Learning an Arousal-Valence Speech Front-End Network using Media Data In-the-Wild for Emotion Recognition

Chih-Chuan Lu
Department of Electrical Engineering
National Tsing Hua University
Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare
Taiwan
lu721@gapp.nthu.edu.tw

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

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

ABSTRACT

Recent progress in speech emotion recognition (SER) technology has benefited from the use of deep learning techniques. However, expensive human annotation and difficulty in emotion database collection make it challenging for rapid deployment of SER across diverse application domains. An *initialization - fine-tuning* strategy help mitigate these technical challenges. In this work, we propose an initialization network that gears toward SER applications by learning the speech front-end network on a large media data collected in-the-wild jointly with proxy arousal-valence labels that are multimodally derived from audio and text information, termed as the Arousal-Valence Speech Front-End Network (AV-SpNET). The AV-SpNET can then be easily stacked simply with the supervised layers for the target emotion corpus of interest. We evaluate our proposed AV-SpNET on tasks of SER for two separate emotion corpora, the USC IEMOCAP and the NNIME database. The AV-SpNET outperforms other initialization techniques and reach the best overall performances requiring only 75% of the in-domain annotated data. We also observe that generally, by using the AV-SpNET as front-end network, it requires as little as 50% of the fine-tuned data to surpass method based on randomly-initialized network with fine-tuning on the complete training set.

KEYWORDS

speech emotion recognition; media data in-the-wild; convolutional neural network; speech front-end network

ACM Reference Format:
Chih-Chuan Lu, Jeng-Lin Li, and Chi-Chun Lee. 2018. Learning an Arousal-Valence Speech Front-End Network using Media Data In-the-Wild for Emotion Recognition. In *2018 Audio/Visual Emotion Challenge and Workshop (AVEC'18), October 22, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3266302.3266306

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
AVEC'18, October 22, 2018, Seoul, Republic of Korea
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-5983-2/18/10…$15.00
https://doi.org/10.1145/3266302.3266306

1 INTRODUCTION

The ability to sense human internal emotion states with measurable data, e.g., speech, video, and physiology, provides key enablers for the next-generation human-machine interface design and offers objective analytic to facilitate experts in decision making across a wide range of application domains, e.g., health [24, 26], education [21, 23, 27], and commerce [15, 32]. Speech emotion recognition (SER) technology has progressed tremendously further due to the fact that speech is one of the most natural form of human daily communication [1]. Recently, SER has also started to adopt the use of deep neural network architecture to achieve further improved recognition performances (e.g., [11, 12, 17, 38]). While the deep learning method is capable of obtaining the state-of-art recognition accuracy, the wide adoption and rapid deployment of SER across applications still hinders by issues of non-scalable emotion annotation collection with inherent high variability in the data collected across diverse domains and situations.

Among successful usage of deep learning, the strategy with the network architecture of '*initialization - fine-tuning*' has been utilized in order to handle these technical issues [33, 39, 40]. *Initialization*, also referred to as pre-training, network is often obtained by learning from a large amount of background corpus in order to derive situational-invariant and robust-cross domain feature representation, and *fine-tuning* is where the network is learned to perform the final *in-domain* recognition tasks. In fact, the generality versus specificity of neurons in deep convolutional neural network (CNN) has been examined to understand the transferability of information for heterogeneous recognition tasks, and it is concluded that the *initialization - fine-tuning* method outperforms random initialization [36]. For example, VGG is a set of well-known pre-trained networks learned from the ImageNet [34]. Various works have demonstrated that by using VGGs as the initialized image network, it can obtain high recognition rates in heterogeneous image recognition tasks, such as prostate cancers recognition using mpMRI scans [6] and thyroid nodules classification using ultrasound images [25]. This particular network architecture not only obtains an improved recognition performance but also reduces human effort in collecting *in-domain* labeled data.

Past works in SER have also utilized a similar strategy in obtaining improved cross-corpora emotion recognition. For example, Fayek et al. used the technique of automatic speech recognition (ASR) to examine the transferability of acoustic features for SER [9].

Table 1: A summary on the key statistics of the three databases used in this work: *Background* database - DaAi, and two *Target* emotion databases: IEMOCAP and NNIME. The label distribution for DaAi is from our derived proxy labels, and the label distribution for the two target emotion databases are the ground truth human annotations.

Database	Total	Total	Time	Label (Arousal)		Label (Valence)		Label (4 Emotion Categories)			
	(utterances)	(segments)	(hrs.)	High	Low	Positive	Negative	Happiness	Anger	Sadness	Neutral
DaAi	20082	62288	14.01	9581	10501	11672	8410	-	-	-	-
IEMOCAP	5538	23360	5.26	2576	2962	2374	3167	788	594	652	897
NNIME	4021	12210	2.75	1950	2071	2151	1870	-	-	-	-

Deng et al. performed feature transfer using sparse autoencoder between emotion corpora [8]; similarly, Huang et al. proposed a PCANet to align feature spaces between emotion corpora [14], and Neumann et al. investigated adaptation technique in transferring the network weights from English to French for cross-lingual SER [30]. While these works have examined cross-corpora SER, most of them focus on transferring between emotion corpus. Most of the existing corpus is often limited in scale making the *initalization* network learns from data potentially with inadequate variability. Furthermore, while abundant media data is easily obtainable nowadays, directly pre-training on this widely-diverse data without proper emotionally-relevant constraint may not benefit the recognition network; for example, Badshah et al. presented a non-successful fine-tuning using pre-trained AlexNet on speech spectrogram for emotion recognition [3].

