

Speech Emotion Recognition Using Semi-supervised Learning with Ladder Networks

Jian Huang

*National Laboratory of Pattern Recognition, (NLPR)
Institute of Automation, CAS,
School of Artificial Intelligence
University of Chinese Academy of Sciences
Beijing, China
jian.huang@nlpr.ia.ac.cn*

Ya Li

*National Laboratory of Pattern Recognition, (NLPR)
Institute of Automation, CAS,
School of Artificial Intelligence
University of Chinese Academy of Sciences
Beijing, China
yli@nlpr.ia.ac.cn*

Jianhua Tao

*National Laboratory of Pattern Recognition, (NLPR)
Institute of Automation, CAS,
School of Artificial Intelligence
University of Chinese Academy of Sciences,
CAS Center for Excellence in Brain Science and Intelligence Technology,
Institute of Automation, CAS
Beijing, China
jhtao@nlpr.ia.ac.cn*

Zhen Lian

*National Laboratory of Pattern Recognition, (NLPR)
Institute of Automation, CAS,
School of Artificial Intelligence
University of Chinese Academy of Sciences
Beijing, China
lianzheng2016@ia.ac.cn*

Mingyue Niu

*National Laboratory of Pattern Recognition, (NLPR)
Institute of Automation, CAS,
School of Artificial Intelligence
University of Chinese Academy of Sciences
Beijing, China
niumigyue2017@ia.ac.cn*

Jiangyan Yi

*National Laboratory of Pattern Recognition, (NLPR)
Institute of Automation, CAS,
School of Artificial Intelligence
University of Chinese Academy of Sciences
Beijing, China
jiangyan.yi@nlpr.ia.ac.cn*

Abstract—As a major branch of speech processing, speech emotion recognition has drawn much attention of researchers. Prior works have proposed a variety of models and feature sets for training a system. In this paper, we propose to use semi-supervised learning with ladder networks to generate robust feature representation for speech emotion recognition. In our method, the input of ladder network is the normalized static acoustic features and is mapped to high level hidden representations. The model is trained to simultaneously minimize the sum of supervised and unsupervised cost functions by back-propagation. The extracted hidden representations are used as emotional features in SVM model for speech emotion recognition. The experimental results, performed on IEMOCAP database, show 2.6% higher performance than denoising auto-encoder, and 5.3% than the static acoustic features.

Index Terms—*speech emotion recognition, semi-supervised learning, ladder network*

I. INTRODUCTION

Speech emotion recognition (SER) is becoming more and more essential for many applications in natural human-computer interaction [1] which requires a good appreciation of the emotional state for users. It is still a challenging task due to complexity of emotional expressions influenced by the factor

of age, gender and language.

SER system consists of feature extraction step followed by classification step. One of the central research issues in SER is how to extract discriminative, affect-salient features from speech signals [2]. Various spectral and prosodic features have been proposed in the literature [3][4], which are different types of low-level descriptors (LLD) processed by various functionals, such as mean, maximum, minimum, variance etc. However, there is still no consensus on appropriate emotional features and classifiers in SER. State of the art conventional methods mostly differ in their choice of features group and of classifiers type [5]. Moreover, the performance varies greatly in different scenarios and databases.

Meanwhile, SER has greatly benefited from boosting development of deep learning. Stuhlsatz et al. [6] propose a generalized discriminant analysis based on deep neural networks to learn low dimension discriminative features from a large set of acoustic features, which slightly rises the benchmark for emotion recognition. Trigeorgis et al. [7] obtain impressive performance by proposing the deep convolutional neural networks based framework that directly inferred emotional states from the raw speech waveform, instead of from the hand-crafted features. Keren and Schuller [8] utilize convolutional recurrent neural networks to enhance feature extraction from emotional speech data, which shows an performance improvement compared with traditional supervised learning methods.

This work is supported by the National Natural Science Foundation of China (NSFC) (No.61425017, No. 61773379), the National Key Research & Development Plan of China (No. 2017YFB1002804) and the Major Program for the National Social Science Fund of China (13&ZD189).

On the other hand, the success of unsupervised learning mainly contributes to the ability of extracting abstract hierarchical non-linear features of the input [9]. Most of researchers utilize auto-encoders to provide salient representation, leading to notable improvement for speech emotion recognition. Mao et al. [10] combine sparse auto-encoder and convolutional neural network to discover common feature representations in an unsupervised way and then feed them as input to a discriminative classifier. Xia [11] use modified denoising auto-encoders (DAE) to learn more robust features and achieves good performance in SER. In addition, auto-encoders have been highly successful in addressing the distribution mismatch issue across different domains in speech emotion recognition [12][13].

