INTEGRATING PERCEIVERS NEURAL-PERCEPTUAL RESPONSES USING A DEEP VOTING FUSION NETWORK FOR AUTOMATIC VOCAL EMOTION DECODING

Wan-Ting Hsieh[1], *Hao-Chun Yang*[1], *Ya-Tse Wu*[1], *Fu-Sheng Tsai*[1], *Li-Wei Kuo*[2], *Chi-Chun Lee*[1]

[1]Department of Electrical Engineering, National Tsing Hua University
[2]Institute of Biomedical Engineering and Nanomedicine, National Health Research Institute, Taiwan
cclee@ee.nthu.edu.tw

ABSTRACT

Understanding neuro-perceptual mechanism of vocal emotion perception continues to be an important research direction not only in advancing scientific knowledge but also in inspiring more robust affective computing technologies. The large variabilities in the manifested fMRI signals among subjects has been shown to be due to the effect of individual difference, i.e., inter-subject variability. However, relatively few works have developed modeling techniques in task of automatic neuro-perceptual decoding to handle such idiosyncrasies. In our work, we propose a novel computation method of deep voting fusion neural network architecture by learning an adjusted weight matrix applied at the fusion layer. The framework achieves an unweighted average recall of 53.10% in a four-class vocal emotion states decoding task, i.e., a relative improvement of 8.9% over a two-stage SVM decision-level fusion. Our framework demonstrates its effectiveness in handling individual differences. Further analysis is conducted to study the properties of the learned adjusted weight matrix as a function of emotion classification accuracy.

Index Terms— individual difference, fMRI, vocal emotion perception, deep voting fusion neural net

1. INTRODUCTION

Investigating human brain's activities using BOLD (blood-oxygen-level-dependent) signal captured from functional magnet resonance imaging (fMRI) has brought a vast amount of valuable insights in understanding the underlying complex neural mechanism of emotion perception (e.g., [1, 2]). Variabilities existed in the BOLD signals consist of multiple complex factors, mostly due to the diversity in humans that makes everyone differs from one another. In fact, studying of neural mechanism has recently emphasized the importance of individual differences. For example, several past research works have demonstrated that by simply averaging neural responses of subjects would inadvertently eliminate important information about brain structures and functions [3, 4, 5]; Canli et al. also indicate that large variabilities between subjects may result in unwanted low significant values leading to unfavorable false interpretations [6]. Furthermore, in an extended study, Hamann et al. have shown that the identification of brain regions responsible for emotion processing is largely affected by individual differences [7].

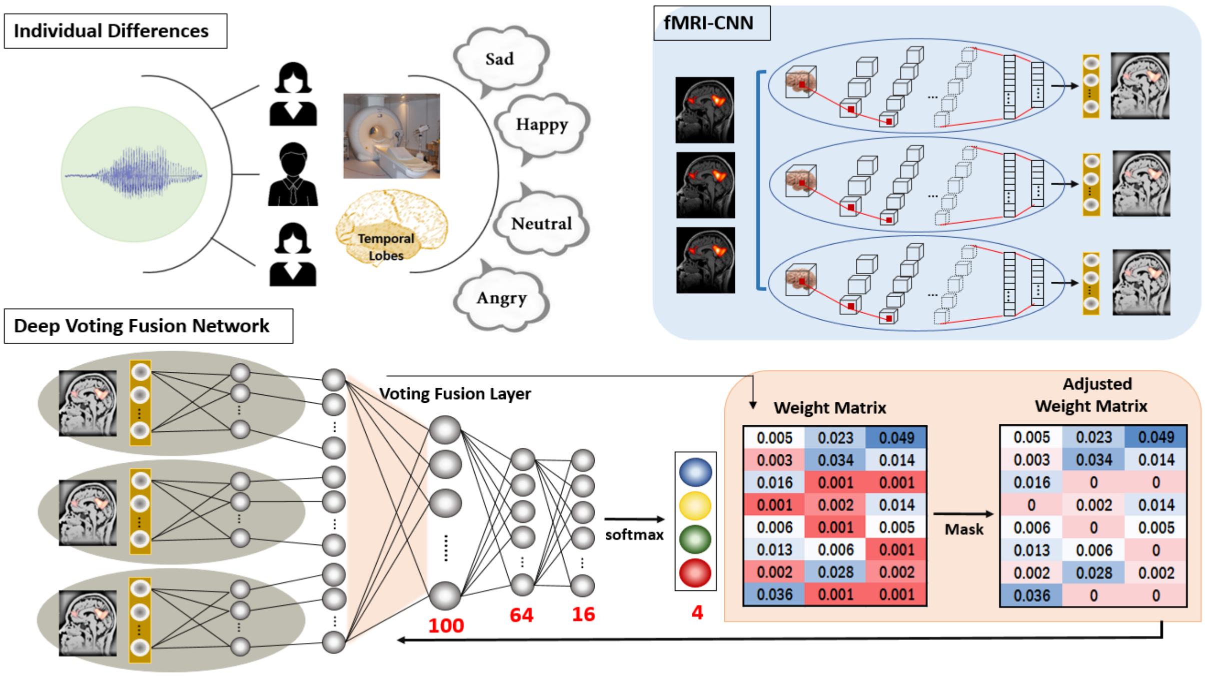
Within the domain the neuro-scientific studies, the method in addressing individual differences is often to perform correlation study at the individual level, e.g., Dubois et al. focus on validating scientific hypothesis and the reliability of the collected fMRI signal by generating correlation plots at the individual level [8]; Parasuraman et al. use similar methods to observe how individual variation influences cognitive processes of working memory and decision-making [9]. These methods reveal the importance of considering individual difference. However, in the context of developing algorithms for automatic decoding humans emotion perception from fMRI data (e.g., [10, 11, 12]), few modeling techniques have integrated this component beyond conventional method of decision-level fusion.