In this work, instead of learning to transfer between emotion corpora, our aim is to derive a meaningful speech front-end network that can easily be used in different emotion contexts (i.e., languages, scenarios, domains, etc) in order to obtain high recognition accuracy with lesser human annotation effort. In this work, we propose an arousal-valence speech front-end network (AV-SpNET) by learning from a large-scale media data collected in the wild to be the *initialization* network. We further introduce the use of multimodally-derived proxy emotion labels, i.e., based on rule-based prosody information and dictionary-based lexical methods, in assigning arousal and valence labels to these originally unlabeled data. The AV-SpNET is then optimized by learning with a combination of reconstruction and proxy label recognition loss criterion in a CNN speech network architecture. The AV-SpNET once learned is frozen and stacked with recognition networks implemented with fully-connected dense layers within each target emotion corpus in order to perform the final speech emotion recognition.

We evaluate this front-end network on two separate emotion corpora, the USC IEMOCAP database [5] and the NNIME database [7]. The use of AV-SpNET obtains the best emotion recognition accuracy compared to other initialization techniques. More importantly, with the use of AV-SpNET, the recognition network requires lesser amount of annotated data to perform fine-tuning. Specifically, it achieves the best 65.1% in arousal classification for the NNIME database using 75% of fine-tuning data (a 10.8% relative improvement compared to random initialization) and 53.0% in valence classification using also 75% of fine-tuning data (a 7.2% relative improvement compared to random initialization). In the IEMOCAP database, it achieves the best 71.1% in arousal classification using 75% of fine-tuning data (a 7.0% relative improvement compared to random initialization), 52.2% in valence classification, and 43.0% in four-class emotion classification with 75% of fine-tuning data (a 9.4% relative improvement compared to random initialization). We also observe that in most of the recognition tasks, approximately 50% of the fine-tuning labeled data is sufficient to achieve a reliable recognition accuracy when using the proposed AV-SpNET.

The rest of the paper is organized as follows. Methodology and databases are detailed in Section 2. Section 3 includes the experimental setup and results. Conclusion is discussed in Section 4.

2 RESEARCH METHODOLOGY

2.1 Databases

This work includes one *background* database, the DaAi Media Corpus, that is used to train the AV-SpNET, and two *target* emotion databases, the USC IEMOCAP database and the NNIME database, which are used to evaluate the emotion recognition. We will briefly describe each in the following. Table 1 summarizes the key statistics of the database used in this work.

2.1.1 Background: The DaAi Media Corpus. The DaAi Media Corpus is composed of a large collection of Chinese television programs with audio recordings and lexical transcripts. The corpus collects three years worth of TV programs with over thousands of hours of audio data. The genre of the programs included is diverse ranges from education, documentary, healthcare, news, talk shows, interviews, to religion and so on. Each of the program differs in the number of speakers, the environment backgrounds, and the show settings. This large collection of TV programs includes a vast diversity of media data. In this work, we use approximately 14 hours from 20 different programs (20082 utterances) of audio and transcript data. The programs included in this work are usually talk shows, where there is often a main host or narrator per episode. This corpus was not originally collected for emotion recognition; hence, neither specific affective conditions were pre-defined or assumed, nor any prior emotion label existed.

2.1.2 Target: The USC IEMOCAP Database. The USC Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [5] is an well-known English emotion database including five sessions of 10 unique speakers engaging in dyadic interactions. It includes 12 hours of audio-visual data segmented into utterance. Each session has both performances of selected emotional scripts and improvisation of hypothetical emotional scenarios. There are six human evaluators been asked to assess the emotional content and each utterance is labeled by three annotators on categorical emotions

