INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Modeling Perceivers Neural-Responses using Lobe-dependent Convolutional Neural Network to Improve Speech Emotion Recognition

Ya-Tse Wu[1], *Hsuan-Yu Chen*[1], *Yu-Hsien Liao*[1], *Li-Wei Kuo*[2], *Chi-Chun Lee*[1]

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]Institute of Biomedical Engineering and Nanomedicine, National Health Research Institute, Taiwan

cclee@ee.nthu.edu.tw

Abstract

Developing automatic emotion recognition by modeling expressive behaviors is becoming crucial in enabling the next generation design of human-machine interface. Also, with the availability of functional magnetic resonance imaging (fMRI), researchers have also conducted studies into quantitative understanding of vocal emotion perception mechanism. In this work, our aim is two folds: 1) investigating whether the neural-responses can be used to automatically decode the emotion labels of vocal stimuli, and 2) combining acoustic and fMRI features to improve the speech emotion recognition accuracies. We introduce a novel framework of lobe-dependent convolutional neural network (LD-CNN) to provide better modeling of perceivers neural-responses on vocal emotion. Furthermore, by fusing LD-CNN with acoustic features, we demonstrate an overall 63.17% accuracies in a four-class emotion recognition task (9.89% and 14.42% relative improvement compared to the acoustic-only and the fMRI-only features). Our analysis further shows that temporal lobe possess the most information in decoding emotion labels; the fMRI and the acoustic information are complementary to each other, where neural-responses and acoustic features are better at discriminating along the valence and activation dimensions, respectively.

Index Terms: speech emotion recognition, convolutional neural network (CNN), affective computing, fMRI

1. Introduction

Imagining humans as complex dynamical systems, i.e., systems that are characterized by multiple interacting layers of hidden states producing expressive multimodal behavior signals (e.g., body gestures, facial expressions, and speech, etc) have sparked a variety of computational effort in modeling these internal states and behaviors using measurable signals resulting in fields such as affective computing [1], social signal processing [2], and behavioral signal processing [3]. In fact, a vast amount of engineering works already exist in automatic recognition of emotion states from external expressive behaviors, e.g., vocal characteristics [4, 5, 6] and facial expression/body language [7, 8, 9]. Past research also shows that physiological signals, e.g., ECG (electrocardiography) or EEG (electroencephalography), are also indicative of emotion states [10, 11].

Researchers in field of neuroscience have been actively exploring the use of the blood-oxygen-level-dependent (BOLD) signal captured during functional magnetic resonance imaging (fMRI). The BOLD signal is a proxy measure of neuron activations providing quantitative evidence into various studies of neuro-perceptual mechanism. There exists several neuroscience studies in understanding which parts of the human brains are responsible for processing vocal emotion stimuli. For example, Ethofer et al. shows that the activity of superior temporal gyrus increases when exposed to voice-based emotional stimuli [12]; Sander et al. identifies multiple brain areas, e.g., the right amygdala and bilateral superior temporal sulcus, that are responsive to anger prosody [13]. In the present work, our goal is to first investigate whether perceivers internal neural-responses, i.e., acquired using brain imaging techniques as these perceivers being auditorily stimulated with external vocal emotion utterances, can be used to decode the emotion labels of these utterances. We further examine whether such internal neural-responses would possess complementary information to acoustic features in tasks of performing emotion recognition.

In the past, principal component analysis (PCA) operated on BOLD signal time series as feature extractor from fMRI data has been successfully applied in a variety of machine learning tasks in neuroscientific studies (e.g., [14, 15, 16]). Recently, convolutional neural nets (CNNs) have been shown to achieve superior performance in image recognition tasks [17]. Since BOLD signals are derived from the 3-D images, and further the neural activations in response to vocal emotion stimuli have been to shown to be concentrated in specific brain regions, we introduce a novel framework of lobe-dependent convolutional neural network (LD-CNN). We utilize LD-CNN to learn the brain region-based, according to the anatomical categorization of lobe system of human brain [18], feature representations from the fMRI 3D-images. We then perform emotion recognition by fusing the LD-CNN features (internal neuro-perceptual responses) with Fisher-vector encoding of acoustic features (expressive acoustic characteristics).

