Generating fMRI-Enriched Acoustic Vectors using a Cross-Modality Adversarial Network for Emotion Recognition

Gao-Yi Chao
Department of Electrical Engineering,
National Tsing Hua University
MOST Joint Research Center for AI
Technology and All Vista Healthcare
Taiwan
ooxx15935720@gmail.com

Chun-Min Chang
Department of Electrical Engineering,
National Tsing Hua University
MOST Joint Research Center for AI
Technology and All Vista Healthcare
Taiwan
cmchang@gapp.nthu.edu.tw

Jeng-Lin Li
Department of Electrical Engineering,
National Tsing Hua University
MOST Joint Research Center for AI
Technology and All Vista Healthcare
Taiwan
cllee@gapp.nthu.edu.tw

Ya-Tse Wu
Department of Electrical Engineering,
National Tsing Hua University
MOST Joint Research Center for AI
Technology and All Vista Healthcare
Taiwan
unreal765@yahoo.com.tw

Chi-Chun Lee
Department of Electrical Engineering,
National Tsing Hua University
MOST Joint Research Center for AI
Technology and All Vista Healthcare
Taiwan
cclee@ee.nthu.edu.tw

ABSTRACT

Automatic emotion recognition has long been developed by concentrating on modeling human expressive behavior. At the same time, neuro-scientific evidences have shown that the varied neuro-responses (i.e., blood oxygen level-dependent (BOLD) signals measured from the functional magnetic resonance imaging (fMRI)) is also a function on the types of emotion perceived. While past research has indicated that fusing acoustic features and fMRI improves the overall speech emotion recognition performance, obtaining fMRI data is not feasible in real world applications. In this work, we propose a cross modality adversarial network that jointly models the bi-directional generative relationship between acoustic features of speech samples and fMRI signals of human percetual responses by leveraging a parallel dataset. We encode the acoustic descriptors of a speech sample using the learned cross modality adversarial network to generate the fMRI-enriched acoustic vectors to be used in the emotion classifier. The generated fMRI-enriched acoustic vector is evaluated not only in the parallel dataset but also in an additional dataset without fMRI scanning. Our proposed framework significantly outperform using acoustic features only in a four-class emotion recognition task for both datasets, and the use of cyclic loss in learning the bi-directional mapping is also demonstrated to be crucial in achieving improved recognition rates.

CCS CONCEPTS

• **Human-centered computing** → *HCI theory, concepts and models*;

KEYWORDS

fMRI, Acoustic Representation, Cross-Modality Adversarial Network, Speech Emotion Recognition

ACM Reference Format:
Gao-Yi Chao, Chun-Min Chang, Jeng-Lin Li, Ya-Tse Wu, and Chi-Chun Lee. 2018. Generating fMRI-Enriched Acoustic Vectors using a Cross-Modality Adversarial Network for Emotion Recognition. In *2018 International Conference on Multimodal Interaction (ICMI '18), October 16–20, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3242969.3242992

1 INTRODUCTION

Emotion governs our behaviors and daily decision making. Integrating robust emotion sensing technologies provides essential analytics of user states and traits in a number of domains geared toward designing human-centered applications. Computing human expressive behaviors using measurable signal recordings, e.g., audio [2], video [6, 10, 37], and physiology[21], to infer an individual's emotion states has been the major focal point of past research effort. In specifics, among these expressive human data streams, speech is one of the most natural form of human communication. Recent advancement in computational models for speech emotion recognition (SER) has been observed in utilizing sophisticated deep learning methods to achieve state-of-art accuracy [18], leveraging cross-corporal databases to improve model robustness [36], and further learning from emotion data in-the-wild to handle real world complications [32]. These algorithmic development in achieving accurate emotion sensing technology using expressive speech cues has made the SER become an integral part of intelligent personalized services, e.g., customer call center[26], human-machine dialog

ICMI '18, October 16–20, 2018, Boulder, CO, USA

ACM ISBN 978-1-4503-5692-3/18/10…$15.00
https://doi.org/10.1145/3242969.3242992

interface [23], and health applications such as depressive disorder assessment [29] and suicidal prevention[13].

Aside from computing human *expressive* acoustics cues for emotion recognition, various development in devices (e.g., electrocardiography (ECG), electroencephalography (EEG), and magnetic resonance imaging (MRI) etc.) has also brought quantitative evidence in aspects of human physiological and neuro-perceptual response to emotion stimuli [7, 12, 21, 39]. The use of fMRI (functional magnetic resonance imaging) technique is especially prevalent in recent years to understand the neuro-*perceptual* mechanism of human brains in response to vocal emotion stimuli using the measured blood-oxygen-level-dependent (BOLD) signal. For example, Johnstone et al. conducted an fMRI study to examine the brain responses to vocal expressions of anger and happiness in order to understand whether specific brain regions would show preferential engagement in the processing of one emotion over the other [20]; Grandjean et al. demonstrated that middle temporal sulcus has an increased activation for angry relative to neutral prosody [16]; Buchanan et al. found that the detection of vocal emotion result in significantly more activity in the right inferior frontal lobe compared to detection of verbal sounds [4].

