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

Hsuan-Yu Chen[1], *Yu-Hsien Liao*[1], *Heng-Tai Jan*[2], *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 Institutes, Taiwan

ABSTRACT

Understanding the underlying neuro-perceptual mechanism of humans' ability to decode emotional content in vocal signal is an important research direction. In this paper, we describe our initial research effort into quantitatively modeling the joint dynamics between measures of vocal arousal and blood oxygen level-dependent (BOLD) signals. We utilize Gaussian mixture regression approach to predict the invoked BOLD signal response as the subject is exposed to various levels of continuous vocal arousal stimuli. The proposed framework is built upon measures of vocal arousal from acoustically-derived features, and we obtain a reasonable predictive correlation to the true BOLD signal for the seven emotionally-related brain regions. Further experiment also demonstrates that there exists a more explanatory power of using signal-derived arousal measure to the internal BOLD signal responses compared to using human annotated arousal in the construction of Gaussian mixture regression modeling.

Index Terms— behavioral signal processing (BSP), vocal arousal score, fMRI, Gaussian Mixture Regression

1. INTRODUCTION

Emotion is a core mechanism in determining how human produces and perceives behaviors and further reacts and responds to each other during an interaction. Human encodes and decodes important affective information through expressive behavioral modulation (encoding) and internal perceptual process (decoding), respectively. In neuroscience, the use of functional magnetic resonance imaging (fMRI) in measuring blood-oxygen-level-dependent (BOLD) signal, i.e., a proxy measure to human brain's neural activities, have enabled a wealth of research in understanding the scientific underpinning of affective cognitive/perceptual process (e.g., [1, 2, 3, 4, 5]). In engineering, researchers in the past decades have worked extensively on extracting affective-meaningful measures from audio-video recordings, i.e., quantifying expressive attributes related to speech acoustics, facial expressions, and body gestures, enabling systems to perform automatic emotion recognition (e.g., [6, 7, 8, 9]). Our research aims at bridging the two seemingly separate fields to in order to provide a more objective understanding of human's neuro-perceptual mechanism when decoding vocal emotion.

Previous studies in neuroscience have demonstrated that through the use of fMRI technology, different brain regions could be identified as *activated* when a subject is presented with vocal-based affective stimuli. For example, Sander *et al.* identified multiple brain areas in processing vocal emotion, e.g., the right amygdala and bilateral superior temporal sulcus both respond to anger prosody [3], Grandjean *et al.* demonstrates that middle temporal sulcus has enhanced activation for angry relative to neutral prosody [4]. These studies have brought insights into understanding the underlying neuro-perceptual mechanism of vocal emotion decoding. Most of the past studies, though, focused mainly on analyzing which brain regions are activated when presented with a *predefined* set of vocal emotion stimuli, often without objective signal quantification. Furthermore, the dynamic variation of vocal arousal within a stimulus is also largely ignored.

In this paper, we describe our initial work on joint modeling of the dynamics between levels of vocal arousal, i.e., measured with acoustic signal, and human brain's perceptual response, i.e., measured with BOLD signal. On of the recent advancements in behavioral signal processing (BSP) [10] have resulted in the development of an objective measure of emotion arousal, i.e., vocal arousal score (VC-AS) [11]. With the availability of VC-AS as an objective measure of emotional arousal and the BOLD signal as objective measures of neural activities, it enables our research into developing a joint model between these two set of time series using Gaussian Mixture Regression (GMR). To the best of our knowledge, there has not be any directly related work in the literature.

We carry out our experiment in a 18 subjects database where each subject is presented with three distinct levels of 5-minute long vocal arousal stimuli. The results show that our framework is capable of achieving a good prediction of the BOLD signal response in areas such as anterior singular cortex, hippocampus, amygdala, middle temporal pole, and inferior temporal pole. Furthermore, GMR achieves better

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

prediction compared to support vector regression and linear regression. Lastly, by comparing the use of VC-AS to annotator rating as measure of emotional arousal, the use of VC-AS leads to a prediction closer to subject's true BOLD response. This result further implicates the viability of using signal-derived measure of emotional arousal in the study of human affective perceptual process.

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.