We carry out our experiment in a 36 subjects (perceivers) database, where each perceiver is presented with three 5-minute long continuous vocal emotion stimuli that are designed from the USC IEMOCAP database [19]. In total there are 251 utterances categorized into four emotion classes. The best fusion of acoustic features and LD-CNN features achieve an unweighted accuracy (UAR) of 63.17%. This result is an improvement of 9.89% and 14.42% relative to the best acoustic-only and the best fMRI-only baselines. Out of the four major lobe systems of human brain, we demonstrate that the temporal lobe carries the most information about the emotional content of the vocal stimuli. Furthermore, our results indicate that these perceivers internal neural-responses seem to possess more discriminatory information along the valence dimension, where the acoustic features are better for discriminating along the activation dimension.

The rest of the paper is organized as follows: section 2 describes about research methodology, section 3 details the experimental setup and results, and section 4 concludes with discussion and future works.

Thanks to Ministry of Science and Technology (103-2218-E-007-012-MY3, 104-2420-H-400-001-MY2) for funding

http://dx.doi.org/10.21437/Interspeech.2017-562

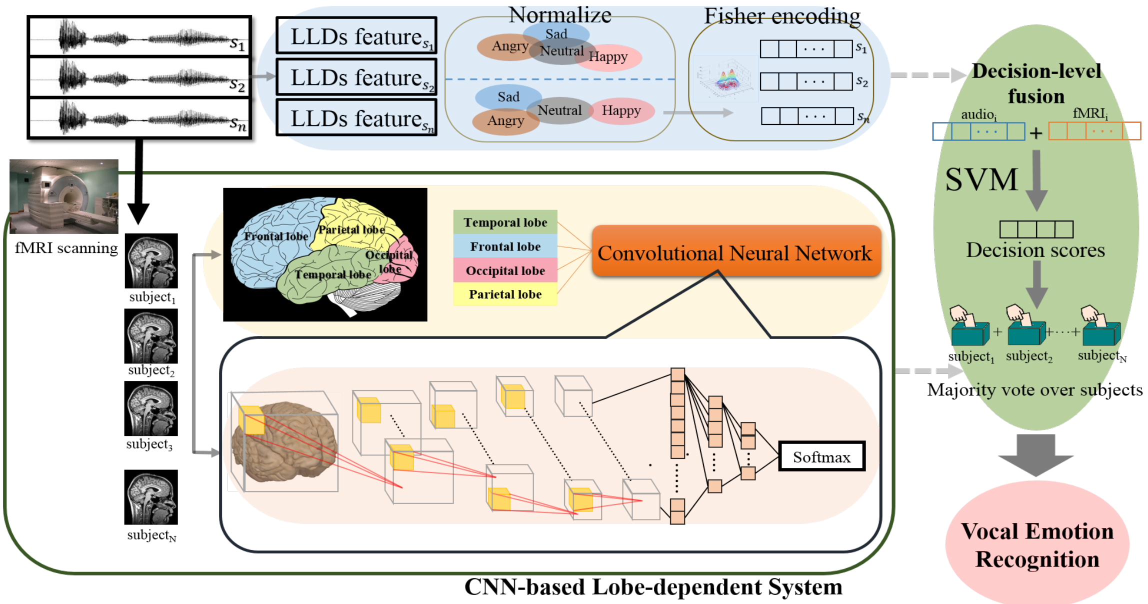

Figure 1: *A schematic of multimodal emotion recognition from audio (Fisher-vector feature representation) and fMRI (Lobe-depedent convolutional neural network-derived feature representation) data*

2. Research Methodology

2.1. Vocal Emotion Stimuli Design and Collection

In this section, we will describe about the dataset used in this work including: vocal stimuli design, relevant emotion labels, and MRI scanning protocols. The vocal emotion stimuli used in our fMRI experiments are from the USC IEMOCAP database [19] and was previously used in the joint modeling between prosody and BOLD signal [20]. They were also used in the study of brain's network connectivity of vocal emotion [21, 22]. There are a total of six different stimuli; each stimulus lasts for 5 minute long. These six different 5-minute long stimuli consist of emotional utterances (put together without context) from a single actor in the database. In total, we use 251 utterances from the database to construct these six stimuli used for MRI scanning and subsequently for this emotion recognition work.