Recent research has further demonstrated that not only does the measured BOLD signal correlates to the emotional vocal stimuli, through development of deep learning frameworks, the types of emotion categories of the vocal samples can be automatically decoded by modeling the measured fMRI-*perceptual* data directly [19, 34]. In fact, by integrating fMRI signals, which represents how each audio samples being *perceived* by multiple subjects, into the speech-based emotion recognition framework, Wu et al. has shown that these two different modalities (*expressive* acoustic cues and *perceptual* neuro-responses) provide complementary information to each other. However, obtaining fMRI scans on multiple subjects in order incorporate measurable perceptual features to improve speech emotion recognition is not feasible in real world application.

Hence, in this work, we present an encoder framework that can be utilized to derive an fMRI-enriched representation for acoustic inputs through a joint generative model learning with adversarial mechanism using data of parallel cross-modality collection (*expressed* emotional vocal stimuli and *perceptual* neuro-responses). The idea is similar to a previous work carried out by Chen et al. [9], where they utilized Gaussian Mixture Regression (GMR) in order to learn the mapping function between prosody and BOLD signal time series. Our proposed cross modality adversarial network with cyclic loss is capable of learning the complex bi-directional generative relationship between acoustic features and fMRI signals. This particular cross-domain adversarial architecture has also been shown to be useful in applications of cardiac image synthesis [8] and compressed sensing MRI reconstruction [27]. With a learned speech-fMRI cross modality adversarial network, we utilize the speech encoder part of the network, which has effectively incorporated the perspective on humans neuro-perceptual responses, to derive the fMRI-enriched acoustic features for emotion classifier.

We conduct our experiment on two different datasets. The first set is a cross modality parallel dataset. This dataset consists of 18 subjects, where each subject is presented with 251 emotional utterances stimuli designed from the USC IEMOCAP database [5]. Our framework achieve 49.58% in a four-class emotion recognition tasks, which improves 7.99% over using only acoustic feature. Further, our proposed cross modality adversarial network outperforms baseline method of GMR (49.58% vs. 44.3%), and it also achieves the best results among different well-known cross-modality networks. We additionally evaluate the framework on a second dataset with a total of 390 utterances without fMRI data. By using the fMRI-enriched speech encoder learned from the parallel dataset, we derive acoustic features for the Test dataset. We obtain a four-class recognition accuracy of 46.49%, which improves 3.93% over using acoustic features only. The generalization and applicability of our proposed framework in real world application is further strengthened by evaluating on this second dataset.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.

Table 1: *Summary of the original and the merged labels of the Parallel set (251) and the Test set (390) used in this work*

Original	Parallel set		Test set		Merged
Sad	33	33	91	91	Class 1
Happy	12	79	28	54	Class 2
Excited	64		21		
Surprise	3		5		
Neutral	69	69	86	86	Class 3
Angry	19	70	80	159	Class 4
Distress	1		0		
Frustrated	50		79		

2 RESEARCH METHODOLOGY

2.1 Datasets

There are two different datasets used in this work: 1) the Audio-fMRI Parallel Set, and 2) the Audio Test Set. All of our audio emotion samples are derived from the USC IEMOCAP database, which includes multimodal (speech, facial expressions, and lexical content) behavior data. The database has been used widely to develop algorithms for emotion recognition [1, 30, 33]. There are a total of 10 subjects in the database, and 641 sentences from a single male actor in the database are collected to construct the two different datasets (the Parallel set and the Test set) in this work. Each of the sentences in the database is annotated with an emotion label of happiness, anger, sadness, frustration, neutral, disgust, fear, excitement, or surprise, and also dimensional ratings of arousal and valence are given. We will briefly discuss the two different datasets and the fMRI scanning set up in the following.

2.1.1 The Audio-fMRI Parallel Set. The parallel set contains a total of 251 sentences. Six different 5-minute long emotion stimuli were designed from these 251 sentences according to their valence and arousal ratings. In order to collect the audio-fMRI parallel set, i.e., the neuro-perceptual responses to these acoustic emotion stimuli, 18 righted-handed healthy subjects between 20 and 35 years old to participate were recruited. Each participant listened to three 5-minute long continuous vocal emotion stimuli and had a 5-minute break in-between. MRI scanning was conducted on a 3T scanner