In this work, we propose a novel deep voting fusion neural network architecture that directly integrates individual's fMRI neural representations for task of automatic decoding of vocal emotion perception. Our proposed deep voting fusion neural network introduces the use of fusion layer and learns an adjusted weighted matrix, then the entire emotion decoding network is fine-tuned after re-initializing the fusion layer with the adjusted weights. We conduct our experiment in a database consists of 18 subjects, where each individual is presented with 251 emotional utterances stimuli gathered from the USC IEMOCAP database [13]. Our framework achieves an unweighed average recall (UAR) of 53.10% in a four-class emotion decoding task, which improves 8.9% relative over a popular two-stage decision-level fusion technique. The rest of paper is organized as follows: section 2 describes about emotion stimuli design and collection and our proposed framework, section 3 includes experimental setups and results, and section 4 concludes with future work.

2. RESEARCH METHODOLOGY

2.1. Vocal Emotion Stimuli Design and Collection

Our vocal emotion stimuli dataset is designed from the USC IEMOCAP database [13]. The same set of stimuli was also



Weight Matrix

0.005	0.023	0.049
0.003	0.034	0.014
0.016	0.001	0.001
0.001	0.002	0.014
0.006	0.001	0.005
0.013	0.006	0.001
0.002	0.028	0.002
0.036	0.001	0.001

Adjusted Weight Matrix

0.005	0.023	0.049
0.003	0.034	0.014
0.016	0	0
0	0.002	0.014
0.006	0	0.005
0.013	0.006	0
0.002	0.028	0.002
0.036	0	0

Fig. 1. *A schematic of our proposed deep voting fusion neural work in performing automatic 4-class vocal emotion decoding.*

used in the study of brain connectivity of vocal emotion [14] and joint modeling between the BOLD signal time series and prosodic features [15]. This dataset consists of six different stimuli, and each lasts for 5 minutes when presented to the subjects. These stimuli are gathered from a single speaker in the USC IEMOCAP database. There is a total of 251 utterances from the database being used as the vocal emotion stimuli presented in this work.

2.1.1. Emotion Classes

Since the original labels offered by the USC IEMOCAP database on these 251 utterance are distributed unevenly across 8 different emotion classes, where some of the classes include only few data samples. Thus, we merge the 8 classes into 4 distinctive emotion classes according to the valence-activation representation of categorical emotion [16]. Table 1 lists the original and the merged emotion label classes with their associated numbers of samples. In the following classification experiments, we use these four emotion classes.

2.1.2. fMRI Data Collection and Preprocessing

All of the participated subjects are aged between 20-35 years old with college-level education (18 subjects). Each participant listens to three 5-minute long continuous vocal emotion stimuli and has a 5-minute break in-between. MRI scanning is performed on a 3T scanner (Prisma, Siemens, Germany). Anatomical images with spatial resolution of $1 * 1 * 1mm^3$ (T1-weighted MPRAGE sequence) are acquired using an EPI sequence (TR/TE= $3000/30$ms, voxel size $=3 * 3 * 3mm^3$, 40 slices, and 100 repetitions). We perform all necessary preprocessing steps using the DPARSF toolbox [17] and interpolate the MRI images to 1 second time interval.