2.1.1. Emotion Labels

In our work, our goal is perform automatic emotion recognition on this set of 251 utterances. While the USC IEMOCAP database provides an emotion label for each utterance, the design of these stimuli was originally for the purpose of understanding neuro-perceptual mechanism at the level of an entire stimulus (5-minute long); hence, the distribution of the original emotion labels is spread across eight different classes. We further merge the original eight different emotion labels into four different classes according to the valence-activation representation of categorical emotion [23]. Table 1 lists the original and merged labels and their associated number of samples. These four emotion classes are the labels of interest for this work.

2.1.2. fMRI Data Collection and Pre-processing

We recruited a total of 36 right-handed healthy subjects (27-male, 9 female, 20-35 years old) with college-level education to participate in our study. 18 of them were stimulated using the same three stimuli, and the rest was stimulated by the remaining three. Each trial included listening to the three 5-minute long continuous vocal emotion stimuli with 5-minute break in between. The subjects were not informed about the details of the experiment a-priori and were only told that this was a study about perception on vocal sounds. They were also required to stay awake during MRI scanning. The order in which the stimulus was presented was random across subjects.

Table 1: *A summary on the number of samples for the original labels and the merged labels (used in this work) of the 251 utterances from the USC IEMOCAP data*

Original	Number	Merged	Number
Sad	33	Class1	33
Happy	12	Class2	79
Excited	64		
Surprise	3		
Neutral	69	Class3	69
Angry	19	Class4	70
Distress	1		
Frustrated	50		

MRI scanning was conducted on a 3T scanner (Prisma, Siemens, Germany). Anatomical images with spatial resolution of $1 \times 1 \times 1\text{mm}^3$ (T1-weighted MPRAGE sequence) were acquired using an EPI sequence (TR/TE= 3000/30ms, voxel size = $3 \times 3 \times 3\text{mm}^3$, 40 slices, and 100 repetitions). We performed all necessary pre-processing steps on the collected MRI data using the DPARSF toolbox [24]. MRI scanning captured one image every 3 seconds, and we additionally performed interpolation to generate an image sample at 1 second time step to handle the varying time-length of utterances within each stimulus.

2.2. Feature Extraction

In this section, we describe briefly our approach of using Fisher-vector encoding on acoustic feature representation, lobe-dependent convolutional neural network on fMRI, and finally the multimodal fusion technique.

2.2.1. Acoustic Feature Representation

We derive a high-dimensional vector as acoustic feature representation for every utterance using two steps: 1) extracting acoustic low-level descriptors (LLDs), and 2) encoding the vari-

Table 2: *The detail list of the structure of convolutional neural network (CNN) for fMRI 3-D brain images used in this work*

Index	Layer 1	Layer 2	Layer 3
Layer	Convolutional	Max-pooling	Convolutional
Parameter	Filter:3,3,3 Node:16 ZeroPad:1,1,1	Pooling:2,2,2 Stride:2,2,2	Filter:3,3,3 Node:32 ZeroPad:1,1,1
Index	Layer 4	Layer 5	Layer 6
Layer	Max-pooling	Convolutional	Convolutional
Parameter	Pooling:2,2,2 Stride:2,2,2	Filter:3,3,3 Node:64 ZeroPad:1,1,1	Filter:3,3,3 Node:64 ZeroPad:1,1,1
Index	Layer 8	Layer 9	Layer 10
Layer	Max-pooling	Fully Connect	Fully Connect
Parameter	Pooling:2,2,2 Stride:2,2,2	Node:2000	Node:1000
Index	Layer 11	Layer 12	
Layer	Fully Connect	Softmax	Dropout
Parameter	Node:500	Node:4	25%