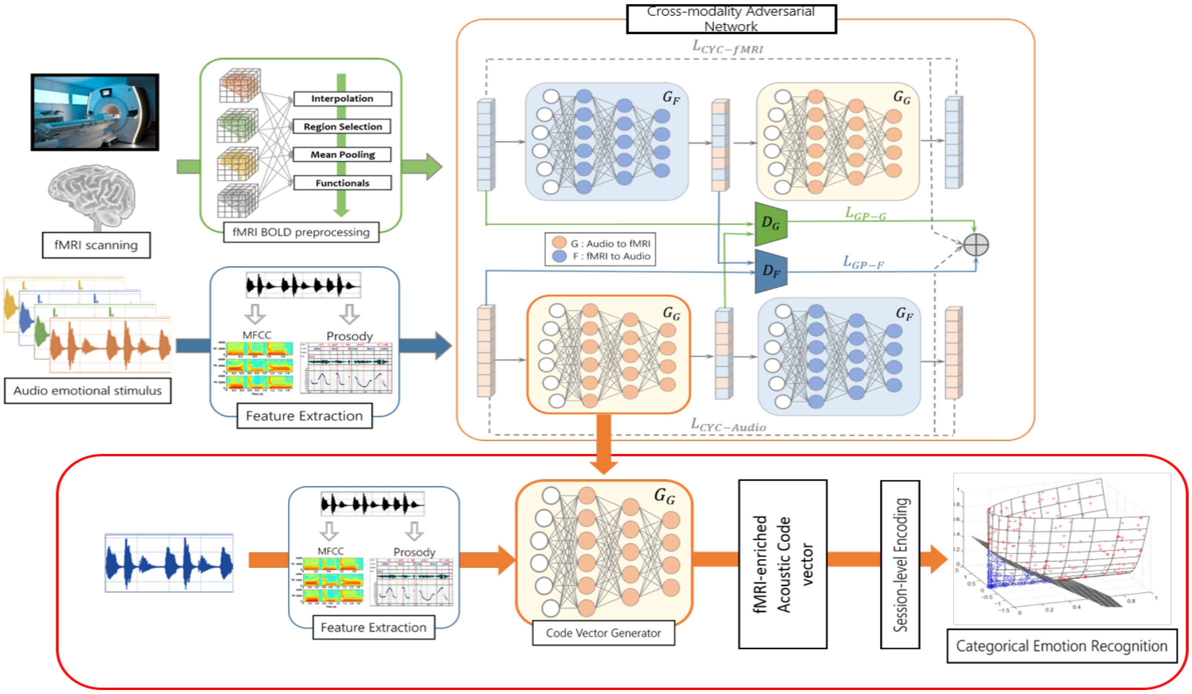

Figure 1: Our cross modality adversarial framework used for emotion recognition can be split into two parts: (upper portion) the first part includes learning a cross modality network, i.e., training $G : X \rightarrow Y$ and $F : Y \rightarrow X$ simultaneously with L_{cyc} (cycle consistency loss) and L_{GP} (adversarial loss); (bottom portion) the second part is to derive fMRI-enriched acoustic vectors from the learned G generator in order to train the final speech emotion recognizer.

(Prisma, Siemens, Germany). Anatomical images with spatial resolution of 1*1*1 mm^3(T1-weighted MPRAGE sequence) were acquired using an EPI sequence ($TR/TE = 3000 = 30ms$, voxel size =$3 * 3 * 3$ mm^3, 40 slices, and 100 repetitions). We performed all necessary pre-processing steps on the collected MRI data using the DPARSF toolbox [35] and additionally performed interpolation to generate a sample at 1 second time step. This parallel dataset is the same dataset used in the previous research [19, 34].

2.1.2 *The Test Set.*

In this work, our aim is to learn a cross-modality network encoder that can incorporate information about perceptual responses to audio emotion stimuli as we encode speech samples through this network. The Audio-fMRI parallel set will be used to derive the cross-modality adversarial network. In order to further test the generalization of our network in deriving fMRI-enriched acoustic vectors, we use 390 unseen sentences from the same male speaker of the IEMOCAP database to be our test set.

2.1.3 *Target Emotion Label.*

The distribution of the original emotion labels provided by the IEMOCAP database on these 641 utterance is spread across eight different classes. According to the valence-activation representation of categorical emotion, we further merge the eight labels into four different classes [28]. Table 1 lists the original and merged labels and their associated number of samples of both the Parallel Set and the Test Set. We use these four emotion classes as our target emotion labels for our experiments.

2.2 Acoustic and fMRI Feature Extraction

In this section, we will describe acoustic and fMRI features used in learning our cross modality adversarial network.

2.2.1 *Acoustic Features.*

We extract 45 low-level acoustic descriptors (LLDs) in total for each audio sample, including 1 pitch, 1 intensity, 13 MFCCs and their associated delta and delta-delta every 10ms using the Praat toolkit [3]. Furthermore, in order to align the audio frame-level features to their corresponding fMRI framerate (1 second), mean pooling is employed on these audio LLDs. The proposed cross modality adversarial network is learned on these 45 acoustic LLDs.

2.2.2 *fMRI Features.*

After preprocessing raw MRI images using the DPARSF, we use the anatomical automatic labeling (AAL) to split the brain into 90 regions resulting in a total of 47636 number of voxels. We further concentrate only on the 20 emotion-related brain regions of interest based on several prior research [14, 31] resulting in a total of 11352 number of voxels. These regions are the left and right of inferior temporal gyrus, middle temporal pole, middle temporal gyrus, superior temporal pole, superior temporal gyrus, precuneus, amygdala, hippocampus, posterior cingulate gyrus, and anterior cingulate gyrus. Within each region, we compute 17 statistical functionals (max, min, mean, median, std, 1 percentile, 99 percentile, 99 percentile - 1 percentile, skewness, kurtosis, min position, max position, 25 percentile, 75 percentile, 75 percentile

- 25 percentile , power, 1st difference) to characterize the region-wise fMRI representation resulting in a feature vector with 340 dimensions (20 × 17) computed at a framerate of 1 second.