Table 1. *Summary of the original and the merged labels of the 251 utterances used in this work*

Original	Number	Merged Classes	Number
Sad	33	Class 1	33
Happy	12	Class 2	79
Excited	64		
Surprise	3		
Neutral	69	Class 3	69
Angry	19	Class 4	70
Distress	1		
Frustrated	50		

2.2. fMRI-CNN Representation

Prior work on the same dataset has shown that the temporal lobe possesses the most vocal emotionally-relevant information [10]. In this work, we use the same exact structure in deriving convolutional neural network (CNN) representation for each subject at their associated temporal lobe region. A brief description of the CNN structure is given below.

There are a total of 11 hidden layers in the training of fMRI-CNN representation: including 4 convolutional layers, 3 pooling layers, 3 fully connected layers, and 1 softmax layer. Hyper-parameter settings are: activation function of Relu, weight decay: 0.000001, momentum: 0.9, learning rate: 0.0001, epoch 20 times. The training accuracy achieves around 88% to 95%. Finally, we extract the output of tenth hidden layer (500 nodes) as the feature for each 3-D scanning image. Since there are multiple images per utterance, max pooling over temporal dimension is used to generate the final representation of each subject at an emotional utterance level.

2.3. Deep Voting Fusion Architecture

Our proposed framework is a deep voting fusion network architecture that learns to fuse multiple subjects neural responses for automatic emotion decoding as illustrated in Figure 1. There are a total of 5 hidden layers. The first dense layer condenses each individual subject-wise 500 dimensional fMRI-CNN representation to a lower-dimension of 100. The second layer simply concatenates all subjects 100-dimensional features in a merged representation. The third layer is a *voting fusion* layer with two additional fully-connected layers (64 and 16 dimensions respectively), and finally a softmax layer for emotion prediction. Categorical cross-entropy is used as the loss function; other hyperparameter settings include activation function of Relu, Adam as optimizer, learning rate set at 0.0001, epoch for 10 times.

2.3.1. Voting Fusion Layer

We introduce the use of voting fusion layer, f, that fuses multiple subjects fMRI neural responses in our architecture. It is defined as below:

$$D_f = W_f \times D_2 \tag{1}$$

where D_f is the output of the fusion layer f, W_f refers to the weight matrix of fusion layer, and D_2 is the concatenated features from each subject output from the second layer. Note the absence of activation function and bias function in this particular layer as compared to the standard neural network in order to represent the weights as voting operation (i.e., contribution from the merged feature output layer, D_2).

2.3.2. Deep Voting Fusion Network (DVFN)

Our proposed deep voting fusion network (DVFN) is a two-pass procedure: 1) masking fusion layer weights after learning the first-pass model, and 2) fine-tuning again the entire network after applying adjusting fusion layer weights from step 1. The weight matrix of fusion layer, W_f can be seem as reflecting the contribution from each individual fMRI-CNN features. In step 1, we apply a mask on W_f by introducing a threshold of τ:

$$f(w) = \begin{cases} 1, & \text{if } |w| \geq \tau \\ 0, & \text{if } |w| < \tau \end{cases} \tag{2}$$

then the adjusted weight can be acquired by,

$$W_{\text{adjusted}} = W_f \times f(w) \tag{3}$$

This step effectively emphasizes the node with higher contribution and resets the low-valued weights as noise. Then in step 2, the entire deep voting fusion network is fine-tuned again using back-propagation by replacing the original W_f with W_{adjusted}.

3. EXPERIMENTAL SETUP AND RESULTS

We carry out experiments for 4-class automatic emotion decoding on the 251 utterances. The evaluation scheme is carried out using leave-one-utterance-out cross-validation, where the learning of fMRI-CNN and deep voting fusion network are all contained within the training set. The performance measure is unweighted average recall (UAR). We compare our framework with the following methods:

- **AVB Average**: perform emotion classification directly using multi-class SVM on fMRI-CNN derived from the average of raw imaging data of all subjects
- **INB Individual**: perform emotion classification directly using multi-class SVM on fMRI-CNN derived from subjects individual raw imaging data
- **INB SVM-Voting**: perform emotion classification by learning a decision-level fusion with a second stage SVM trained on the decision functions outputted from each individual subjects multi-class SVM [10]
- **INB DNN-Fusion**: perform emotion classification using our voting fusion network without adjustment in the fusion layer weight
- **INB DNN-SVM-Voting**: perform emotion classification by using SVM-voting (two-stage late-decision fusion technique mentioned above) operating on the prediction output of the voting fusion network with adjustment in the fusion layer weight
- **INB DVFN DNN-Fusion**: our proposed deep voting fusion network architecture (section 2.3)

3.1. Emotion Classification Results

Table 2 summarizes all the results in our experiments. Our proposed deep voting fusion network (DVFN) obtains the best accuracies among all methods compared. It achieves 53.1% in 4-class UAR, which is 8.9% relative improvement over previously-published work, i.e., the method of INB SVM-Voting, on the same task [10]. There are several points to make from our results. By simply treating individual subject separately (INB vs. AVB), there is already an improvement in the accuracy - a result reinforcing the trend in stressing the importance of individual-level modeling of neural responses.