able length sequence of LLDs using Gaussian Mixture Model (GMM) based Fisher-vector (FV) encoding. The list of LLDs includes the first thirteen of MFCCs (Mel-scale Frequency Cepstral Coefficients), pitch, intensity and their first and second order extracted at 60 Hz framerate using the Praat toolkit [25]. We employ a feature scaling approach based on z-normalizing these LLDs with respect to the neutral utterances [26, 27]. Since each utterance is of different lengths, we further adopt the use of GMM-FV approach, i.e., a method that has been shown to obtain competitive accuracy in various computer vision tasks [28, 29] and has also recently been demonstrated to achieve promising accuracies in speech-related tasks [30]. Fisher-vector encoding is operated by first training an overall background GMM and further calculates the gradient vector using FIM (Fisher Information Matrix) approximation to describe the direction changed needed for the trained GMM parameters, i.e., means and variances, to obtain a better fit on the data sample of interest, i.e., a sequence of LLDs per utterance. By employing GMM-FV, we encode the temporal information on the sequence of LLDs into a fixed length vector representation at an utterance level. We set the mixture number equals to four generating the final feature dimension of 45 LLDs $\times$ 2 parameters $\times$ 4 mixtures $= 360$ per utterance.

2.2.2. fMRI Feature Representation

We derive our fMRI feature representation by training convolutional neural network (CNN) on 3-D MRI images of each participant's brain scans. We train five different CNNs per subject: the whole brain and four major human lobe systems (temporal, frontal, occipital, parietal lobe). Each individual lobe system is obtained by applying AAL (anatomical automatic labeling) mask to first split the whole brain into 90 regions (the entire brain has a total of 47636 number of voxels) and further merging the regions into their associated lobe system. The detail list of our CNN structure is shown in Table 2. We use a total of eleven hidden layers: including four convolutional layers, three pooling layers, three fully connected layers, and one softmax layer. We train the CNNs using error propagation and stochastic gradient decent with cross entropy as the loss function; dropout (25%) and regularization are applied to avoid overfitting. Other hyper-parameters are: activation function: Relu, weight decay: 0.000001, momentum: 0, learning rate: 0.0001, epoch 20 times. The training accuracy achieved is around 88% to 95%. We extract the output of the tenth hidden layer (500 nodes) as the feature per 3-D image scan. Each utterance corresponds to multiple time points of CNN-features, we then use *max* pooling over the temporal dimension to derive the final fixed length representation at an utterance level (500 dimensions).

2.3. Multimodal Fusion Paradigm

Since LD-CNN fMRI representations are derived per stimulated perceiver, the technique that we employ to fuse between fMRI and acoustic data is based on two stage late fusion technique. For every subject, the first-layer fusion is carried out using decision score derived from audio and fMRI modalities. Then, in the second layer, we use majority vote over N-fused subjects to generate our final predictions. The classifier of choice is one-versus-all multiclass support vector machine.

3. Experimental Setup and Results

We setup 4-class emotion recognition experiments on the 251 utterances using audio, fMRI, and fusion of audio and fMRI. The evaluation is carried out using leave-one-utterance-out cross-validation. The CNNs are trained within each of the training set, and the decision-level fusions are learned solely on the training set to prevent contamination.

Aside from FV-based representation and LD-CNN-based representation of acoustic and fMRI information, we further compare the performances with respect to the following two conventional baseline systems:

- **Audio:** *EmoLarge-method*
  Computing exhaustive acoustic features using opensmile toolkit [31] with the emolarge configuration
- **fMRI:** *PCA-method*
  Performing fMRI feature extraction using the conventional principal component analysis method

EmoLarge-method, i.e., exhaustive acoustic features of 6506 dimensions are computed per utterance, is a common baseline used in speech-based paralinguistic recognition. PCA is a standard method in dimensional reduction that has been widely used for machine learning tasks in neuroscience. We use PCA as baseline feature extractor for fMRI data using Minka's MLE method to automatically determine the number of dimensions retained [32]. Further temporal pooling over an utterance-length is carried out using max, min, and mean pooling.