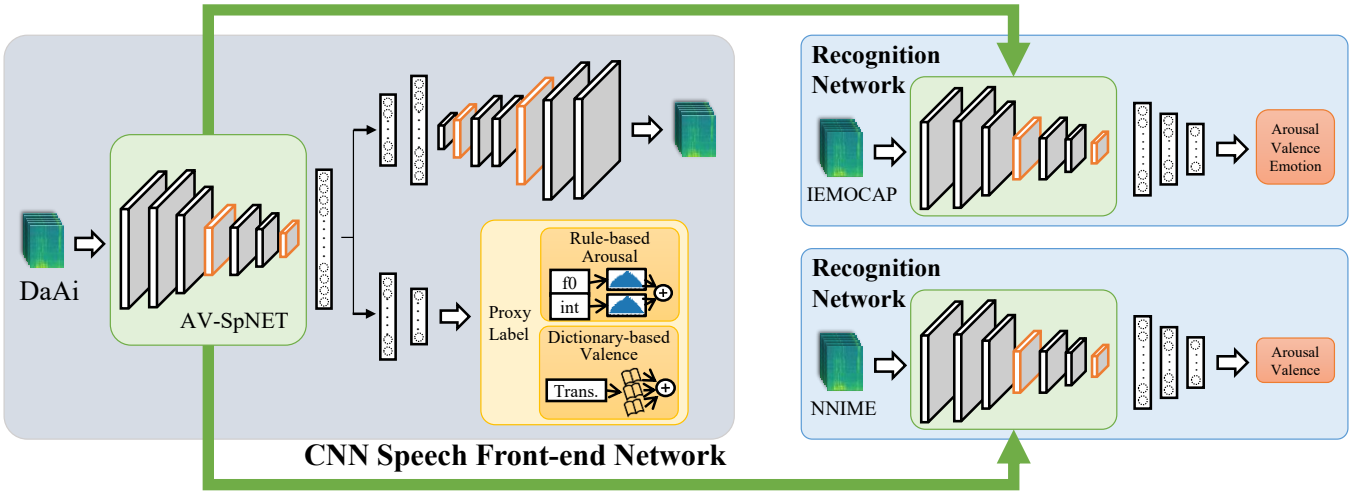


Figure 1: A complete schematic of our initialization - fine-tuning framework for SER. The left shows our proposed network architecture of arousal-valence speech front-end network (AV-SpNET) that is learned from the background DaAi media corpus, and the right shows the recognition network by stacking AV-SpNET with fully connected dense layers to perform the final emotion recognition in the given target emotion databases.

and two annotators on dimensional attributes. The database originally includes five emotion categories: *happiness*, *anger*, *sadness*, *neutral*, and *frustration*, where the raters were free to mark other emotions while annotating. In this work, following previous studies on the same database [10, 28], we focus on the four emotion categories: *happiness* merged with *excitement*, *anger*, *sadness*, and *neutral*. In terms of dimensional annotation, this database includes three primitive attributes: *arousal*, *valence*, and *dominance*; each rating ranges between [1, 5] in step of 0.5. We utilize *arousal* and *valence* attributes in this work, and we further use the average of all utterances for each speaker to binarize the attribute into binary classes: high and low. In this work, we utilize 5538 from 10039 available utterances where each utterance is long enough for further processing and no overlapping talks occur.

2.1.3 Target: The NNIME Database. The NTHU-NTUA Chinese Interactive Multimodal Emotion (NNIME) corpus [7] is a newly-released Chinese emotion corpus designed based on situating the dyadic interactions in daily-life scenario settings. This database consists of 11 hours of audio-video-physiology data with 44 actors grouped in pairs to engage in spontaneous dyadic interactions. Each session is an approximately 3-minute long affective interaction steering along the pre-defined affect atmosphere. In this paper, we take use of a subset from the 3-hour released version which is 4021 out of 6699 utterances that are long enough for further processing. The continuous-in-time *arousal* and *valence* attributes are annotated by four different native raters in a scale ranged from $[-1, 1]$. We average the time-continuous values for each utterance and binarize it into either high or low to be our emotion label of interest.

2.2 Arousal-Valence Speech Front-End Network (AV-SpNET)

We propose to learn an arousal-valence speech front-end network (AV-SpNET) on the background DaAi corpus. Then AV-SpNET is a multi-task structure with a task of CNN autoencoder and an additional task of proxy arousal and valence emotion labels recognition that is learned on inputs of speech spectrogram. The multi-task structure is designed in order to embed emotionally-relevant information in this initialization front-end network. The left side of Figure 1 shows a schematic of our AV-SpNET. We will describe each component in details in the following.

2.2.1 Spectrogram Features. We compute spectrogram as our raw feature input to the network. The audio sampling rate for all of the databases is set at 16kHz. The spectrogram is computed using short-time Fourier transform (STFT) with DFT filter size of 800 *samples* to extract spectrogram of size $400 \times N_f$, where N_f denotes the total number of frames per utterances. Each frame is obtained using 20-ms Hamming window with 10-ms overlap. We also emphasize the low-frequency information by passing it through a low-pass filter with a cutoff frequency set at 4kHz and finally apply logarithm on the spectrum magnitude to obtain our final spectrogram features. Since each utterance varies in their duration, we chunk each utterance into 80 frames a segment: $(20 - 10)\text{ms} \times 80 + 10\text{ms} = 810\text{ms}$ - a duration that has been shown to contain enough emotion information [18]. Each segment shares the same emotion label as an utterance and is considered as a sample in our training. The exact number of utterances and segments for each database used is also listed in Table 1.

2.2.2 Rule-based Arousal Label from Audio. Since there is no emotion annotation in the DaAi media corpus, in order to embed emotion-related information in our AV-SpNET, we need to generate a proxy label for the corpus without any explicit human annotation. Acoustics information has long been shown to contain more information along the dimension of arousal than the lexical modality (e.g., [2, 16]). In fact, a robust unsupervised rule-based arousal indicator has been proposed by Bone et al. [4], which is capable of obtaining reliable emotion arousal index across databases without any human annotation.