2. RESEARCH METHODOLOGY

2.1. Vocal Emotion Stimulus Design

Vocal emotion stimulus is designed from a well-known emotional database, the USC IEMOCAP database [12]. Each of the sentences in the database is annotated with an activation (a.k.a., arousal) and a valence score from the scale of 1 to 5 by at least two naive evaluators. We choose a total of 175 sentences from a single male subject in the IEMOCAP database to construct our stimuli. These sentences are first grouped into three different emotional arousal levels, i.e., *high*, *mid*, and *low*, according to the annotated score; *high* corresponds to activation value greater than 3.5, *low* corresponds to value less than 2.5, and the rest is termed as *mid*. All utterances are restricted to have valence score between 2.5 and 3.5. The stimulation paradigm is a random presentation of three different distinct arousal-level of 5-minute long continuous vocal stimuli. Each of the three stimuli is essentially a collection out of the 175 utterances from the group of *high*, *mid*, and *low*, respectively.

2.2. fMRI Data Collection

We recruited a total of 18 right-handed healthy subjects (14 male, 4 female) to participate in our study. Subjects did not know about the experimental details a-priori. They were only informed that this was an experiment about hearing perception, and they were required to stay awake during MRI scanning while listening to the three vocal stimuli in random order. MRI scanning was conducted on a 3T scanner (Prisma, Siemens, Germany). Anatomical images with spatial resolution of $1 \times 1 \times 1\ mm^3$ (T1-weighted MPRAGE sequence) were acquired using an EPI sequence (TR/TE$=3000/30ms$, voxel size $= 3 \times 3 \times 3\ mm^3$, 40 slices, and 100 repetitions). We performed all necessary pre-processing steps on the collected MRI data using the DPARSF toolbox [13].

2.3. Vocal Stimuli's Arousal Ratings

Figure 1 demonstrates the overall vocal stimuli's design (section 2.1). There are a total of three 5-minute long continuous vocal stimuli. Each of which is associated with two different types of arousal ratings:

- **Global arousal rating (GA)**: a single arousal label, i.e., *high*, *mid*, *low*, for each stimulus

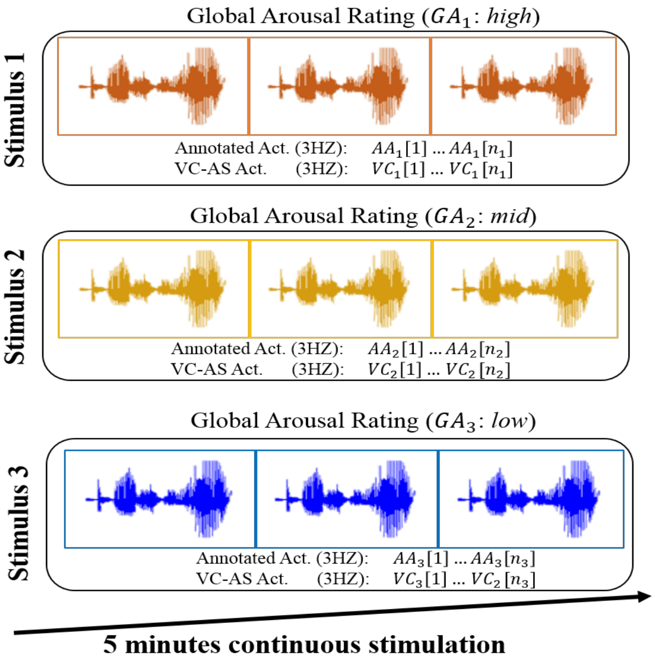
Global Arousal Rating (GA_1: *high*)
Stimulus 1
Annotated Act. (3HZ): $AA_1[1] \dots AA_1[n_1]$
VC-AS Act. (3HZ): $VC_1[1] \dots VC_1[n_1]$
Global Arousal Rating (GA_2: *mid*)
Stimulus 2
Annotated Act. (3HZ): $AA_2[1] \dots AA_2[n_2]$
VC-AS Act. (3HZ): $VC_2[1] \dots VC_2[n_2]$
Global Arousal Rating (GA_3: *low*)
Stimulus 3
Annotated Act. (3HZ): $AA_3[1] \dots AA_3[n_3]$
VC-AS Act. (3HZ): $VC_3[1] \dots VC_2[n_3]$
5 minutes continuous stimulation