3.1. Multimodal Recognition Results

Table 3 summarizes all of our experimental results. Several notable recognition results are summarized below. In the audio modality, FV-encoding on acoustic LLDs achieve an improved UAR (53.28%) compared to using Emo-Large baseline (48.84%), i.e., 4.44% relative improvement. In the fMRI modality, our proposed CNN-based feature representations show dominantly better recognition rates compared to the widely-used PCA-based methods in the literature. The best fMRI-CNN based method is learned from the temporal lobe system (48.75%), i.e., 9.06% relative improvement over the best PCA-based method (fMRI-PCA with max temporal pooling: 39.69%). We also observe that fMRI-based features are significantly skewed better at recognizing Class2 (happy, excited, surprise) and Class 3 (neural).

Examining the columns of "Audio and fMRI Multimodal Fusion" in table 3, we observe that the fusion between the acoustic information and the fMRI data improves the recognition rates in all cases. Further, the four different lobe systems achieve similar recognitions when using fMRI-only features; however, when fusing with audio information, the temporal lobe (TL) provides the most complementary information

Table 3: *It provides a summary of our recognition results using audio-only, fMRI-only, and fusion of the two modalities. The accuracy is measured in unweighted average recall (UAR). max, min, mean indicates the temporal function that PCA-based method used. Finally, ALL, TL, FL, OL, PL indicates temporal, frontal, occipital, parietal lobe respectively.*

Emotion	Audio		fMRI-PCA			fMRI-CNN					Audio and fMRI Multimodal Fusion					
	Emo-Large	**Fisher-V**	**max**	**min**	**mean**	**ALL**	**TL**	**FL**	**OL**	**PL**	**PCA**	**ALL**	**TL**	**FL**	**OL**	**PL**
Class1	45.45	60.61	9.09	12.12	6.06	18.18	15.15	9.09	15.15	15.15	60.61	24.24	57.58	48.48	45.45	51.52
Class2	51.90	44.30	64.56	59.49	68.35	86.08	84.81	82.28	83.54	83.54	41.77	87.34	79.75	78.48	78.48	74.68
Class3	60.87	76.81	59.42	63.77	26.09	49.28	55.07	49.28	53.62	52.17	76.81	66.67	73.91	73.91	78.26	76.81
Class4	37.14	31.42	25.71	17.14	25.71	40.00	40.00	40.00	38.57	44.29	37.14	40.00	41.43	32.86	35.71	38.57
UAR	48.84	53.28	39.69	38.13	31.55	48.38	48.75	45.16	47.72	48.79	54.08	54.56	**63.17**	58.43	59.48	60.40

to the acoustic features. The best fusion accuracy achieved is 63.17% by using acoustic FV representations with CNN-based representation learned from the temporal lobe, which is 9.89% and 14.42% relative improvement to the best acoustic-only and fMRI-only, respectively. Furthermore, we observe that there exists a wide variability in the inter-subject neural responses to the vocal emotion stimuli. Therefore, the two-stage fusion techniques that we employ in this work is essential in obtaining good recognition accuracies. It relies on learning a CNN representation per stimulated subject and performing major votes over the classifiers trained on each *audio-fMRI*-fused subject. This particular methodology outperforms learning a single CNN from all of the stimulated subjects.

3.2. Analysis and Discussion

In this work, we demonstrate that perceivers' neural responses, measured through fMRI, of vocal emotion stimuli indeed possess discriminative power in decoding different emotion classes. One important thing to note is that since these utterances usually last only seconds long, the sequence of brain images used for recognition include little temporal information. The discriminative power, i.e., through the use of CNNs, is a result from modeling the multi-scaled and the nonlinear spatial-connectivity between the local regions of voxels within selected parts of brain. A similar finding is recently shown by using network-based analysis in the study of the relationship between brain's spatial connectivity and vocal emotion stimuli [22]. Furthermore, we show that the temporal lobe possess the most vocal emotion-related information among the four major lobe systems. Aside from the fact that since our emotion stimuli are vocal sounds and the temporal lobe has been known to be in charge of hearing perception in the brain, this result further corroborates well-known research in identifying several sub-parts of temporal lobe, e.g., superior temporal sulcus and amygdala, hold important functions in processing emotion [33, 34].