In this work, we use the same framework to assign an arousal index for every utterance in the DaAi corpus. Bone et al. presented the use of five different acoustic features in their framework: *pitch*, *vocal intensity*, *HF500*, *speaking rate*, and *jitter*; however, due to the noisy conditions in our DaAi media corpus, we compute only *median pitch* and *median vocal intensity* for each utterance. Then, a baseline model for each speaker of feature type i, $N^{(i)}$, is built for each of the DaAi TV program episodes. The arousal score of each feature type i, denoted as $a_u^{(i)}$ for u^{th} utterance with feature value $x_u^{(i)}$, is given by:

$$a_u^{(i)} = 2 \times E[x_u^{(i)} >= N^{(i)}] - 1 \tag{1}$$

where $E[x_u^{(i)} >= N^{(i)}]$ represents the percentage of utterances which are larger than baseline model, $N^{(i)}$, and the rest are adjustment to bound the arousal score between $[-1, 1]$.

Summing and normalizing score for all feature type i for each utterance results in the final arousal score, a'_u for u^{th} utterance.

$$a'_u = \frac{1}{I} \sum_{i=1}^{I} a_u^{(i)} \tag{2}$$

where I is the number of feature types included ($I = 2$ in this case).

2.2.3 Dictionary-based Valence Label from Text. We further generate a proxy valence label for the DaAi media corpus. Lexical modality is known to capture valence dimension better than the acoustic features [2, 16]; hence, we leverage the availability of lexical transcripts in the corpus to generate the valence proxy label for every utterance used. The method is based on dictionary-based sentiment analysis relying on the assumption that there exists a text sentiment polarity (i.e., positive, negative, or neutral) for every word [31].

Therefore, in order to assign a proxy valence score for each utterance in the DaAi corpus, we utilize three different sentiment dictionaries: NTU Sentiment Dictionary (NTUSD) [20], the Chinese Valence-Arousal Words (CVAW) [37], and the Chinese Linguistic Inquiry and Word Count dictionary (CLIWC) [13]; each includes a list of Chinese words with their corresponding sentiment label (positive: +1 or negative: −1). Thus, for each dictionary i, and j^{th} word of u^{th} sentence denoted as $w_{u,j}$, the valence score can be calculated using the following:

$$v_u = \sum_{i=1}^{I} \sum_{j=1}^{J_u} D_{w_{u,j}}^{(i)} \tag{3}$$

where I is number of dictionaries, J_u is number of words in u^{th} utterance, and $D_{w_{u,j}}^{(i)}$ represents the sentiment value of $w_{u,j}$ given in each dictionary.

We use a similar strategy as in [4], i.e., to build a speaker-wise baseline model in order to derive the valence score in this case. Similar to (1), the valence score v_u can be seen as a feature value and convert to desired score for each utterance in reference to the corresponding speaker baseline model using the following equation (generating a bounded score between $[-1, 1]$):

$$v'_u = 2 \times E[v_u >= N] - 1 \tag{4}$$

where N denotes baseline model constructed by v_u per speaker.

2.2.4 Speech Network in CNN Architecture. The AV-SpNET is an encoder network structure constructed based on CNN architecture. It includes 2-stage convolution and pooling layers. Both stages consist of two convolutional layers with batch normalization and rectified linear units (ReLU) as activation function, and one max-pooling at the end (similar to the ones proposed in [22])

The AV-SpNET is learned using a multi-task structure, i.e., at the decoding there are two different tasks that need to be simultaneously optimized. One of the tasks is learned to perform input reconstruction (auto-encoder) and another task is learned to perform proxy label recognition. A schematic is shown in Figure 1, the upper task is essentially a CNN auto-encoder structure, and the lower task is stacked with fully-connected layers to perform recognition on the proxy labels.

Hence, denoting the input and reconstruction output as x_i, $\tilde{x}_i$, where $i = 1...n$, the hidden layers are learned by minimizing the reconstruction mean-squared loss $\mathcal{L}_m$. For proxy label recognition, we denote proxy label and prediction output as y_i and $\tilde{y}_i$. To approximate $\tilde{y}_i$, we minimize cross-entropy loss $\mathcal{L}_x$. Thus, the total loss used in learning the AV-SpNET is $\mathcal{L}$ in (7).

$$\mathcal{L}_m = \sum_{i=1}^{n} \|x_i - \tilde{x}_i\|^2 \tag{5}$$

$$\mathcal{L}_x = -\sum_{i=1}^{n} y_i \ln \tilde{y}_i + (1 - y_i) \ln (1 - \tilde{y}_i) \tag{6}$$

$$\mathcal{L} = \alpha \mathcal{L}_m + (1 - \alpha) \mathcal{L}_x \tag{7}$$

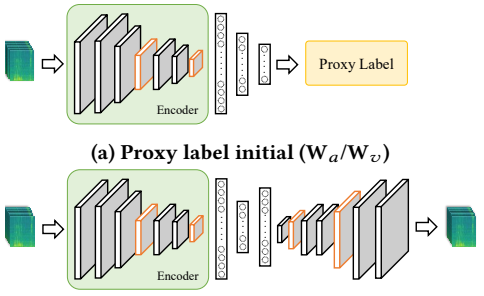
where α is a balancing factor between two losses to prevent undesirable convergence. In this particular structure, the AV-SpNET can be seen as a front-end encoder operated on the speech spectrogram, where the encoder network is learned to embed both the essential latent speech structures (achieved by the use of autoencoder) and the emotion-related information (derived from learning to recognize proxy labels).