However, there is an underlying problem when using supervised learning or unsupervised learning for SER separately. The unsupervised learning, which is used to extract emotional features, aims to retain all the information that is needed to perfectly reconstruct the input examples. However, maybe we don't need all the information but just emotional relevant information. Whereas, supervised learning preserves only important information that is useful to predict the class label, and drops redundant information which is maybe useful for SER. To address this problem, an emerging area of deep semi-supervised learning has attracted growing interest recently [14][15], which supports to simultaneously deploy both unsupervised and supervised learning.

One of the main attractions to use deep semi-supervised learning for speech emotion recognition is the fact that it facilitates the construction of deep structures which represent many complex functions more concisely than common shallow models. Xue et al. [16] make efforts to disentangle the emotion-specific features from some other factors by employing a semi-supervised feature learning framework. The ladder network [14] is an auto-encoder with skip connections from the encoder and the learning task is similar to that in denoising auto-encoders but applied at every layer, not just the inputs. The skip connections and layer-wise unsupervised targets effectively turn auto-encoders into hierarchical latent variable models which are known to be well suited for semi-supervised learning. In this paper, we propose to use semi-supervised learning with ladder networks to generate robust feature representation for speech emotion recognition.

In the following, Section 2 presents database and acoustic features. Section 3 briefly introduces the proposed method. Section 4 describes experimental results and discussion. Section 5 concludes this paper.

II. DATABASE AND FEATURE SET

A. Database

In this study, we use Interactive Emotional Dyadic Motion Capture (IEMOCAP) [17] to evaluate our proposed method. This corpus has approximately 12 hours of audiovisual data, including video, speech, motion capture of face, and text transcriptions [17]. It has 10 professional actors (5 males and 5 females) acting in two different scenarios: scripted play and spontaneous dialog, in their dyadic interactions. Each

interaction is around 5 minutes in length, and is segmented into sentence levels. Every sentence corresponds a discrete emotion label and a three-dimensional point rating. We use four emotion categories in this study: "angry", "happy", "sad", and "neutral", similar to most prior studies using this corpus [11][18]. Note that we merge "happy" and "excited" in the original annotation into the "happy" class. Only the utterances with majority agreement are used in the experiments. In total we use 5,531 utterances. The class distribution is: 20.0% "angry", 19.6% "sad", 29.6% "happy", and 30.8% "neutral". The experiment protocol for IEMOCAP data is leave-one-speaker-out which means there is no speaker overlap between training and testing set.

B. Acoustic Features

The features used as the input of the ladder network are static acoustic features, which have been successfully applied to emotion recognition task. These features are baseline features of INTERSPEECH 2009 Emotion Challenge [19] extracted by OpenSMILE toolkit [20], whose feature set of 12 functionals applied to 32 acoustic Low-Level Descriptors (LLDs) including their first order delta regression coefficients as shown in Table I. In detail, the 16 LLDs are zero-crossing-rate (ZCR) from the time signal, root mean square (RMS) frame energy, pitch frequency (normalised to 500 Hz), harmonics-to-noise ratio (HNR) by autocorrelation function, and Mel-frequency cepstral coefficient (MFCC) 1–12. Then, 12 functionals – mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, and ranges as well as two linear regression coefficients with their mean square error (MSE) – are applied on the chunk level. Thus, the total feature vector per chunk contains 384 attributes.

TABLE I. FEATURES SETS: 32 LOW-LEVEL DESCRIPTORS (LLD) AND 12 FUNCTIONS.

LLD	Functions
(△) ZCR	Mean
(△) RMS Energy	Standard deviation
(△) F0	Kurtosis, skewness
(△) HNR	Extremes: value, real, position, range
(△) MFCC 1-12	Linear regression: offset, slope, MSE

III. PROPOSED METHOD

In this section, we describe the ladder network architecture, as shown in Fig. 1. In the case of the ladder network, this function is a deep DAE in which noise is injected into all hidden layers. Since all layers are corrupted by noise, another encoder path with shared parameters is responsible for providing the clean reconstruction targets. Through lateral skip connections, each layer of the noisy encoder is connected to its corresponding layer in the decoder. This enables the higher layer features to focus on more abstract and emotional-specific features. Hence, at each layer of the decoder, two signals, one from the layer above and the other from the corresponding layer in the encoder are combined. Formally, the ladder network is defined as follows:

$$\tilde{x}, \tilde{z}^{(1)}, \dots, \tilde{z}^{(L)}, \tilde{y} = \text{Encoder}_{\text{noisy}}(x) \quad (1)$$

$$x, z^{(1)}, \dots, z^{(L)}, y = \text{Encoder}_{\text{clean}}(x) \quad (2)$$

$$x, \hat{z}^{(1)}, \dots, \hat{z}^{(L)}, y = \text{Decoder}(\tilde{z}^{(1)}, \dots, \tilde{z}^{(L)}) \quad (3)$$