Another point to make is that by examining the confusion matrices of Audio Fisher-V and fMRI-CNN-TL (Table 4), it is evident that expressive acoustic features and internal neural responses hold complementary information. Audio features are better at discriminating between Class1 vs. Class3, where fMRI features are better at discriminating between Class2 vs. Class 4. Acoustic features possess more emotion information along the axis of arousal dimension compared to the internal neural responses, and fMRI measurements possess more information along the axis valence dimension. It is an interesting finding that may point to the underlying cognitive functioning of higher-level valence assessment in the brain.

Table 4: *Confusion Matrices of Audio-FV and fMRI-CNN-TL*

Audio	Class1	Class2	Class3	Class4
Class1	20	7	0	6
Class2	4	35	4	36
Class3	0	0	53	16
Class4	11	30	7	22

fMRI-TL	Class1	Class2	Class3	Class4
Class1	5	2	18	8
Class2	1	67	6	5
Class3	11	9	38	11
Class4	9	18	15	28

4. Conclusions

In summary, we present a novel study into automatic decoding of vocal emotion states by fusing expressive acoustic features and a novel framework of deriving internal neural responses with lobe-dependent convolutional neural networks (LD-CNN). The multimodal fusion achieves an improved and promising accuracy in a four-class emotion recognition task, and the LD-CNN is shown to possess enhanced modeling power compared to the conventional PCA-based method currently used in many neuroscientific studies. Our recognition results also corroborate the finding that processing of vocal-based emotion information is mostly concentrated in the temporal lobe system. Further analysis reveals that the complementary nature between acoustic and fMRI features; our fMRI features, i.e., the CNN-representations, are better at discriminating vocal emotion states along the dimension of valence, where acoustic features are better along the dimension of activation. This result seems to implicate the more complex and higher-level functioning in the assessment of valence is encoded more in the local spatial coordination (connectivity) of the neural responses within particular brain regions than just the neural activation in isolation.

There are several future directions. One of them is on technically deriving and improving the region-based CNNs from brain imagining with data-driven approach. Our proposed local region segmentation in this work is from broad anatomical structures of human brain. With continuous data collection and larger availability of vocal emotion stimuli-based brain imaging, our next aim is derive data-driven segmentations through further algorithmic development on CNNs in order to uncover the spatial segmentation based on components of emotion functioning in the human brain. Currently, the use of acoustic features in affective computing tasks has shown its robustness mostly in assessing the arousal dimensions. Our analysis implicates the possibility that neural responses may relate more toward the valence dimensions. This insights provides yet another algorithmic venue in deriving robust acoustic representations of valence substantiated by the quantitative evidence of brain imagining to further enhance the modeling power of automatic speech emotion recognizers. Lastly, we plan to collect vocal emotion stimuli using multi-lingual speech sounds with associated perceivers neural responses of multi-cultural backgrounds with an overarching goal to bring additional scientific insights on neuro-perceptual mechanism in vocal emotion decoding with novel algorithmic advancement.

5. References

[1] R. W. Picard and R. Picard, *Affective computing*. MIT press Cambridge, 1997, vol. 252.

[2] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 69–87, 2012.

[3] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[4] A. B. Kandali, A. Routray, and T. K. Basu, "Emotion recognition from assamese speeches using mfcc features and gmm classifier," in *TENCON 2008-2008 IEEE Region 10 Conference*. IEEE, 2008, pp. 1–5.

[5] K. K. Kishore and P. K. Satish, "Emotion recognition in speech using mfcc and wavelet features," in *Advance Computing Conference (IACC), 2013 IEEE 3rd International*. IEEE, 2013, pp. 842–847.

[6] N. Sato and Y. Obuchi, "Emotion recognition using mel-frequency cepstral coefficients," *Information and Media Technologies*, vol. 2, no. 3, pp. 835–848, 2007.

[7] G. Castellano, S. D. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2007, pp. 71–82.

[8] M. Pantic and L. J. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, 2003.