2.3 Cross Modality Adversarial Network

Our proposed cross modality adversarial network is shown in Figure 1. The core idea of cross modality adversarial network is to learn the bi-directional mapping functions between two modalities by discovering the common representation space between these heterogeneous data samples. We utilize this network to jointly learn the relationship between acoustic features (section 2.2.1) and fMRI responses (section 2.2.2), denoted as X and Y. Given paired training samples $\{x_i\}_{i=1}^{N}$ where $x_i \in X$ and $\{y_j\}_{j=1}^{N}$ where $y_j \in$ Y. We denote the data distribution as $x \sim p_{data}(x)$ and $y \sim p_{data}(y)$.

2.3.1 Adversarial Loss. Our network includes two mapping functions in a structure of generative adversarial network architecture: one is from acoustic feature space to fMRI representation $G : X \rightarrow Y$ and another one is from fMRI to acoustic $F : Y \rightarrow X$. For the mapping function G and its discriminator D_Y, we can express the network objectives by defining an adversarial loss [15] as:

$$L_{GAN}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{data}(y)}[log(D_Y(y))]$$
$$+\mathbb{E}_{x \sim p_{data}(x)}[log(1 - D_Y(G(x)))]$$

G_G learns from the source domain X in order to generate a code vector $G_G(x)$ that looks similar to the feature vector derived from domain Y, and D_Y aims to ensure that generated samples $G(x)$ to be distribution-ally similar to the real samples y by trying to identify fake samples. G_G iteratively aims to minimize this objective functional against an adversary D_Y that tries to maximize it: $min_G max_{D_Y}(L_{GAN}, G, D_Y, X, y)$. The similar adversarial strategy can be used to to learn the mapping function $F : Y \rightarrow X$ and its discriminator D_X, i.e. $min_G max_{D_X}(L_{GAN}, G, D_X, Y, x)$.

The two mapping functions are jointly optimized in the training process by chaining the parameters of two generators in order to learn the bi-directional mapping relationships simultaneously. To further constrain the space in identifying the possible mapping functions, we use a cycle consistency loss for the cross modality adversarial network, which ensures that these bi-directional functions need to learn a mapping that is cycle consistent [38]. The cycle consistent loss is defined as:

$$L_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)}[\|F(G(x)) - x\|_1]$$
$$+\mathbb{E}_{y \sim p_{data}(y)}[\|G(F(y)) - y\|_1]$$

The network can integrate the cycle loss into its objective function defined as:

$$L(G, F, D_x, D_y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_x, Y, X)$$
$$+\lambda_{cyc} L_{cyc}(G, F)$$

where λ_{cyc} affects the relative importance between the two objectives. In summary, our system involves two set of adversarial networks which are trained simultaneously to minimize the following loss:

$$G^*, F^* = \arg\min_{G,F} \max_{D_X, D_Y} L(G, F, D_X, D_y)$$

2.3.2 Improved Training. In order to further avoid problems of gradient vanish and model collapse, we utilize an improved WGAN [17] objective function with gradient penalty defined as:

$$L_{GP} = \mathbb{E}_{\tilde{x} \sim P_g}[D(\tilde{x})] - \mathbb{E}_{x \sim P_r}[D(x)]$$
$$+\lambda_{gp}\mathbb{E}_{\tilde{x} \sim P_{\tilde{x}}}[(\| \nabla_{\tilde{x}} D(\tilde{x}) - 1\|)^2]$$

where $\tilde{x}$ is sampled uniformly along a straight line between a real and a generated feature. We set all λ_{gp} to be 5 after greed search. Finally, our proposed cross modality adversarial network is optimized using the following complete objective function:

$$L(G, F, D_x, D_y) = L_{GP}(G, D_Y, X, Y)$$
$$+L_{GP}(F, D_x, Y, X) + \lambda L_{cyc}(G, F)$$

2.4 fMRI-Enhanced Emotion Recognition

As Figure 1 demonstrates, our work involves first learning a mapping function using the cross modality adversarial network (section 2.3) learned between paired frame-level acoustic features and and fMRI features in the parallel set, and then we utilize the network to generate fMRI-enriched features for each audio sample to train the final acoustic-based emotion recognizer. The cross modality adversarial network include two generators: acoustic-to-fMRI and fRMI-to-acoustic generators G_G and G_F. These generators can be intuitively thought as encoder networks that map the single modality feature space to a common cross-modality feature space. Hence, by inputting frame-level acoustic low-level features into the generator G_G, we effectively transform the original acoustic feature and obtain an fMRI-enriched acoustic vector.

Our approach in constructing the emotion recognizer is illustrated in the bottom of Figure 1. We first input frame-level acoustic features into the generator G_G to obtain fMRI-enriched acoustic features at every second. Since an emotion label is defined at the utterance level and each audio sample is different in its duration, we conduct session-level encoding using 15 functionals (max, min, mean, median, std, 1 percentile, 99 percentile, 99 percentile - 1 percentile, skewness, kurtosis, min position, max position, 25 percentile, 75 percentile, 75 percentile - 25 percentile) to derive a final feature vector (675 dimensions) as input to support vector machine.

