MODELING GENDER INFORMATION FOR EMOTION RECOGNITION USING
DENOISING AUTOENCODER

Rui Xia\textsuperscript{1,2}, Jun Deng\textsuperscript{2}, Björn Schuller\textsuperscript{2,3}, Yang Liu\textsuperscript{1}

\textsuperscript{1} Computer Science Department, The University of Texas at Dallas, USA
\textsuperscript{2} Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Germany
\textsuperscript{3} Department of Computing, Imperial College London, U.K.

ABSTRACT
The Denoising autoencoder (DAE) has been successfully applied
to acoustic emotion recognition lately. In this paper, we adopt the
framework of the modified DAE introduced in [1] that projects the
input signal to two different hidden representations, for neutral
and emotional speech respectively, and uses the emotional representa-
tion for the classification task. We propose to model gender information
for more robust emotional representation in this work. For
neutral representation, male and female dependent DAEs are built
using non-emotional speech with the aim of capturing distinct information
between the two genders. The emotional hidden representation
is shared for the two genders in order to model more emotion
specific characteristics, and is used as features in a back-end clas-
sifier for emotion recognition. We propose different optimization
objectives in training the DAEs. Our experimental results show im-
provement on unweighted accuracy compared with previous work
using the modified DAE method and the classifiers using the stan-
dard static features. Further performance gain can be achieved by
structural level system combination.

Index Terms— Emotion recognition, Denoising autoencoder,
Gender

1. INTRODUCTION
There has been a lot of research efforts lately on identifying paraling-
guistic information in human speech (information beyond words).
Many challenges [2, 3, 4] related to paralinguistic tasks have been
organized and attracted many researchers. Automatic emotion
recognition is one of such paralinguistic tasks (others include vari-
ous speaker states, age, etc.). To accurately detect emotion, front-end
feature extraction and back-end classification are two major parts.
In the front-end, it is important to extract a robust feature representa-
tion which captures emotional cues. Previous work (e.g., [5]) has
shown that static features extracted by applying various functionals
to large amounts of low level descriptors (LLD) can yield competi-
tive performance on emotion recognition tasks. Many studies have
been conducted to investigate complementary features in addition to
these static features, such as gaussian mixture model (GMM) related
features [6, 7], bag-of-word sentiment categories as lexicon features
[8], and facial related features [9]. In the classification stage, stan-
dard classifiers such as support vector machines (SVM) have been
very popular. In addition, ensemble methods have been used to take
advantages of strength of multiple classifiers and shown good re-
sults. For example, in [10], a particle filtering based method is used

\textsuperscript{2} Correspondence should be addressed to {rx, yang}@hlt.utdallas.edu,
{jun.deng, schuller}@tum.de

for fusion of audio, visual and lexicon features. Besides these, prior
studies also investigated transfer learning [11] and active learning
[12, 13] approaches for the emotion recognition task. With growing
interest in deep neural network (DNN) recently, deeper structure by
stacking autoencoders or Restricted Boltzmann Machines (RBM)
has also been successfully used in many fields including emotion
recognition task [14, 15, 16, 17, 18, 19]. In our previous work
[1], we proposed to use the denoising autoencoder (DAE) and its
modified version for emotion recognition. By introducing two hid-
den representations, one used to capture neutral information and
the other for emotional cues, we demonstrated that a more robust
feature representation can be extracted, yielding a performance gain
on emotion recognition.

In this study, we adopt the modified DAE framework as in [1],
but propose to better model the neutral projection in order to con-
sider gender information. In speech recognition, gender dependent
acoustic models are sometimes used in order to model the huge dif-
fERENCE between male and female speech (vocal tract characteris-
tics, pitch, etc.). For emotion recognition, there has been little prior
work on modeling gender information. In [20], gender-dependent
emotion recognizers are trained. In our method, we train female
and male dependent DAEs separately by using their corresponding
non-emotional data. These gender dependent parameters are used in
the modified DAE framework to estimate the emotional projection.
In addition, we propose different cost functions when pre-training
the gender dependent DAEs, which are meant to either capture the
shared information between the two genders or the distinct features
between them. Our experimental results show that our proposed
method has better performance compared with results in previous
work.