Fig. 1: Three continuous 5-minute long vocal affect stimuli were designed. Each stimulus is associated with an global arousal rating (*high*, *mid*, and *low*) and two sequences of local arousal ratings (human annotated arousal, and signal-derived vocal arousal) at every 3 minutes in-sync with MRI scanning

- **Local arousal rating (AA, VC)**: a sequence of arousal ratings occurred at every 3 seconds, i.e., 100 arousal scores aligned with 100 scanning time point from the MRI, for each of stimulus

Global arousal rating is determined based on whether the collection of the sentences that make up that particular stimulus is from the *high*, *mid*, or *low* annotated arousal group (section 2.1). Furthermore, there are two categories of local arousal rating: human annotated activation (AA) and signal-derived VC-AS (VC). Annotated activation is based on human judgment given in the database. VC-AS is a signal-derived arousal score that consists of three simple and knowledge-inspired acoustic features, i.e., pitch, intensity, HF500 (details in [11]); each of the features provides a measure of arousal, and the common single metric used is the fusion of the three scores. Global arousal rating indicates the overall arousal level for the entire stimulus, and local arousal rating indicates the variation of emotional arousal within each stimulus across time.

2.4. Gaussian Mixture Regression

We utilize Gaussian mixture regression approach to jointly model the times series of the arousal ratings and the BOLD signal. The core idea of GMR is to first train a Gaussian mixture model (GMM) to characterize the joint probability density function over a set of variables, and then by utilizing regression function to retrieve desired output variables by specifying the known input variables' values. Researchers have utilized GMR successfully in modeling humanoid robot's motion trajectories [14], performing acoustic-to-articulatory inversion [15], and even tracking continuous-valued emotional attributes in dyadic interactions [16]. We will briefly describe GMR in the following.

Assuming two random variables, X, Y. It is well known that when (X, Y) is a joint Gaussian distribution, the conditional density is also Gaussian. The regression function, $m(x)$, is hence a linear function whose slope is determined by $\sum_X$, the variance of X, $\sum_{YX}$, the covariance of Y and X, and μ_x, μ_y, the mean of X and Y:

$$m(x) = E[Y|X = x] = \mu_y + \Sigma_{YX}\Sigma_X^{-1}(x - \mu_x)$$

$$\sigma^2 = Var[Y|X = x] = \Sigma_Y - \Sigma_Y X \Sigma_X^{-1}\Sigma_X Y$$

the joint density function can then be further partitioned as:

$$\phi(x, y; \mu, \Sigma) = f_{Y|X}(y|x)f_X(x) = \phi(y; m(x), \sigma^2)\phi(x; \mu_x, \Sigma_x)$$

Assuming now the joint density function, f_{XY}, is a mixture of Gaussians with the number of mixture, k, we can easily derive the marginal density of X:

$$f_X(x) = \int f_{XY}(x, y)dy = \sum_{j=1}^{k} \pi_j \phi(x; \mu_{jx}, \Sigma_{jx})$$

and the conditional density, $f_{Y|X}$,

$$f_{Y|X}(y|x) = \frac{\phi(x, y; \mu, \Sigma)}{f_X(x)} = \sum_{j=1}^{k} w_j(x)\phi(y; m_j(x), \sigma_j^2)$$

where,

$$w_j(x) = \frac{\phi(x; \mu_{jx}, \Sigma_{jx})}{\sum_{j=1}^{k} \pi_j \phi(x; \mu_{jx}, \Sigma_{jx})}$$

π_j indicates the mixture weight with the constraint, $\sum_{j=1}^{k} \pi_j$. The final GMR regression function, $m(x)$, is of the form:

$$m(x) = E[Y|X = x] = \sum_{j=1}^{k} w_j(x)m_j(x) \qquad (1)$$

Equation 1 is the core regression function that by specifying $X = x$, we can obtain $Y = m(x)$. Note that the weight function $w_j(x)$ is not determined by the local structure of the data but by the components of the global Gaussian mixture model; this makes GMR a global parametric model with non-parametric flexibility.