The voting neural network strategy, i.e., the voting fusion layer (section 2.3.1), which learns to vote jointly as part of the neural network architecture is beneficial in improving emotion classification. In specifics, the method that uses voting fusing layer outperforms the conventional two-staged SVM-Voting based decision-level fusion technique. Lastly, by re-training the network after applying adjusted voting weight matrix improves 6.9% relative (DVFN DNN-Voting vs. DNN-Fusion) compared to the one without. The use of adjusted weight matrix ensures the network to concentrate more heavily on relevant subjects and their contributed features to improve the overall automatic emotion decoding accuracy.

Table 2. *It presents the 4-class emotion classification results of our proposed deep voting fusion neural network and other fusion techniques. The accuracy is measured in UAR (%).*

4-Class	**AVB: Average**	**INB: Individual**	**INB** **SVM-Voting**	**INB** **DNN-Fusion**	**INB** **DNN-SVM-Fusion**	**INB** **DVFN DNN-Voting**
Class 1	15.15	12.24	15.15	15.15	15.15	24.24
Class 2	72.15	77.43	84.81	89.87	87.34	87.34
Class 3	44.93	46.41	55.07	55.07	57.97	56.52
Class 4	37.14	40.87	40.00	38.57	47.14	44.29
UAR	42.34	44.31	48.75	49.67	51.90	**53.10**

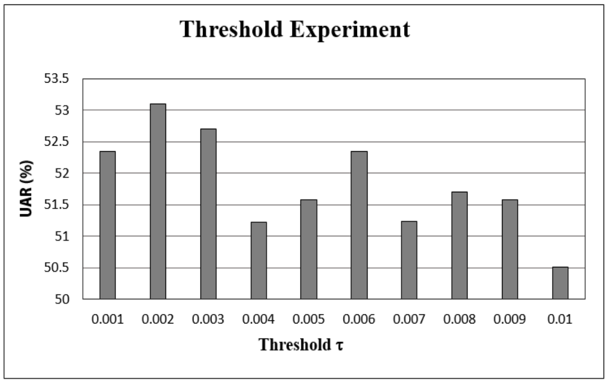


Fig. 2. *The 4-class motion classification performance measured in UAR (%) as a function of threshold (τ) value*

3.1.1. Threshold Analysis

In the work, we additionally present different accuracies obtained on our proposed deep voting fusion network by altering thresholds, τ, values ranging from 0.001 to 0.01 (see Figure 2). By setting the threshold to be 0.002, the best accuracy can be obtained in the 4-class emotion classification experiment. In general, we observe that setting the threshold to be a lower value (e.g., ≤ 0.003) seems to result in high accuracies compared to otherwise. Forcing more weights to be zero at the fusion layer, in this case, provides a more robust modeling of multi-perceivers neural responses.

4. CONCLUSION AND FUTURE WORK

The individual difference existed in the humans neural responses not only poses challenge in the study of emotional perception and other higher-cognitive functioning of human brains, but also requires continuous advancements in the modeling technique to handle such a complexity. In this work, we propose a novel technical framework of deep voting fusion network to integrate multiple subjects neural responses in task of automatic vocal emotion decoding. By introducing the use of voting layer with adjusted weight matrix, we demonstrate that it obtains a significant improvement over conventional approach of two-stage late fusion technique. Furthermore, our analysis reveals that by forcing more weights to be zero at the fusion layer is likely to increase the accuracy in modeling human brains fMRI neural responses .

There are multiple future directions. One of the immediate directions is to incorporate a sparsity constraint on weight matrix at the fusion layer such that the optimal weights of the proposed deep voting fusion architecture can be optimized jointly. Secondly, the use of time-series model, such as recurrent neural network (RNN) and long-short term memory neural network (LSTM), in modeling EEG and other similar brain signals have been shown to be useful in tasks of automatic emotion decoding [18, 19]. While fMRI images are scanned with coarser temporal resolution, we plan to incorporate the temporal aspect of fMRI brain responses into our modeling. Lastly, one of the future goals is to investigate in details about the effect of individual difference on the underlying vocal emotion perception as a function of perceiver's background, which will hopefully bring further quantitative insights in understanding the source of variabilities in the measured fMRI signals.