[9] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on affective computing*, vol. 1, no. 1, pp. 18–37, 2010.

[10] J. Cai, G. Liu, and M. Hao, "The research on emotion recognition from ecg signal," in *Information Technology and Computer Science, 2009. ITCS 2009. International Conference on*, vol. 1. IEEE, 2009, pp. 497–500.

[11] M. Li, Q. Chai, T. Kaixiang, A. Wahab, and H. Abut, "Eeg emotion recognition system," in *In-vehicle corpus and signal processing for driver behavior*. Springer, 2009, pp. 125–135.

[12] T. Ethofer, D. Van De Ville, K. Scherer, and P. Vuilleumier, "Decoding of emotional information in voice-sensitive cortices," *Current Biology*, vol. 19, no. 12, pp. 1028–1033, 2009.

[13] D. Sander, D. Grandjean, G. Pourtois, S. Schwartz, M. L. Seghier, K. R. Scherer, and P. Vuilleumier, "Emotion and attention interactions in social cognition: brain regions involved in processing anger prosody," *Neuroimage*, vol. 28, no. 4, pp. 848–858, 2005.

[14] L. Zhang, D. Samaras, D. Tomasi, N. Volkow, and R. Goldstein, "Machine learning for clinical diagnosis from functional magnetic resonance imaging," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 1211–1217.

[15] S.-y. Xie, R. Guo, N.-f. Li, G. Wang, and H.-t. Zhao, "Brain fmri processing and classification based on combination of pca and svm," in *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*. IEEE, 2009, pp. 3384–3389.

[16] A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Muller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux, "Machine learning for neuroimaging with scikit-learn," *arXiv preprint arXiv:1412.3919*, 2014.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[18] M. K. Wolf, "Neuroanatomy text and atlas," 1997.

[19] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[20] H.-Y. Chen, Y.-H. Liao, H.-T. Jan, L.-W. Kuo, and C.-C. 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.

[21] H. Jan, S.-Y. Lin, S.-Y. Chen, Y.-H. Liao, Y.-P. Chao, C.-C. Lee, and L.-W. Kuo, "Voxel-based graph-theoretical analysis (vga) of brain networks modulated by external vocal emotional expressions," in *21st Annual Meeting of the Organization for Human Brain Mapping, Honolulu*, p. 3814.

[22] S.-Y. Lin, C.-P. Lin, L.-L. Liao, C.-C. Lee, and L.-W. Kuo, "Brain network re-configuration during emotional speech assessed by graph theoretical analysis," in *23rd Annual Meeting of the Organization for Human Brain Mapping*, 2017.

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

[24] C. Yan and Y. Zang, "Dparsf: a matlab toolbox for" pipeline" data analysis of resting-state fmri," *Frontiers in systems neuroscience*, vol. 4, p. 13, 2010.

[25] P. P. G. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, 2002.

[26] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE transactions on speech and audio processing*, vol. 13, no. 2, pp. 293–303, 2005.

[27] C. Busso, A. Metallinou, and S. S. Narayanan, "Iterative feature normalization for emotional speech detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5692–5695.

[28] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[29] X. Peng, C. Zou, Y. Qiao, and Q. Peng, "Action recognition with stacked fisher vectors," in *European Conference on Computer Vision*. Springer, 2014, pp. 581–595.

[30] H. Kaya, A. Karpov, and A. A. Salah, "Fisher vectors with cascaded normalization for paralinguistic analysis." in *INTERSPEECH*, 2015, pp. 909–913.

[31] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[32] T. P. Minka, "Automatic choice of dimensionality for pca," in *Nips*, vol. 13, 2000, pp. 598–604.

[33] D. Grandjean, D. Sander, G. Pourtois, S. Schwartz, M. L. Seghier, K. R. Scherer, and P. Vuilleumier, "The voices of wrath: brain responses to angry prosody in meaningless speech," *Nature neuroscience*, vol. 8, no. 2, pp. 145–146, 2005.

[34] R. Adolphs, D. Tranel, H. Damasio, and A. Damasio, "Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala," *Nature*, vol. 372, no. 6507, p. 669, 1994.