3. EXPERIMENTAL SETUP AND RESULTS

3.1. Experimental Setup

We conduct the following two experiments and result their results in the following sections:

- **Experiment I**: Classification of the three different levels of global arousal rating (GA) using voxel-wise BOLD signals
- **Experiment II**: Prediction of the average BOLD signal response for the 20 selected emotion-related brain regions using GMR

Experiment I aims at understanding whether there exists a significant relationship between the changes of BOLD signals in brain regions and human perceptual arousal rating. We adopt the use of anatomical automatic labeling (AAL) to split the brain into 90 regions ($ROI_{\textbf{AAL90}}$) [17], resulting in a total of 47636 number of voxels. Furthermore, we selected 20 emotion-related brain regions-of-interest ($ROI_{\textbf{EMO20}}$) based on a prior research [18], resulting in a total of 11550 number of voxels. Experiment II aims at demonstrating that the predictive ability of GMR to generate the invoked BOLD signal time series when given a sequence of desired VC or AA arousal values. We focus on predicting the average BOLD signal response for the $ROI_{\textbf{EMO20}}$.

The $ROI_{\textbf{EMO20}}$ are the following: the left (L) and the right (R) region of anterior cingulate cortex (ACC), posterior cingulate (PCC), hippocampus (HIPPO), amygdala (AMYG), precuneus (PREC), superior temporal pole (ST_POLE), middle temporal pole (MT_POLE), and inferior temporal (IT).

3.2. Experiment I Results and Discussions

The classification experiment is carried out using leave-one-subject out cross-validation. We train a multi-class linear support vector machine classifier to differentiate between the three classes (levels) of global arousal using either the entire $ROI_{\textbf{AAL90}}$ regions or the $ROI_{\textbf{EMO20}}$ regions.

Table 1: *Summary of Experiment I results*

Classification Accuracies		
	$ROI_{\textbf{AAL90}}$	$ROI_{\textbf{EMO20}}$
PCA	0.50	0.72
Feature Selection	**0.72**	**0.91**

Table 1 summarizes our classification accuracies. The dimensionality of the feature vector for each sample of stimulus is extremely large (100 time points × total # of voxels). Principle component analysis (PCA) approach is performed to reduce the dimensionality to 51 (17 training subjects × 3 stimuli) [19]. We also use feature selection, i.e., univariate selection, to keep only the top 10% of all features.

It is evident there indeed is a strong indication that BOLD signal is distinct when the subject is presented with these three levels of vocal arousal stimuli; by using feature selection within the 20 ROI$_{\textbf{EMO20}}$ areas, we can obtain an accuracy of 91%. We further examine the most important regions within the ROI$_{\textbf{EMO20}}$ by computing a ratio, i.e., the number of voxels selected divided by the total number of voxels within each of the 20 areas. The top three most important areas are AMYG$_R$ (30%), PCC$_L$ (26%), and AMYG$_L$ (25%); the number is the parenthesis indicate the percentage of voxels used in that particular area in the final classification model.

3.3. Experiment II Results and Discussions

In Experiment II, we first train an GMM ($m = 32$) on 29 variables: 1 time index, 20 ROI$_{\textbf{EMO20}}$ average BOLD sig-

Table 2: *Summary of Experiment II results (average Pearson correlation over 18-fold cross-validation)*

ROI	ACC$_L$	ACC$_R$	PCC$_L$	PCC$_R$	HIPPO$_L$	HIPPO$_R$	AMYG$_L$	AMYG$_R$	PREC$_R$	PREC$_L$
LinearR	0.141	0.214	0.210	0.051	0.199	0.196	0.221	0.132	0.151	0.061
SVR	0.165	0.244	0.250	0.022	0.237	0.231	0.265	0.132	0.175	0.084
GMR	0.214	**0.327**	**0.372**	0.041	**0.331**	**0.302**	**0.333**	0.162	0.211	0.101
ROI	ST$_L$	ST$_R$	ST_POLE$_L$	ST_POLE$_R$	MT$_L$	MT$_R$	MT_POLE$_L$	MT_POLE$_R$	IT$_L$	IT$_R$
LinearR	0.018	0.011	0.010	0.030	0.016	0.029	0.227	0.134	0.192	0.205
SVR	0.040	-0.012	0.018	0.040	0.016	0.047	0.268	0.152	0.211	0.228
GMR	0.030	0.020	0.013	0.075	0.005	0.039	**0.381**	0.160	0.283	**0.322**