3 EXPERIMENTAL SETUP AND RESULTS

3.1 Experimental Setup

We setup two different four-class emotion recognition experiments in this work. **Experiment I**: the Parallel Set, and **Experiment II**: the Test Set. The evaluation is carried out using 10-fold cross validation scheme, and the evaluation metric is unweighted average recall (UAR). The recall is the ratio $tp/(tp + fn)$ where tp is the number of true positives and fn the number of false negatives.

Linear support vector machine is used as the final classifier. The parameters of the cross modality adversarial network are listed below: the learning rate, λ_{cyc}, and the number of epoch is set to be 0.00005, 25, 268 respectively. All the models are 5-layer DNN architecture, both generators utilize activation function of relu, and discriminators use leaky_relu.

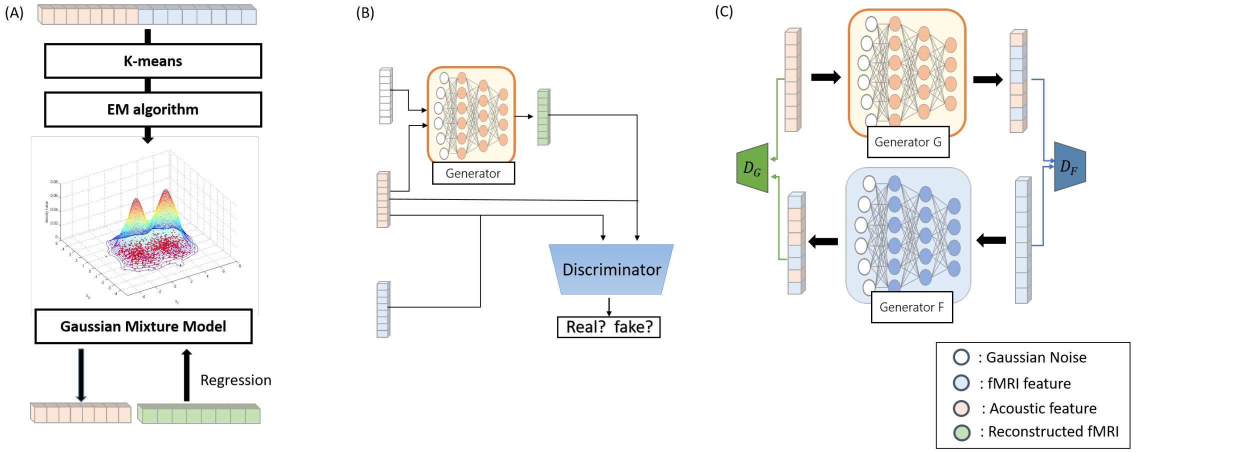

Figure 2: (A) GMR-fMRI(Gaussian Mixture Regression): Using Gaussian Mixture Regression to derive fMRI-enriched acoustic feature. (B) cGAN-fMRI (Conditional GAN): Given acoustic feature to be constrain in generator and discriminator to compute fMRI-enriched acoustic feature). (C) No-Cycle consistence loss (Two pairs of GAN without cyclic mapping): Only chain the two generators without reconstructed mapping and cyclic loss.

3.1.1 ***Experiment I:***. Since this experiment is conducted on a parallel dataset, we also report accuracy obtained by using the original feature set computed from the collected the audio and the fMRI data. The Audio feature set is a 675 dimensional feature vector computed for every audio sample, the fMRI feature set is computed for every utterance by mean pooling on the frame-level 340 dimensional fMRI features (both extraction are described in section 2). We also report accuracy obtained by representing fMRI using principal component analyses on the voxel-wise values (PCA-fMRI), which is a widely employed feature extraction method for MRI [11].

We further compare our proposed cross modality adversarial network with the following methods, which can also be used to learn the relationship between fMRI and audio features in order to generate the fMRI-enriched acoustic vectors:

- Gaussian Mixture Model Regression [25] (GMR-fMRI): Using Gaussian Mixture Regression to generate fMRI-enriched acoustic vectors
- Conditional Generative Adversarial Network [24] (cGAN-fMRI): Using conditional GAN to generate fMRI-enriched acoustic vectors
- Cross Modality Adversarial Netowrk without Cycle-Loss (Code-NoCyc): Using the proposed cross modality adversarial network without the cycle loss to generate fMRI-enriched acoustic vectors
- Cross Modality Adversarial Netowrk with $L2$ Cycle-Loss (Code-l2-fMRI): Using the proposed cross modality adversarial network with L2-norm cycle loss to generate fMRI-enriched acoustic vectors
- Cross Modality Adversarial Netowrk with $L1$ Cycle-Loss and unpaired fMRI-Audio data (Code-unpair): Using the proposed cross modality adversarial network with L1-norm cycle loss by learning on unpaired fMRI-Audio samples to generate fMRI-enriched acoustic vectors
- Cross Modality Adversarial Netowrk with $L1$ Cycle-Loss and paired fMRI-Audio data (Code-l1-fMRI): Using the proposed cross modality adversarial network with L1-norm cycle loss by learning on paired fMRI-Audio samples to generate fMRI-enriched acoustic vectors

A schematic of different cross modality relationship learning algorithms is also depicted in Figure 2.