2. METHOD
In this section, we first briefly introduce the method in previous work
[1] that uses the denoising autoencoder (DAE) for emotion recogni-
tion. Then, based on this previous approach we propose a method
considering gender variability.

2.1. Previous Work: Modified DAE for Emotion Recognition
The traditional DAE introduced by Vincent et al. [21] aims to learn
a mapping function from an input to a hidden representation, which
can capture the main variation of the data. There are two stages for
training DAE, pre-training and fine-tuning. The unsupervised pre-
training stage iteratively minimizes the loss function between the
original input and the reconstructed input. To make a prediction, a
softmax layer can be added on top of the hidden layer. Given the
predicted labels and the ground truth, a supervised fine-tuning stage is applied to further update parameters. The details of the learning algorithm can be found in [21].

![Diagram](image)

**Fig. 1.** Modified DAE structure with two hidden representations for emotion recognition.

In [1], we proposed a modified DAE framework for emotion recognition, as shown in Figure 1. The major difference between this and the traditional DAE is that in the hidden layer we proposed to project the input into two hidden representations, $y_h$ and $y_e$. $y_h$ is called neutral hidden and designed to capture neutral information that may be contained in all emotional speech. The other one $y_e$ is called emotional hidden, which encodes emotional information.

During training, two parameter sets need to be estimated. The parameter set, $\theta_h(W_n, b_n, b'_n)$, associated with the projection to neutral hidden is pre-learned by a traditional DAE using a large neutral based corpus. The parameter set of emotional hidden representation, $\theta_e(W_e, b_e, b'_e)$, is estimated via pre-training and fine turning. It is initialized with random values and pre-trained based on the following steps:

- **Encoding**: project corrupted input to hidden representations
  
  \[
  y_h = s(W_n x + b_n),
  \]
  \[
  y_e = s(W_e x + b_e).\tag{2}
  \]

- **Decoding**: reconstruct inputs from hidden representations
  
  \[
  z_h = s(W_n^T y_h + b'_n),
  \]
  \[
  z_e = s(W_e^T y_e + b'_e).\tag{4}
  \]

- **Combine**: make new reconstructed input with linearly weighted combination
  
  \[
  z = \alpha \cdot z_e + (1 - \alpha) \cdot z_h
  \]
  \[
  \tag{5}
  \]

- **Learning**: minimize the loss function and update $\theta_e$
  
  \[
  L(z, x) = |z - x|^2.
  \tag{6}
  \]

where $s$ is sigmoid function ($s(x) = (1 + \exp(-x))^{-1}$). The loss function $L(z, x)$ is defined as the squared error between the new reconstructed input and the original input. Stochastic gradient descent algorithm is applied to minimize the cost. Note that here we only update the parameter set $\theta_e$ and fix the other parameter set $\theta_h$.

After pre-training, a softmax layer is added on top of the emotional hidden layer for classification. Parameters in $\theta_e$ are fine-tuned based on the predictions and the corresponding ground truth. For emotion recognition, the emotional hidden representations are used as features with support vector machine (SVM) as the back-end classifier. Experimental results in [1] showed improved emotion recognition performance using this modified DAE method, suggesting the emotional hidden projection in this framework is a better feature representation.

### 2.2. Proposed Method

Based on research in speech recognition, it is expected that there may be significant differences between male and female speech, including emotional speech. Therefore, we investigate whether gender information benefits speech emotion recognition. One way to do this would be using the same framework for male and female speech separately, i.e., build gender dependent neutral and emotion models. However, since the typical emotional speech data sets are rather small, we expect that splitting the data this way will result in too little data for training the emotion models. Therefore in this study, we propose to model gender information by using gender dependent neutral models, but shared emotion models, based on the modified DAE framework described above. Additionally, when training the gender specific neutral models, we propose to consider the relationship between the two genders. The following describes our method in details.

#### 2.2.1. Gender Dependent Neutral DAEs

Assume we have non-emotional speech training utterances for female, $x_f$, and male, $x_m$. From these, the corrupted input sets, $\tilde{x}_f$ and $\tilde{x}_m$, are generated by adding Gaussian noise. To build gender dependent DAEs, instead of simply estimating model parameters using the two sets separately, we propose to consider the relationship between the male and female data. Figure 2 shows our proposed method—we use the female model as an example here to explain the method. Note that this DAE corresponds to the left part of Figure 1, i.e., the neutral model.