Table 3: *Comparison of the ROI$_{\textbf{EMO20}}$ BOLD signal prediction accuracy between using vocal arousal score (VC) and annotated activation (AA) as input to the GMR model*

	$GMR_{\textbf{VC}}$	$GMR_{\textbf{AA}}$
ACC$_R$	0.327	0.249
PCC$_L$	0.372	0.292
HIPPO$_L$	0.331	0.251
HIPPO$_R$	0.302	0.239
AMYG$_L$	0.333	0.264
MT_POLE$_L$	0.381	0.280
IT$_R$	0.322	0.263

nals, and 8 VC-AS scores. 20 ROI$_{\textbf{EMO20}}$ BOLD signals are obtained by averaging over voxel values within the specified regions. 8 VC-AS scores are obtained by computing pitch-based, intensity-based, HF500-based, and final fused VC-AS arousal scores individually, and their respective deltas ($4 \times 2 = 8$). VC-AS scores are first convolved with a canonical hemodynamic response function (HRF) to compensate for the delay occurs between receiving stimulus and showing effect in the physical BOLD responses. Then, for the testing subject, we use GMR to obtain 20 sequences of BOLD signals by inputting time index and 8 VC-AS score sequences. The accuracy is measured here by Pearson correlation. Two other baseline methods, i.e., support vector regression (SVR) and linear regression (LinearR), are also carried out to compare with GMR.

Table 2 shows the average correlation between predicted BOLD signal in each of the 20 regions and the true BOLD signal over 18 subjects. First thing to note is that GMR consistently outperforms SVR and LinearR. Secondly, it is encouraging to see that with simply 8 values of vocal arousal scores, the proposed framework is able to achieve good predictive performance (average $r > 0.3$) for a number of brain regions: ACC$_R$, PCC$_L$, AMYG$_L$, HIPPO$_{LR}$, MT_POLE$_L$, and IT$_R$; these areas have all been demonstrated in past works to serve major functions in emotion processing [1, 5, 20]. Figure 2 shows an exemplary predicted versus true BOLD signal for one of our subjects at the region of MT_POLE$_L$.

Furthermore, we conduct additional experiment but building a model between annotated arousal (AA) and ROI$_{\textbf{EMO20}}$ BOLD instead. The comparison between this model and the vocal arousal score-based model for the best 7 areas is shown is Table 3. Our result indicates that acoustically-derived arousal measures obtain a better modeling power.

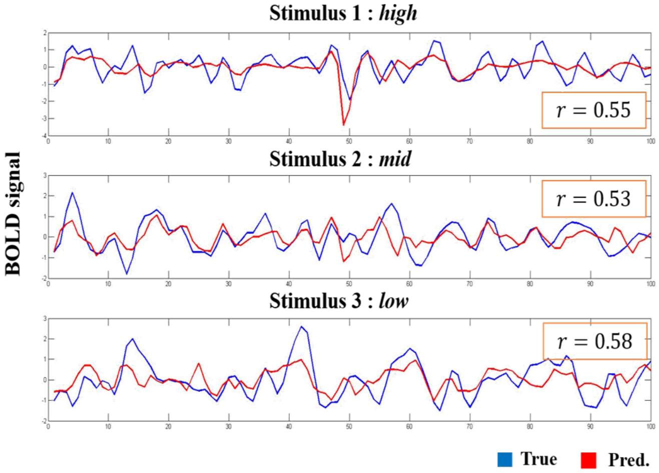

Fig. 2: An example of predicted BOLD signal (in red) versus true invoked BOLD signal (in blue) in the middle temporal pole (left) region using vocal arousal scores in the GMR for the three different stimuli

It may simply due to the fact there are more variables in the use of GMR$_{VC}$ compared to GMR$_{AA}$; however, we would further investigate whether it is indeed that the internal brain's BOLD signal responses are more associated with acoustically-derived signals than with human's annotation.

4. CONCLUSIONS AND FUTURE WORK

In this work, we present a Gaussian mixture regression approach toward objectively model the dynamics between acoustically-derived vocal arousal score and internal brain's BOLD signal response. We demonstrate that the proposed computational framework is capable of tracking the changes of BOLD signal in response to vocal emotion stimuli, measured and quantified acoustically. While initial, it is encouraging to see that it is indeed viable to statistically model the dynamics between expressive vocal behaviors and internal emotion perceptual process, both measured by objective signals.