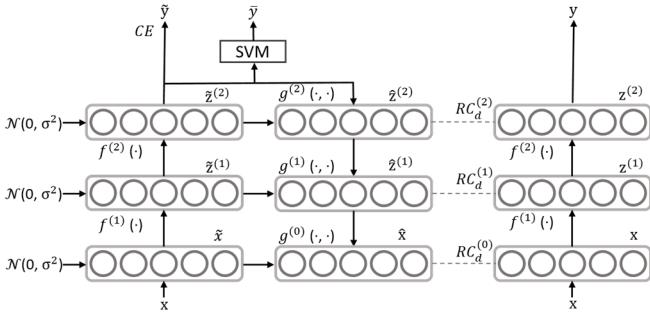


Fig. 1. The ladder network consists of two encoders (on each side of the figure) and one decoder (in the middle). At each layer of both encoders (equations 4 to 8), $z^{(l)}$ and $\tilde{z}^{(l)}$ are computed by applying a linear transformation and normalization on $h^{(l-1)}$ and $\tilde{h}^{(l-1)}$ respectively. The noisy version of the encoder (left) has an extra Gaussian noise injection term. Batch normalization correction ($\gamma^{(l)}, \beta^{(l)}$) and non-linearity are then applied to obtain $h^{(l)}$ and $\tilde{h}^{(l)}$. At each layer of the decoder, two streams of information, the lateral connection $\tilde{z}^{(l)}$ and the vertical connection $u^{(l+1)}$, are required to reconstruct $\hat{z}^{(l)}$ (equations 9 to 13). Acronyms CE and RC stand for Cross Entropy and Reconstruction Cost respectively. The final objective function is a weighted sum of all Reconstruction costs and the Cross Entropy cost.

The variables x , y and \tilde{y} are the input, the noiseless output, and the noisy output respectively. The true target is denoted as y^* . The variables $z^{(l)}$, $\tilde{z}^{(l)}$ and $\hat{z}^{(l)}$ are the hidden representation, its noisy version, and its reconstructed version at layer l .

A. Encoder

In the forward path, individual layers of the encoder are formalized as a linear transformation followed by Batch Normalization and then application of a nonlinear activation function:

$$\tilde{z}_{pre}^{(l)} = W^{(l)} \cdot \tilde{h}^{(l-1)} \quad (4)$$

$$\mu^{(l)} = \text{mean}(\tilde{z}_{pre}^{(l)}) \quad (5)$$

$$\sigma^{(l)} = \text{std}(\tilde{z}_{pre}^{(l)}) \quad (6)$$

$$\tilde{z}^{(l)} = \frac{\tilde{z}_{pre}^{(l)} - \mu^{(l)}}{\sigma^{(l)}} + \mathcal{N}(0, \sigma^2) \quad (7)$$

$$\tilde{h}^{(l)} = \Phi(\gamma^{(l)}(\tilde{z}^{(l)} + \beta^{(l)})) \quad (8)$$

where $\tilde{h}^{(l-1)}$ is the post-activation at layer $l-1$ and $W^{(l)}$ is the weight matrix from layer $l-1$ to layer l . Batch Normalization is applied to the pre-normalization $\tilde{z}_{pre}^{(l)}$ using the mini-batch mean $\mu^{(l)}$ and standard deviation $\sigma^{(l)}$. The next step is to add Gaussian noise with mean 0 and variance σ^2 to compute pre-activation $\tilde{z}^{(l)}$. The parameters $\beta^{(l)}$ and $\gamma^{(l)}$ are responsible for shifting and scaling before applying the nonlinearity $\Phi(\cdot)$. Note that the above equations describe the noisy encoder. If we remove noise $\mathcal{N}(0, \sigma^2)$ and replace \tilde{h} and \tilde{z} with h and z respectively, we will obtain the noiseless version of the encoder.