2.3 Recognition Network

The AV-SpNET can be easily used as the front-end to perform desired target emotion recognition task (in this case, for the IEMOCAP and the NNIME database). We take the learned AV-SpNET encoder and stack three fully-connected layers with an output layer as our final recognition network. We keep the AV-SpNET weights frozen when fine-tuning on the emotion database of interest. In order to avoid overfitting, dropout is used before the output layer. Lastly, since the training and the recognition both occur at the segment-level, a majority voting scheme is used to derive the final utterance-level emotion labels.

Table 2: Lists of other initialization techniques compared and their abbreviations.

Abbrev.	Description
R	Random initial
$\mathbf{W}_a/\mathbf{W}_v$	Proxy label initial
AE	AutoEncoder initial
$\mathbf{W}_{AE_a}/\mathbf{W}_{AE_v}$	AutoEncoder then proxy label initial
AV-SpNET$_a$/**AV-SpNET**$_v$	AV-SpNET initial



(a) Proxy label initial (W_a/W_v)

(b) AutoEncoder initial (AE)

Figure 2: Illustration of two specific models compared in this work. *"Encoder"* denotes initialization front-end network.

3 EXPERIMENTAL SETUP AND RESULTS

3.1 Experimental Setups

3.1.1 Network Details. The raw spectrogram input is of size 200 × 80. The convolution and the pooling layer has 32 kernels of size 3×3. We pad only the first convolution layer in each stage of encoder to maintain the size of feature maps. After the CNN encoder, feature maps are flattened into 1-D and fed into fully-connected layer of 1024 nodes. Then, for reconstruction task, a fully-connected layer of 256 node acts as the latent layer and the decoding part is reverse image of the encoder structure (a 1024 node fully-connected layer and 2-stage deconvolution layers). At last, one more convolution layer of 1 kernel is use to merge feature maps into one output. For proxy label recognition task in the learning of AV-SpNET and recognition network, after the first 1024-node fully-connected layer, two more layers of nodes 512 and 128 are stacked, and an output layer is used to classify either binary or 4-emotion classes depending on the target task. We utilize ADAM optimizer [19] with an initial learning rate of 0.001 in learning the AV-SpNET, and Stochastic Gradient Descent (SGD) with an initial learning rate of 1e-5 in the target emotion database's fine-tuning phase. The size of mini-batch is 32.

3.1.2 Other Initialization Techniques. We compare the effectiveness of AV-SpNET with other initialization network structures. There are four other different initialization settings. The first setting is the standard random initialization without using any pre-training model. The next one is the standard CNN auto-encoder trained on the DaAi corpus, and another one is a CNN structure learns to recognize the proxy labels (these two can be seen as the separate task in the AV-SpNET; a schematic is depicted in Figure 2). We further compare a two-stage initialization technique, i.e., learning an auto-encoder followed by fine-tuning on the proxy labels. All these methods aim at deriving encoder weight to be used as the frozen *front-end* that are simply stacked with the recognition layers. Table 2 provides a list of these techniques with their abbreviations used in this work.

3.1.3 Evaluation Metrics and Scheme. The metric used in this works is unweighted accuracy (UAR). We used a 5-fold speaker independent cross validation scheme for all of our experiments. Except for fine-tuning on all available training set within each fold, we also evaluate the effective of our framework in reducing the amount of fine-tuned emotion labels required. Specifically, within each fold, we fine-tune using either 25%, 50%, 75%, and 100% of available annotated data and report their results accordingly.

3.2 Results and Discussions

A summary of our emotion recognition accuracy obtained for both the IEMOCAP and the NNIME databases is listed in Table 3. The best accuracy obtained using our proposed front-end speech network (AV-SpNET) for the IEMOCAP database is: 0.711 (arousal), 0.522 (valence), 0.431 (4-emotion classes), and for the NNIME database is: 0.651 (arousal) and 0.530 (valence). There are several observations to note. By comparing **AV-SpNET**$_x$ with techniques of $\mathbf{W}_x$ and **AE**, it is evident that in order to learn an emotionally-relevant speech front-end network from large collection of diverse background media data, the initialization network requires more than just an unsupervised auto-encoder or a simple proxy emotion label learning. Furthermore, our proposed joint optimization framework is also critical as evident in its better performance obtained over separate AE then proxy label learning, $\mathbf{W}_{AE_x}$. In fact, we also observe that most of these other initialization networks obtain an accuracy worse than simply randomly initialize the speech CNN network. The vast discrepancy between background corpus and the emotion target databases may actually result in a negative transfer of information (a similar finding is also indicated in other transfer learning task [35]).