Uncovering the *transfer function* between exposure to vocal emotion stimuli and the internal brain's response remains a challenging research direction. With the continuing advancement in the signal-based emotion recognition systems and the growing amount of neuroscience research in emotional perception, we will continue to develop modeling such a transfer function with a goal to inspire both new BSP analytics and to bring objective evidence to advance science.

5. REFERENCES

[1] Tony W Buchanan, Kai Lutz, Shahram Mirzazade, Karsten Specht, N Jon Shah, Karl Zilles, and Lutz Jäncke, "Recognition of emotional prosody and verbal components of spoken language: an fmri study," *Cognitive Brain Research*, vol. 9, no. 3, pp. 227–238, 2000.

[2] Patrik Vuilleumier, Jorge L Armony, Jon Driver, and Raymond J Dolan, "Effects of attention and emotion on face processing in the human brain: an event-related fmri study," *Neuron*, vol. 30, no. 3, pp. 829–841, 2001.

[3] David Sander, Didier Grandjean, Gilles Pourtois, Sophie Schwartz, Mohamed L Seghier, Klaus R Scherer, and Patrik Vuilleumier, "Emotion and attention interactions in social cognition: brain regions involved in processing anger prosody," *Neuroimage*, vol. 28, no. 4, pp. 848–858, 2005.

[4] Didier Grandjean, David Sander, Gilles Pourtois, Sophie Schwartz, Mohamed L Seghier, Klaus R Scherer, and Patrik Vuilleumier, "The voices of wrath: brain responses to angry prosody in meaningless speech," *Nature neuroscience*, vol. 8, no. 2, pp. 145–146, 2005.

[5] Ingrid R Olson, Alan Plotzker, and Youssef Ezzyat, "The enigmatic temporal pole: a review of findings on social and emotional processing," *Brain*, vol. 130, no. 7, pp. 1718–1731, 2007.

[6] Chul Min Lee and Shrikanth S Narayanan, "Toward detecting emotions in spoken dialogs," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 2, pp. 293–303, 2005.

[7] Zhihong Zeng, Maja Pantic, Glenn Roisman, Thomas S Huang, et al., "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.

[8] Dimitrios Ververidis and Constantine Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[9] Rafael Calvo, Sidney D'Mello, et al., "Affect detection: An interdisciplinary review of models, methods, and their applications," *Affective Computing, IEEE Transactions on*, vol. 1, no. 1, pp. 18–37, 2010.

[10] Shrikanth Narayanan and Panayiotis 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.

[11] Daniel Bone, Chi-Chun Lee, and Shrikanth Narayanan, "Robust unsupervised arousal rating: A rule-based framework withknowledge-inspired vocal features," *Affective Computing, IEEE Transactions on*, vol. 5, no. 2, pp. 201–213, 2014.

[12] 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. 335–359, 2008.

[13] 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.

[14] Sylvain Calinon, Florent Guenter, and Aude Billard, "On learning, representing, and generalizing a task in a humanoid robot," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 37, no. 2, pp. 286–298, 2007.

[15] Tomoki Toda, Alan W Black, and Keiichi Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[16] Angeliki Metallinou, Athanasios Katsamanis, and Shrikanth Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing*, vol. 31, no. 2, pp. 137–152, 2013.

[17] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Olivier Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot, "Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain," *Neuroimage*, vol. 15, no. 1, pp. 273–289, 2002.

[18] Sascha Frühholz, Wiebke Trost, and Didier Grandjean, "The role of the medial temporal limbic system in processing emotions in voice and music," *Progress in neurobiology*, vol. 123, pp. 1–17, 2014.

[19] Elia Formisano, Federico De Martino, and Giancarlo Valente, "Multivariate analysis of fmri time series: classification and regression of brain responses using machine learning," *Magnetic resonance imaging*, vol. 26, no. 7, pp. 921–934, 2008.

[20] Karine Sergerie, Caroline Chochol, and Jorge L Armony, "The role of the amygdala in emotional processing: a quantitative meta-analysis of functional neuroimaging studies," *Neuroscience & Biobehavioral Reviews*, vol. 32, no. 4, pp. 811–830, 2008.