B. Decoder

At each layer of the decoder in the backward path, the signal from the layer $\tilde{z}^{(l+1)}$ and the noisy signal $\tilde{z}^{(l)}$ are combined into the reconstruction $\hat{z}^{(l)}$ by the following equations:

$$u_{pre}^{(l+1)} = V^{(l)} \cdot \hat{z}^{(l+1)} \quad (9)$$

$$\mu^{(l+1)} = \text{mean}(u_{pre}^{(l+1)}) \quad (10)$$

$$\sigma^{(l+1)} = \text{std}(u_{pre}^{(l+1)}) \quad (11)$$

$$u^{(l+1)} = \frac{u_{pre}^{(l+1)} - \mu^{(l+1)}}{\sigma^{(l+1)}} \quad (12)$$

$$\hat{z}^{(l)} = g(\tilde{z}^{(l)}, u^{(l+1)}) \quad (13)$$

where $V^{(l)}$ is a weight matrix from layer $l+1$ to layer l . We call the function $g(\cdot, \cdot)$ the combinator function as it combines the vertical $u^{(l+1)}$ and the lateral $\tilde{z}^{(l)}$ connections in an element-wise fashion. The original ladder network proposes the following design for $g(\cdot, \cdot)$:

$$g(\tilde{z}^{(l)}, u^{(l+1)}) = b_0 + w_{0z} \odot \tilde{z}^{(l)} + w_{0u} \odot u^{(l+1)} + w_{0zu} \odot \tilde{z}^{(l)} \odot u^{(l+1)} + w_{0\sigma} \odot \text{Sigmoid}(b_1 + w_{1z} \odot \tilde{z}^{(l)} + w_{1u} \odot u^{(l+1)} + w_{1zu} \odot \tilde{z}^{(l)} \odot u^{(l+1)}) \quad (14)$$

Where \odot is an element-wise multiplication operator and each per-element weight is initialized as: $w_{\{0,1\}z}$ and w_σ is 0, $w_{\{0,1\}u}$, $w_{\{0,1\}zu}$ and $b_{\{0,1\}}$ is 0.

C. Objective Function

The objective function is a weighted sum of the supervised cross entropy cost on the top of the encoder and the unsupervised denoising reconstruction costs at each layer of the decoder.

$$\text{Cost} = - \sum_{n=1}^N [\log P(\tilde{y}(n) = y^*(n) | x(n)) + \sum_{l=1}^L \lambda_l \text{ReconsConst}(z^{(l)}(n), \hat{z}^{(l)}(n))] \quad (15)$$

$$\text{ReconsConst}(z^{(l)}, \hat{z}^{(l)}) = \|\frac{\hat{z}^{(l)} - u^{(l)}}{\sigma^{(l)}} - z^{(l)}\|^2 \quad (17)$$

Where $\hat{z}^{(l)}$ is normalized using $u^{(l)}$ and $\sigma^{(l)}$ which are the encoder's sample mean and standard deviation statistics of the current mini batch respectively. The reason for this second normalization is to release the effect of unwanted noise introduced by the limited batch size of Batch Normalization.

Compared to deep DAE, the most obvious addition is the extra reconstruction cost for every hidden layer and the input layer in ladder network. A second important change is lateral

skip connections, that is each layer of the noisy encoder is connected to its corresponding layer in the decoder.

D. Feature Extraction

In our method, we utilize semi-supervised learning with ladder networks to generate robust feature representation. The input of ladder network is the normalized static acoustic features and is mapped to high level hidden representations on the top of the ladder network. The last layer of encoder is attached with SVM model. The extracted hidden representations are used as emotional features in SVM model to get final results \bar{y} for SER.

IV. EXPERIMENTS

A. Experimental Setup

In the experiments, we choose the layer sizes of ladder network to be 384-500-300-100-4. For models with denoising cost multiplier $\lambda^{(l)}$, we optimize them with search grid $\{0.1, 0.2, 0.5, 1, 2, 5\}$. The total search of all $\lambda^{(l)}$ has much larger search space since every layer needs a cost function. Therefore, we optimize it one layer by one layer respectively. We use ADAM optimization algorithm [21] and the dropout rate is 0.5. The batch size is set as 32. The initial learning rate is 0.02 for 50 iterations followed by 25 iterations with a learning rate decaying linearly to 0. We utilize rbf SVM to evaluate the performance of hidden representations extracted from the top layer of ladder network. To determine the parameters of SVM, we use a grid search in the range of $[1.0, 100.0]$ and $[0.0001, 0.1]$ for C and g respectively.