3.1.2 ***Experiment II:***. The main purpose of this work is to observe our system's generalization on a separate data set (the Test Set) without fMRI scanning, and the emotion distribution between the two datasets are also different. We first learn the cross modality network using the Parallel Set. We generate fMRI-enriched acoustic feature using the the cross modality encoding network (acoustic-to-fMRI), which is then fed into support vector machine.

3.2 Experiment I Results and Discussions

Table 2 summarizes the Experiment I results. The baseline methods, i.e., using the original features only, achieve around 40% UARs in the four-class emotion recognition tasks (chance level is 25%). In specifics, Audio-only feature achieves 41.59%, fMRI-only feature achieves 42.78%, and fMRI-PCA based method obtains 41.79% UARs respectively. By comparing between these original features and the fMRI-enriched acoustic feature sets, we observe, generally, that the fMRI-enriched acoustic features are capable of achieving a higher emotion recognition rates as compared with the original single modality features.

Among different methods in generating the fMRI-enriched acoustic vectors, our proposed method based on cross modality adversarial network achieves the highest recognition rates (49.58%). There are several observations to be made. Firstly, by comparing it with the GMR-based method (44.3%), while GMR-based method outperforms single modality features, the modeling power of GMR is more limited due to its relatively stronger probabilistic parametrization

Table 2: Results of Experiment I: GMR-fMRI: reconstructed feature using Gaussian Mixture Regression; cGAN-fMRI: reconstructed feature using conditional GAN; Code-l1: cross modality adversarial network with $L1$ norm used in the consistence loss; Code-l2: cross modality adversarial network with $L2$ norm used in the consistence loss; Code-NoCyc: cross modality adversarial network without consistence loss; Code-unpair: cross modality adversarial network learning from unpaired data

	Original Features (Audio and fMRI)			fMRI-Enriched Acoustic Features					
	Audio	fMRI	PCA-fMRI	GMR-fMRI	cGAN-fMRI	Code-l1-fMRI	Code-l2-fMRI	Code-unpair	Code-NoCyc
Class 1	30.3	0	6.06	36.36	60.60	66.66	87.87	57.57	51.51
Class 2	46.84	74.68	64.56	54.43	53.16	46.83	31.64	37.97	45.56
Class 3	46.37	52.16	59.42	50.72	34.78	37.68	7.24	42.02	39.13
Class 4	42.58	44.28	37.14	35.71	35.71	47.14	5.71	58.57	44.28
Average	**41.59**	**42.78**	41.79	44.30	46.06	**49.58**	33.12	49.03	45.12

constraint that GMR impose when learning the generative relationship between acoustic stimuli and fMRI responses. The difference in the modeling power of learning complex generative density functions when using adversarial mechanism versus conventional Gaussian-based assumption has also been identified previously [22].

Secondly, our proposed cross modality network learns a bi-directional mapping between acoustic and fMRI. This bi-directional learning is demonstrated to be beneficial by comparing with accuracy obtained by method of cGAN-fMRI (46.06%). The method of cGAN-fMRI essentially learns a one-sided mapping function (acoustic to fMRI) by using real fMRI samples as conditions in the adversarial process. However, we also observe that the appropriate choice on the type of cycle-consistent loss when learning the bi-directional mapping in the adversarial networks is crucial in obtaining the best performances. Using a $L2$-norm in the cycle-consistent loss is detrimental in this context (33.12%), and bi-directional learning without constraint of cycle consistent loss results in an accuracy of only 45.12% potentially due to convergence to less appropriate mapping function. We initially believe this may be due to these 15 common functionals used to encode the fMRI data. The L1 loss tend to be more robust, and some of these functionals are not all robust statistics. The training in GAN also tend to be more sensitive to outlier, this may be the reason why the L2 loss lead to a collapse in training.

Lastly, we also see an interesting effect that while the best accuracy is obtained by learning the cross modality network from the pair acoustic-fMRI data, a similar significant boost in recognition accuracy (49.03%) is also observed when learning the same cross modality network using unpaired data (two domains of features are originally mismatched with time index). This may indicate that the additional modeling power in recognizing emotion may largely come from learning the common representation space between fMRI and acoustic signals, where the actual one-to-one mapping provides only secondary information; however additional analysis is required to understand this phenomenon in detail. In summary, we observe that fMRI-enriched acoustic vectors are useful in improving emotion recognition, and our proposed bi-directional cross modality adversarial network obtains the best recognition accuracy.

3.3 Experiment II Results and Discussions

We further evaluate our framework in the Test Set, where there is no fMRI data available, and the Test Set has a different emotion distribution than the Parallel set (see Table 1). Table 3 summarizes experimental results among different methods. Our proposed method (Code-fMRI) obtains the best performance overall (46.29%), which improves 3.92% over acoustic-only method (42.56%).