5. REFERENCES

[1] Tiantong Zhou, Hailing Wang, Ling Zou, Renlai Zhou, and Nong Qian, "A study of neural mechanism in emotion regulation by simultaneous recording of eeg and fmri based on ica," in *International Symposium on Neural Networks*. Springer, 2013, pp. 44–51.

[2] Philippe Fossati, Stephanie J Hevenor, Simon J Graham, Cheryl Grady, Michelle L Keightley, Fergus Craik, and Helen Mayberg, "In search of the emotional self: an fmri study using positive and negative emotional words," *American Journal of Psychiatry*, vol. 160, no. 11, pp. 1938–1945, 2003.

[3] John Darrell Van Horn, Scott T Grafton, and Michael B Miller, "Individual variability in brain activity: a nuisance or an opportunity?," *Brain imaging and behavior*, vol. 2, no. 4, pp. 327, 2008.

[4] Stuart WS MacDonald, Lars Nyberg, Johan Sandblom, Håkan Fischer, and Lars Bäckman, "Increased response-time variability is associated with reduced inferior parietal activation during episodic recognition in

aging," *Journal of Cognitive Neuroscience*, vol. 20, no. 5, pp. 779–786, 2008.

[5] Ryota Kanai and Geraint Rees, "The structural basis of inter-individual differences in human behaviour and cognition," *Nature reviews. Neuroscience*, vol. 12, no. 4, pp. 231, 2011.

[6] Turhan Canli, Heidi Sivers, Susan L Whitfield, Ian H Gotlib, and John DE Gabrieli, "Amygdala response to happy faces as a function of extraversion," *Science*, vol. 296, no. 5576, pp. 2191–2191, 2002.

[7] Stephan Hamann and Turhan Canli, "Individual differences in emotion processing," *Current opinion in neurobiology*, vol. 14, no. 2, pp. 233–238, 2004.

[8] Julien Dubois and Ralph Adolphs, "Building a science of individual differences from fmri," *Trends in cognitive sciences*, vol. 20, no. 6, pp. 425–443, 2016.

[9] Raja Parasuraman and Yang Jiang, "Individual differences in cognition, affect, and performance: Behavioral, neuroimaging, and molecular genetic approaches," *Neuroimage*, vol. 59, no. 1, pp. 70–82, 2012.

[10] Ya-Tse Wu, Hsuan-Yu Chen, Yu-Hsien Liao, Li-Wei Kuo, and Chi-Chun Lee, "Modeling perceivers neural-responses using lobe-dependent convolutional neural network to improve speech emotion recognition," *Proc. Interspeech 2017*, pp. 3261–3265, 2017.

[11] Lucy Alba-Ferrara, Markus Hausmann, Rachel L Mitchell, and Susanne Weis, "The neural correlates of emotional prosody comprehension: disentangling simple from complex emotion," *PLoS one*, vol. 6, no. 12, pp. e28701, 2011.

[12] Annett Schirmer and Sonja A Kotz, "Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing," *Trends in cognitive sciences*, vol. 10, no. 1, pp. 24–30, 2006.

[13] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[14] Leimin Tian, Johanna D Moore, and Catherine Lai, "Emotion recognition in spontaneous and acted dialogues," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 698–704.

[15] Hsuan-Yu Chen, Yu-Hsien Liao, Heng-Tai Jan, Li-Wei Kuo, and Chi-Chun Lee, "A gaussian mixture regression approach toward modeling the affective dynamics between acoustically-derived vocal arousal score (vc-as) and internal brain fmri bold signal response," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5775–5779.

[16] JA Ressel, "A circumplex model of affect," *J. Personality and Social Psychology*, vol. 39, pp. 1161–78, 1980.

[17] Yan Chao-Gan and Zang Yu-Feng, "Dparsf: a matlab toolbox for pipeline data analysis of resting-state fmri," *Frontiers in systems neuroscience*, vol. 4, 2010.

[18] Xiang Li, Dawei Song, Peng Zhang, Yuexian Hou, and Bin Hu, "Deep fusion of multi-channel neurophysiological signal for emotion recognition and monitoring," *International Journal of Data Mining and Bioinformatics*, vol. 18, no. 1, pp. 1–27, 2017.

[19] Mohammad Soleymani, Sadjad Asghari-Esfeden, Yun Fu, and Maja Pantic, "Analysis of eeg signals and facial expressions for continuous emotion detection," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 17–28, 2016.