To evaluate the performance of our proposed method, we conduct the experiments using DAE for comparison. The DAE has same structure as the left and middle part with supervised learning part of Fig. 1. The hidden representations extracted from the last layer are used as emotional features to be fed into SVM model. The other experimental setups are similar. Besides, we also utilize SVM to evaluate the performance of static acoustic features as baseline.

B. Experimental Results

In DAE, an auto-encoder is trained to reconstruct the original observation x from a corrupted version \tilde{x} . Learning is based simply on minimizing the norm of the difference of the original x and its reconstruction \hat{x} from the corrupted \tilde{x} , and the cost is $\|\hat{x} - x\|^2$. In ladder network, since the cost function needs both the clean $z^{(l)}$ and corrupted $\tilde{z}^{(l)}$ during training, the encoder is run twice: a clean pass for $z^{(l)}$ and a corrupted pass for $\tilde{z}^{(l)}$. Besides, each layer has a skip connection between the encoder and decoder in ladder network. This feature mimics the inference structure of latent variable models and makes it possible for the higher levels of the network to leave some of the details for lower levels. We conduct the experiments described in previous part, and the experimental results are shown in Table II.

TABLE II. CLASSIFICATION ACCURACY OF DIFFERENT MODELS. THE RESULTS OF THE BRACKETS ARE THE RESULTS OF CORRESPONDING MODELS WITHOUT SUPERVISED LEARNING.

Model	Accuracy
-------	----------

Static acoustic features	0.538
DAE	0.564(0.551)
Ladder network	0.591(0.579)

We can observe from Table II that the accuracy of static acoustic features is 0.538, which is baseline results for other models. The extracted emotional features from DAE achieves nearly 2.6% higher than baseline results. The proposed method achieves best performance 0.591, almost 2.7% higher than DAE, which shows the advantage of ladder network than DAE.

In the following, we will explore the influence resulted from the difference between DAE and ladder network for SER. Firstly, we figure out the effect of supervised learning which is added to DAE and ladder network during training. The experimental results without supervised learning are shown in brackets in Table II, which shows the models with supervised learning achieves better performance than the models without supervised learning. Therefore, the supervised learning part of ladder network contributes to generate more robust features, which can enhance the performance of SER.

Then, we explore the performance from low layers to high layers of DAE and ladder network. The feature representations extracted from different layers are fed into SVM model to evaluate their performance. The experimental results are shown in Fig. 2. The first layer ‘384’ represents the experimental result of static acoustic features and the last but one layer “100” represents the results of Table II. It’s worth noting that the final layer ‘4’ represents the accuracy of supervised learning. We can observe from Fig. 2 that the higher hidden layer is, the better the accuracy is for both DAE and ladder network. However, the ladder network has faster growth than DAE from low layers to high layers. Thus, the high layer of ladder network can provide salient emotional representation, leading to notable improvement for SER. In addition, the accuracy of supervised learning is worse than the models which feed the features extracted from last hidden layer into SVM, showing the superior performance of our proposed method.

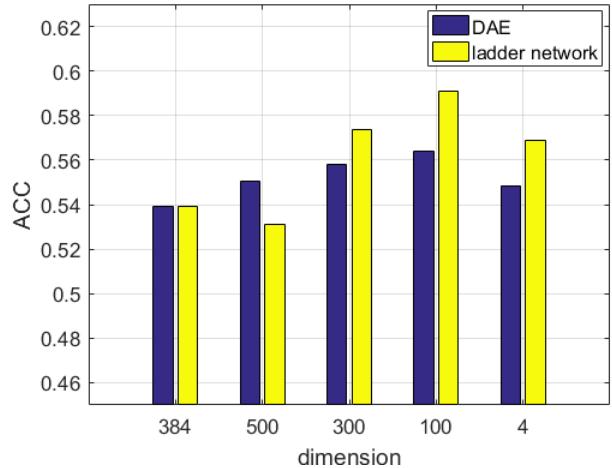


Fig. 2. The Accuracy of different layers for DAE and ladder network.

Having realized the advantages of ladder network over DAE, we explore their influence on specific speech emotion

category. Table III shows each category accuracy of different models corresponding to Table II. When only using static acoustic features, “angry” and “neutral” has better performance than DAE and ladder network. However, DAE and ladder network has more balanced performance than baseline, therefore achieving better performance. Besides, the ladder network achieve better performance than DAE in “angry”, “happy” and “neutral”.

TABLE III. CLASSIFICATION ACCURACY OF EACH EMOTION CATEGORY FOR DIFFERENT MODELS.