We observe that while audio-only recognition accuracy remain similar across the two different datasets (the Parallel Set and the Test Set), the recognition rates obtained from using fMRI-enriched acoustic vectors degrade slightly (49.58% versus 46.29%). It is likely due to the limitation in the size of the Parallel Set, where there is not enough data for each emotion category (e.g., only 33 utterances in Class 1). However, it is still encouraging to see that by transforming the original acoustic features to *fMRI-enriched* acoustic vectors using a cross modality network learned from a different dataset would help improve the emotion recognition of the current dataset.

4 CONCLUSIONS AND FUTURE WORKS

Expressive aspect of acoustic information has been leveraged to develop computational approaches in realizing emotion sensing technologies. Recently, the perceptual responses of human toward affective vocal stimuli can be measured using the MRI technique, and it has been shown that these neuro-perceptual responses can provide complementary information to the acoustic information in terms of improving emotion recognition systems of audio samples. However, obtaining these neuro-perceptual measurements to aid emotion recognition system is infeasible in real world application. In this work, we propose a cross modality adversarial network that learns a bi-directional mapping between these two modalities. Then, by leveraging the learned acoustic to fMRI generator, we can obtain fMRI-enriched acoustic vectors as an enhanced version

Table 3: Results of Experiment II: GMR-fMRI: reconstructed feature using Gaussian Mixture Regression; cGAN-fMRI: reconstructed feature using conditional GAN; Code-fMRI: cross modality adversarial network with $L1$ norm used in the consistence loss; Code-NoCyc: cross modality adversarial network without consistence loss

Test Set	Audio	GMR-fMRI	cGAN-fMRI	Code-fMRI	Code-NoCyc
Class 1	60.43	65.93	64.83	60.43	68.13
Class 2	27.77	29.62	29.62	29.62	25.92
Class 3	34.88	23.25	23.25	26.74	27.90
Class 4	47.16	52.83	58.49	69.18	61.00
Average	**42.56**	42.27	44.05	**46.49**	45.74

of the original acoustic features. We evaluate our framework on two different datasets, i.e., a Parallel Set and a Test Set. Our experiments demonstrate that our proposed network indeed is capable of generating fMRI-enriched acoustic vectors that help improve the recognition rates over using audio-only features.

There are several future directions. One of them is that the current input frame-level audio and fMRI features are hand-crafted descriptors computed using statistical functionals. We will investigate methods toward end-to-end learning while incorporating temporal aspects of these two feature streams as we continue to collect a larger scale of the parallel database and evaluate the generalization of the proposed method across more datasets. Also, while fMRI provides a good spatial resolution of our brain responses, it lacks granular temporal resolution. The affective information encoded in the vocal cues often changes rapidly over time, and while it affects our neuro-perceptual responses, it can not be adequately captured by MRI. We will also explore the use of other brain responses measurements, e.g., electroencephalography, to obtain perceptual responses at a more fine-grained time-scale in order to further enhance our cross modality network. On the other hand, it's interesting topic to test our model on multi-speakers, thus, the expansion of our dataset and the way to solve such issues will be put on our to-do list.

REFERENCES

[1] Mohammed Abdelwahab and Carlos Busso. 2015. Supervised domain adaptation for emotion recognition from speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 5058–5062.

[2] Matthew Black, Athanasios Katsamanis, Chi-Chun Lee, Adam C Lammert, Brian R Baucom, Andrew Christensen, Panayiotis G Georgiou, and Shrikanth S Narayanan. 2010. Automatic classification of married couples' behavior using audio features. In *Eleventh Annual Conference of the International Speech Communication Association*.

[3] P Boersma and D Weenink. 2001. Praat speech processing software. *Institute of Phonetics Sciences of the University of Amsterdam. http://www. praat. org* (2001).

[4] Tony W Buchanan, Kai Lutz, Shahram Mirzazade, Karsten Specht, N Jon Shah, Karl Zilles, and Lutz Jäncke. 2000. Recognition of emotional prosody and verbal components of spoken language: an fMRI study. *Cognitive Brain Research* 9, 3 (2000), 227–238.

[5] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.

[6] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 205–211.

[7] Jing Cai, Guangyuan Liu, and Min Hao. 2009. The research on emotion recognition from ECG signal. In *Information Technology and Computer Science, 2009. ITCS 2009. International Conference on*, Vol. 1. IEEE, 497–500.

[8] Agisilaos Chartsias, Thomas Joyce, Rohan Dharmakumar, and Sotirios A Tsaftaris. 2017. Adversarial Image Synthesis for Unpaired Multi-modal Cardiac Data. In *International Workshop on Simulation and Synthesis in Medical Imaging*. Springer, 3–13.

[9] Hsuan-Yu Chen, Yu-Hsien Liao, Heng-Tai Jan, Li-Wei Kuo, and Chi-Chun Lee. 2016. 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, 5775–5779.

[10] Liyanage C De Silva, Tsutomu Miyasato, and Ryohei Nakatsu. 1997. Facial emotion recognition using multi-modal information. In *Information, Communications and Signal Processing, 1997. ICICS., Proceedings of 1997 International Conference on*, Vol. 1. IEEE, 397–401.

