiology*, vol. 47, no. 2, pp. 299–314, 2010.

[18] T. Kubo, T. Toda, M. Yoshida, T. Hattori, and K. Ikeda, “Vowel recognition based on surface electromyography with electrode grid on submental region,” *Trans Jpn Soc Med Biol Eng*, vol. 50, no. 1, pp. 38–46, 2012.

[19] T. Kubo, M. Yoshida, T. Hattori, and K. Ikeda, “Feature selection for vowel recognition based on surface electromyography derived with multichannel electrode grid,” in *2011 Sino-foreign-interchange Workshop on Intelligence Science and Intelligent Data Engineering*, ser. Lecture Notes in Computer Science 7202, Xi’an, China, Oct 2011, pp. 242–249.

[20] L. Clemmensen, T. Hastie, D. Witten, and B. Ersbøll, “Sparse discriminant analysis,” *Technometrics*, vol. 53, no. 4, pp. 406–413, 2011.

[21] L. Clemmensen, Line Clemmensen.
Online: `http://www2.imm.dtu.dk/~lhc/index.html`. [accessed January 17, 2013].

[22] N. G. Kingsbury, “Shift invariant properties of the dual-tree complex wavelet transform,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, Phoenix, Az, 1999, pp. 1221–1224.

[23] ——, “The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters,” in *Proc. 8th IEEE DSP Workshop*, vol. 8, Bryce Canyon, UT, 1998, p. 86.

[24] ——, “The dual-tree complex wavelet transform: a new efficient tool for image restoration and enhancement,” in *Proc. EUSIPCO*, vol. 98, Rhodes, Greece, pp. 319–322.

[25] I. W. Selesnick, R. G. Baraniuk, and N. G. Kingsbury, “The dual-tree complex wavelet transform,” *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 123–151, 2005.

[26] T. Hattori, T. Sato, K. Minato, H. Nakamura, and M. Yoshida, “An identification method of motor units using a 3D template from grid surface electromyography,” *Trans Jpn Soc Med Biol Eng*, vol. 46, no. 2, pp. 268–274, 2008.

[27] S. Cai, K. Li, and I. W. Selesnick, Matlab Implementation of Wavelet Transforms.
Online: `http://eeweb.poly.edu/iselesni/WaveletSoftware/index.html`. [accessed January 17, 2013].

[28] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society, Series B-Statistical Methodology*, vol. 67, pp. 301–320, 2005.

[29] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society, Series B-Statistical Methodology*, vol. 39, no. 1, pp. 1–38, 1977.

[30] K. Murphy, Hidden Markov Model (HMM) Toolbox for Matlab.
Online: `http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html`. [accessed November 2, 2012].