The use of **AV-SpNET**$_x$ not only obtains generally the best accuracy over all methods across these emotion recognition tasks but also requires lesser amount of fine-tuning labeled data on the target database in order to reach its maximum accuracy. In specifics, it achieves the best arousal classification accuracy in the NNIME database using 75% of fine-tuning data and the best valence classification rate using also 75% of fine-tuning data. In the IEMOCAP database, we observe a similar trend, i.e., the best arousal classification happens at using 75% of fine-tuning in-domain annotated data, and 75% of data for the four-class emotion classification. It is exciting to see that our proposed front-end network embeds meaningful emotion-related information that is capable in improving emotion recognition accuracy for the two different emotion databases at the same time requiring less human annotations.

At last, Table 3 also demonstrates an interesting evidence that when using the **AV-SpNET**$_x$ as the speech front-end network, with roughly 50% of fine-tuned *in-domain* data, the recognition accuracy obtained is already similar to method of fine-tuning on all available data with random network weight initialization. The **AV-SpNET**$_x$ not only helps in obtaining improved accuracy, it provides a robust feature representation power as a front-end network. Furthermore, learning **AV-SpNET**$_x$ with arousal proxy label works better for the IEMOCAP but not with the valence proxy labels; while learning with valence proxy label benefits in recognition tasks for the NNIME database. We hypothesize this may due to the fact that the our valence proxy label is derived from the Chinese sentiment dictionary; however, further research into the potential language difference and other suitable emotion proxy labels will be an important direction to investigate. In short, our experimental results demonstrate that our proposed AV-SpNET learned from the media data collected in-the-wild provides a robust and an effective

Table 3: A summary of emotion recognition results reported using unweighted accuracy (UAR) for different initialization networks. The percentages indicate the amount fine-tuning data involved in learning the recognition network (25%, 50%, 75%, or 100%).

The IEMOCAP: Interactive Emotional Dyadic Motion Capture Database												
	Arousal				**Valence**				**4 Emotion Categories**			
	25%	50%	75%	100%	25%	50%	75%	100%	25%	50%	75%	100%
R	0.685	0.696	0.701	0.712	0.505	0.498	0.510	0.519	0.386	0.385	0.407	0.417
$\mathbf{W}_a$	0.665	0.685	0.686	0.696	0.507	0.504	0.505	0.500	0.335	0.329	0.355	0.372
$\mathbf{W}_v$	0.675	0.685	0.686	0.685	0.499	0.504	0.500	0.500	0.308	0.346	0.361	0.374
AE	0.636	0.649	0.660	0.654	0.510	0.502	0.500	0.496	0.285	0.350	0.362	0.361
$\mathbf{W}_{AE_a}$	0.633	0.667	0.678	0.685	0.503	0.501	0.500	0.505	0.311	0.347	0.337	0.349
$\mathbf{W}_{AE_v}$	0.644	0.664	0.657	0.684	0.500	0.497	0.501	0.500	0.312	0.308	0.347	0.349
AV-SpNET$_a$	0.692	0.700	**0.711**	0.711	0.514	0.513	0.509	**0.522**	0.387	0.419	**0.430**	**0.431**
AV-SpNET$_v$	**0.693**	0.699	0.691	0.703	0.513	0.502	**0.509**	0.506	0.375	0.380	0.386	0.410

The NNIME: The NTHU-NTUA Chinese Interactive Multimodal Emotion Corpus								
	Arousal				**Valence**			
	25%	50%	75%	100%	25%	50%	75%	100%
R	0.575	0.630	0.634	0.640	0.503	0.515	0.516	0.516
$\mathbf{W}_a$	0.580	0.603	0.608	0.607	0.505	0.516	0.518	0.492
$\mathbf{W}_v$	0.586	0.601	0.623	0.621	0.498	0.515	0.506	0.503
AE	0.574	0.589	0.608	0.624	0.492	0.516	0.508	0.495
$\mathbf{W}_{AE_a}$	0.585	0.601	0.607	0.612	0.488	0.522	0.500	0.500
$\mathbf{W}_{AE_v}$	0.593	0.602	0.610	0.612	0.487	0.503	0.504	0.501
AV-SpNET$_a$	0.614	0.633	0.634	**0.636**	**0.518**	0.512	0.517	0.516
AV-SpNET$_v$	0.633	0.642	**0.651**	0.632	0.509	0.522	**0.530**	0.523

speech front-end network that can be easily be utilized, i.e., by simply stacking with a couple fully-connected layers as recognition network, to different targeted emotion databases of interest.