[11] George H Dunteman. 1989. *Principal components analysis*. Number 69. Sage.

[12] Thomas Ethofer, Dimitri Van De Ville, Klaus Scherer, and Patrik Vuilleumier. 2009. Decoding of emotional information in voice-sensitive cortices. *Current Biology* 19, 12 (2009), 1028–1033.

[13] Daniel Joseph France, Richard G Shiavi, Stephen Silverman, Marilyn Silverman, and M Wilkes. 2000. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on Biomedical Engineering* 47, 7 (2000), 829–837.

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

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[16] Didier Grandjean, David Sander, Gilles Pourtois, Sophie Schwartz, Mohamed L Seghier, Klaus R Scherer, and Patrik Vuilleumier. 2005. The voices of wrath: brain responses to angry prosody in meaningless speech. *Nature neuroscience* 8, 2 (2005), 145.

[17] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*. 5769–5779.

[18] Kun Han, Dong Yu, and Ivan Tashev. 2014. Speech emotion recognition using deep neural network and extreme learning machine. In *Fifteenth Annual Conference of the International Speech Communication Association*.

[19] Wan-Ting Hsieh, Hao-Chun Yang, Ya-Tse Wu, Fu-Sheng Tsai, Li-Wei Kuo, and Chi-Chun Lee. 2018. Integrating Perceivers Neural-Perceptual Responses Using A Deep Voting Fusion Network For Automatic Vocal Emotion Decoding. In *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE.

[20] Tom Johnstone, Carien M Van Reekum, Terrence R Oakes, and Richard J Davidson. 2006. The voice of emotion: an FMRI study of neural responses to angry and happy vocal expressions. *Social Cognitive and Affective Neuroscience* 1, 3 (2006), 242–249.

[21] Ma Li, Quek Chai, Teo Kaixiang, Abdul Wahab, and Hüseyin Abut. 2009. EEG emotion recognition system. In *In-vehicle corpus and signal processing for driver behavior*. Springer, 125–135.

[22] Josh Merel, Yuval Tassa, Sriram Srinivasan, Jay Lemmon, Ziyu Wang, Greg Wayne, and Nicolas Heess. 2017. Learning human behaviors from motion capture by adversarial imitation. *arXiv preprint arXiv:1707.02201* (2017).

[23] Wolfgang Minker and Samir Bennacef. 2004. *Speech and human-machine dialog*. Vol. 770. Springer Science & Business Media.

[24] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[25] Bengt Muthén and Kerby Shedden. 1999. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* 55, 2 (1999), 463–469.

[26] Valery Petrushin. 1999. Emotion in speech: Recognition and application to call centers. In *Proceedings of Artificial Neural Networks in Engineering*, Vol. 710.

[27] Tran Minh Quan, Thanh Nguyen-Duc, and Won-Ki Jeong. 2018. Compressed Sensing MRI Reconstruction using a Generative Adversarial Network with a Cyclic Loss. *IEEE Transactions on Medical Imaging* (2018).

[28] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[29] Somchanok Tivatansakul, Michiko Ohkura, Supadchaya Puangpontip, and Tiranee Achalakul. 2014. Emotional healthcare system: Emotion detection by facial expressions using Japanese database. In *Computer Science and Electronic Engineering Conference (CEEC), 2014 6th*. IEEE, 41–46.

[30] Samarth Tripathi and Homayoon Beigi. 2018. Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning. *arXiv preprint arXiv:1804.05788* (2018).

[31] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Olivier Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15, 1 (2002), 273–289.

[32] Valentin Vielzeuf, Stéphane Pateux, and Frédéric Jurie. 2017. Temporal multimodal fusion for video emotion classification in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 569–576.

[33] Martin Wöllmer, Angeliki Metallinou, Florian Eyben, Björn Schuller, and Shrikanth Narayanan. 2010. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *Proc. INTERSPEECH 2010, Makuhari, Japan*. 2362–2365.

[34] Ya-Tse Wu, Hsuan-Yu Chen, Yu-Hsien Liao, Li-Wei Kuo, and Chi-Chun Lee. 2017. Modeling Perceivers Neural-Responses using Lobe-dependent Convolutional Neural Network to Improve Speech Emotion Recognition. *Proc. Interspeech 2017* (2017), 3261–3265.

[35] Chaogan Yan and Yufeng Zang. 2010. DPARSF: a MATLAB toolbox for" pipeline" data analysis of resting-state fMRI. *Frontiers in systems neuroscience* 4 (2010), 13.

[36] Jingwei Yan, Wenming Zheng, Zhen Cui, Chuangao Tang, Tong Zhang, Yuan Zong, and Ning Sun. 2016. Multi-clue fusion for emotion recognition in the wild. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 458–463.

[37] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence* 31, 1 (2009), 39–58.

[38] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. 2016. Learning dense correspondence via 3d-guided cycle consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 117–126.

[39] Tiantong Zhou, Hailing Wang, Ling Zou, Renlai Zhou, and Nong Qian. 2013. 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, 44–51.