Model	Category			
	Angry	Happy	Neutral	Sad
Static acoustic features	0.759	0.303	0.693	0.567
DAE	0.647	0.525	0.534	0.625
Ladder network	0.682	0.554	0.573	0.620

V. CONCLUSIONS

In this paper, we propose to use semi-supervised learning with ladder networks to generate robust feature representation for speech emotion recognition. The input of ladder network is the normalized static acoustic features and is mapped to high level hidden representations. The model is trained to simultaneously minimize weighted sum of the supervised cross entropy cost on the top of the encoder and the unsupervised denoising reconstruction costs at each layer of the decoder. The extracted hidden representations are used as emotional features in SVM model for emotion recognition. The experimental results, performed in IEMOCAP database, show 2.6% higher performance than denoising auto-encoder, and 5.3% than the static acoustic features. Besides, we evaluate the effectiveness of supervised learning during training and higher layer has the ability of generating discriminating features for ladder network. Also, the ladder network can achieve balanced performance than baseline results. In the future, we will explore other semi-supervised methods to extract more discriminative emotional features to improve the performance of SER.

REFERENCES

- [1] C. Vinola, K. Vimaladevi, “A survey on human emotion recognition approaches, databases and applications,” *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, vol. 14, no. 2, pp. 24–44, 2015.
- [2] E. A. Moataz, K. M. S., and K. Fakhri, “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recog.*, vol. 44, no. 3, pp. 572–587, 2011.

- [3] K. S. Rao, S. G. Koolagudi, “Robust emotion recognition using spectral and prosodic features,” Springer Science & Business Media, 2013.
- [4] I. Luengo, E. Navas, and I. Hernandez, “Feature analysis and evaluation for automatic motion identification in speech,” *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 490–501, Oct. 2010.
- [5] M. El Ayadi, M. S. Kamel, F. Karay, “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.
- [6] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, “Deep neural networks for acoustic emotion recognition: raising the benchmarks,” in Proc. ICASSP. Prague, Czech Republic: IEEE, pp. 5688–5691, 2011.
- [7] G. Trigeorgis, F. Ringeval, R. Brueckner, et al, “Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network,” *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, pp. 5200–5204, 2016.
- [8] G. Keren, B. Schuller, “Convolutional RNN: an enhanced model for extracting features from sequential data,” *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, pp. 3412–3419, 2016.
- [9] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural Networks*, vol. 61, pp. 85–117, 2015.
- [10] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, “Learning salient features for speech emotion recognition using convolutional neural networks,” *Multimedia, IEEE Transactions on*, vol. 16, no. 8, pp. 2203–2213, 2014.
- [11] R. Xia, Y. Liu, “Using denoising autoencoder for emotion recognition,” *Interspeech*, pp. 2886–2889, 2013.
- [12] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, “Sparse autoencoder based feature transfer learning for speech emotion recognition,” in Proc. ACII, Geneva, Switzerland, pp. 511–516, 2013.
- [13] J. Deng, Z. Zhang, and B. Schuller, “Linked source and target domain subspace feature transfer learning –Exemplified by speech emotion recognition,” in Proc. ICPR, Stockholm, Sweden, pp. 761–766, 2014.
- [14] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, “Semisupervised learning with ladder networks,” in Proc. NIPS, Montreal, Canada, pp. 3546–3554, 2015.
- [15] J. Weston, F. Ratle, H. Mobahi, and R. Collobert, “Deep learning via semi-supervised embedding,” in *Neural Networks: Tricks of the Trade - Second Edition*, pp. 639–655, 2012.
- [16] W. Xue, Z. Huang, X. Luo, and Q. Mao, “Learning speech emotion features by joint disentangling-discrimination,” in Proc. ACII, Xi'an, China, pp. 374–379, 2015.
- [17] C. Busso, M. Bulut, C. C. Lee, et al, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.
- [18] R. Xia, Y. Liu, “DBN-ivector Framework for Acoustic Emotion Recognition,” *Interspeech 2016*, pp. 480–484, 2016.
- [19] B. Schuller, S. Steidl, and A. Batliner, “The INTERSPEECH 2009 Emotion Challenge,” in Proc. INTERSPEECH, Brighton, UK, pp. 312–315, 2009.
- [20] F. Eyben, M. Wöllmer, and B. Schuller, “openSMILE – the Munich versatile and fast open-source audio feature extractor,” in Proc. MM, Florence, Italy, pp. 1459–1462, 2010.
- [21] D. Kingma, J. Ba, “Adam: A Method for Stochastic Optimization,” *Computer Science*, 2014.