4 CONCLUSIONS AND FUTURE WORKS

A rapid deployment of SER technology for a wide range of applications, while important, remains challenging due to the natural difficulties in both obtaining large scale annotated emotion data and also handling the variability in the collected speech data across different domains. In this work, we present a *initialization - fine-tuning* strategy in mitigating these issues. In specifics, we propose to learn an arousal-valence speech front-end network (AV-SpNET) from a large scale unlabeled media data collected in-the-wild. AV-SpNET is learned with a multi-task structure in a CNN architecture, where one of the tasks is to perform input reconstruction and another task is to perform proxy label recognition. The proxy labels are derived multimodally from audio and text data without any human annotation. The learned AV-SpNET is then used as the speech front-end, i.e., no additional weights adaptation needed, in carrying out SER tasks in the target emotion database. We conduct our recognition experiments on two separate and different emotion databases. It demonstrates that the AV-SpNET not only obtains the best recognition accuracy compared to other initialization methods but also require lesser amount of annotated *in-domain* data to achieve the maximum SER recognition rates.

There are multiple future directions. First, the framework in deriving proxy emotion labels from the available media data collected in-the-wild plays an important role in the effectiveness of this speech front-end network. We will immediately investigate emotion recognition accuracy obtained as a function of the differences in language, culture, and other contextual factors existed between the derived proxy labels in the background corpus and the target emotion databases. Further technical development in the front-end CNN network structure to capture supra-segmental information, such as prosody intonation, and also the inclusion of other relevant emotion information in our media corpus will be explored in order to obtain an improved framework. Through continuous advancement in our proposed network, our aim is to provide a publicly-available speech front-end network that is not only robust and reliable in emotion recognition but also can easily be adapted to a wide range of scenarios of human-centered research and applications [29].

REFERENCES

[1] MA Anusuya and Shriniwas K Katti. 2009. Speech Recognition by Machine: A Review. *International journal of computer science and Information Security (IJCSIS)* 6, 3 (December 2009), 181–205.

[2] M. Asgari, G. Kiss, J. van Santen, I. Shafran, and X. Song. 2014. Automatic measurement of affective valence and arousal in speech. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 965–969. https://doi.org/10.1109/ICASSP.2014.6853740

[3] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik. 2017. Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. In *2017 International Conference on Platform Technology and Service (PlatCon)*. 1–5. https://doi.org/10.1109/PlatCon.2017.7883728

[4] Daniel Bone, Chi-Chun Lee, and Shrikanth Narayanan. 2014. Robust Unsupervised Arousal Rating:A Rule-Based Framework withKnowledge-Inspired Vocal Features. *IEEE Transactions on Affective Computing* 5, 2 (April 2014), 201–213. https://doi.org/10.1109/TAFFC.2014.2326393

[5] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation* 42, 4 (dec 2008), 335–359. https://doi.org/10.1007/s10579-008-9076-6

[6] Quan Chen, Xiang Xu, Shiliang Hu, Xiao Li, Qing Zou, and Yunpeng Li. 2017. A transfer learning approach for classification of clinical significant prostate cancers from mpMRI scans. In *Medical Imaging 2017: Computer-Aided Diagnosis*, Vol. 10134. International Society for Optics and Photonics, 101344F.

[7] Huang-Cheng Chou, Wei-Cheng Lin, Lien-Chiang Chang, Chyi-Chang Li, Hsi-Pin Ma, and Chi-Chun Lee. 2017. NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus. In *2017 International Conference on Affective Computing and Intelligent Interaction (ACII)*. 292–298. https://doi.org/10.1109/ACII.2017.8273615

[8] Jun Deng, Zixing Zhang, Erik Marchi, and Bjorn Schuller. 2013. Sparse Autoencoder-Based Feature Transfer Learning for Speech Emotion Recognition. In *2013 International Conference on Affective Computing and Intelligent Interaction (ACII)*. 511–516. https://doi.org/10.1109/ACII.2013.90

[9] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. 2016. On the Correlation and Transferability of Features Between Automatic Speech Recognition and Speech Emotion Recognition.. In *Proceedings of the International Speech Communication Association (Interspeech)*. 3618–3622. https://doi.org/10.21437/Interspeech.2016-868

[10] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. 2017. Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Networks* 92 (2017), 60–68.

[11] Kun Han, Dong Yu, and Ivan Tashev. 2014. Speech emotion recognition using deep neural network and extreme learning machine. In *Proceedings of the International Speech Communication Association (Interspeech)*.

[12] Pavol Harár, Radim Burget, and Malay Kishore Dutta. 2017. Speech emotion recognition with deep learning. In *2017 4th International Conference on Signal Processing and Integrated Networks (SPIN)*. 137–140. https://doi.org/10.1109/SPIN.2017.8049931

[13] Chin-Lan Huang, Cindy K Chung, Natalie Hui, Yi-Cheng Lin, Yi-Tai Seih, Ben CP Lam, Wei-Chuan Chen, Michael H Bond, and James W Pennebaker. 2012. The development of the Chinese linguistic inquiry and word count dictionary. *Chinese Journal of Psychology* (2012).

[14] Zhengwei Huang, Wentao Xue, Qirong Mao, and Yongzhao Zhan. 2017. Unsupervised domain adaptation for speech emotion recognition using PCANet. *Multimedia Tools and Applications* 76, 5 (01 Mar 2017), 6785–6799. https://doi.org/10.1007/s11042-016-3354-x

[15] Nicolae Jascanu, Veronica Jascanu, and Severin Bumbaru. 2008. Toward Emotional E-Commerce: The Customer Agent. In *Knowledge-Based Intelligent Information and Engineering Systems*, Ignac Lovrek, Robert J. Howlett, and Lakhmi C. Jain (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 202–209.

[16] Seliz Gülsen Karadoğan and Jan Larsen. 2012. Combining semantic and acoustic features for valence and arousal recognition in speech. In *Cognitive Information Processing (CIP), 2012 3rd International Workshop on*. IEEE, 1–6.

[17] Jaebok Kim, Gwenn Englebienne, Khiet P Truong, and Vanessa Evers. 2017. Deep Temporal Models using Identity Skip-Connections for Speech Emotion Recognition. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 1006–1013.

[18] Y. Kim and E. M. Provost. 2013. Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 3677–3681. https://doi.org/10.1109/ICASSP.2013.6638344

[19] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[20] Lun-Wei Ku and Hsin-Hsi Chen. 2007. Mining opinions from the Web: Beyond relevance retrieval. *Journal of the American Society for Information Science and Technology* 58, 12 (2007), 1838–1850.

[21] Krithika L.B and Lakshmi Priya GG. 2016. Student Emotion Recognition System (SERS) for e-learning Improvement Based on Learner Concentration Metric. *Procedia Computer Science* 85 (2016), 767 – 776. https://doi.org/10.1016/j.procs.2016.05.264 International Conference on Computational Modelling and Security (CMS 2016).

[22] Wootaek Lim, Daeyoung Jang, and Taejin Lee. 2016. Speech emotion recognition using convolutional and Recurrent Neural Networks. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. 1–4. https://doi.org/10.1109/APSIPA.2016.7820699

[23] Kuan Cheng Lin, Tien-Chi Huang, Jason C. Hung, Neil Y. Yen, and Szu Ju Chen. 2013. Facial emotion recognition towards affective computing-based learning. *Library Hi Tech* 31, 2 (2013), 294–307. https://doi.org/10.1108/07378831311329068

[24] Christine Lisetti and Cynthia LeRouge. 2004. Affective computing in tele-home health. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the.* 8 pp.–. https://doi.org/10.1109/HICSS.2004.1265373

[25] Tianjiao Liu, Shuaining Xie, Jing Yu, Lijuan Niu, and Weidong Sun. 2017. Classification of thyroid nodules in ultrasound images using deep model based transfer learning and hybrid features. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 919–923. https://doi.org/10.1109/ICASSP.2017.7952290

[26] Andrej Luneski, Panagiotis D Bamidis, and Madga Hitoglou-Antoniadou. 2008. Affective computing and medical informatics: state of the art in emotion-aware medical applications. *Studies in health technology and informatics* 136 (2008), 517.

[27] Qi Luo. 2009. Emotion Recognition in Modern Distant Education System by Using Neural Networks and SVM. In *Applied Computing, Computer Science, and Advanced Communication*. Springer, 240–247.

[28] Soroosh Mariooryad and Carlos Busso. 2013. Exploring Cross-Modality Affective Reactions for Audiovisual Emotion Recognition. *IEEE Transactions on Affective Computing* 4, 2 (April 2013), 183–196. https://doi.org/10.1109/T-AFFC.2013.11

[29] Shrikanth Narayanan and Panayiotis G Georgiou. 2013. Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proc. IEEE* 101, 5 (2013), 1203–1233.

[30] Michael Neumann and Ngoc Thang Vu. 2018. Cross-lingual and Multilingual Speech Emotion Recognition on English and French. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[31] Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2, 1–2 (2008), 1–135.

[32] Mohana Shanmugam, Shiwei Sun, Asra Amidi, Farzad Khani, and Fariborz Khani. 2016. The Applications of Social Commerce Constructs. *Int. J. Inf. Manag.* 36, 3 (June 2016), 425–432. https://doi.org/10.1016/j.ijinfomgt.2016.01.007

[33] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging* 35, 5 (2016), 1285–1298.

[34] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[35] Lisa Torrey and Jude Shavlik. 2009. Transfer Learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques: Algorithms, Methods, and Techniques*, Emilio Soria Olivas (Ed.). IGI Global, Chapter 11, 242–264.

[36] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Advances in neural information processing systems*. 3320–3328.

[37] Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K Robert Lai, and Xuejie Zhang. 2016. Building Chinese affective resources in valence-arousal dimensions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 540–545.

[38] W. Q. Zheng, J. S. Yu, and Y. X. Zou. 2015. An experimental study of speech emotion recognition based on deep convolutional neural networks. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. 827–831. https://doi.org/10.1109/ACII.2015.7344669

[39] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*. 487–495.

[40] Xiaodan Zhuang, Arnab Ghoshal, Antti-Veikko Rosti, Matthias Paulik, and Daben Liu. 2017. Improving DNN Bluetooth Narrowband Acoustic Models by Cross-Bandwidth and Cross-Lingual Initialization. In *Proceedings of the International Speech Communication Association (Interspeech)*. 2148–2152. https://doi.org/10.21437/Interspeech.2017-1129