![Diagram](image)

**Fig. 2.** Proposed DAE for the female neutral model.

To train the female dependent DAE, we first encode female corrupted instances $\tilde{x}_f$ to the hidden representation and reconstruct it as follows:

\[
  y_f = s(W_n \tilde{x}_f + b_n),
  \tag{7}
  \]

\[
  z_f = s(W_e^T y_f + b'_e).
  \tag{8}
  \]

Here, we use the pre-learned parameter set $\theta_n$ (the same for female and male models) as the initial value instead of training from scratch.
To train the female model parameters, we consider male speech information as well, and define two loss functions \( \text{Loss}_1 \) and \( \text{Loss}_2 \) as follows:

\[
\text{Loss}_1(f_x, f_y) = |f_x - f_y|^2 - \beta \ast KL(p||q),
\]

\[
\text{Loss}_2(f_x, f_y) = |f_x - f_y|^2 + \gamma \ast KL(p||q),
\]

where \( \beta \) and \( \gamma \) are hyper parameters, \( p \) and \( q \) are calculated as follows:

\[
p = \frac{1}{M} \sum_{i=1}^{M} s(W_n x_{f_i} + b_n),
\]

\[
q = \frac{1}{N} \sum_{k=1}^{N} s(W_n x_{m_k} + b_n),
\]

where \( M \) and \( N \) represent the number of instances for female and male sets. With \( p \) and \( q \), \( KL(p||q) \) is defined as:

\[
KL(p||q) = \frac{1}{S} \sum_{i=1}^{S} (p_i \log \frac{p_i}{q_i} + (1 - p_i) \log \frac{1 - p_i}{1 - q_i}),
\]

where \( S \) means the number of the hidden nodes.

There are two components in these loss functions (Equations 9 and 10). The first part is the standard squared loss function – it calculates the reconstruction cost between the original input \( x_i \) and the reconstructed \( f_x \). This is the same as the standard or the modified DAE. The second term in the above loss functions considers information from the other gender. \( p \) and \( q \), as calculated in Equation 11 and 12, represent the average of the hidden representations over the training set for female and male respectively. We introduce a distance function (Equation 13) for these two average representations, based on KL-divergence: it is the average KL-divergence between each of the corresponding hidden nodes for the two genders.

The two loss functions are based on two different considerations: \( \text{Loss}_1 \) (Equation 9) aims to increase \( KL(p||q) \), which means increasing the difference between the two hidden representations, i.e., forcing the models to learn differences between genders. \( \text{Loss}_2 \) (Equation 10) tries to minimize \( KL(p||q) \) in order to make the DAEs to encode some shared information in both genders. The motivation behind these two is based on the assumption that female and male speech may have shared information, as well as distinct gender specific information. Note that the added KL term in the loss function compared to the standard squared loss can also be treated similarly as a regularization term in many optimization problems.

During training, the stochastic gradient descent algorithm is applied to minimize the loss function and update parameters. Since we use batch mode to train the DAE, when training the female models, \( M \) is equal to the number of instances in each minibatch. \( N \) is the total number of instances in the male set \( x_n \). After training, two groups of the estimated parameters based on different loss functions can be obtained. \( \theta_{f(m)l_2}(W_{f(m)l_2}, b_{f(m)l_2}, b_{f(m)l_2}) \) denote female or male parameter sets trained by using \( \text{Loss}_1 \). \( \theta_{f(m)l_2}(W_{f(m)l_2}, b_{f(m)l_2}, b_{f(m)l_2}) \) are for DAEs learned based on \( \text{Loss}_2 \).

2.2.2. Emotional Hidden Representation and Emotion Recognition

For building the emotion model, we use the same modified DAE framework. Different from the previous work, for projection to the neutral hidden representation, we apply gender specific parameter sets (different depending on the loss functions used) on inputs with the known gender label in the training set. Projection parameter sets from input to the emotional hidden are initialized with randomized value and iteratively estimated. Pre-training minimizes the squared loss between the reconstructed \( z \) and the input, and fine-tuning updates parameters for the emotional hidden projection to minimize the emotion classification error.

After pre-training and fine-tuning, emotional hidden representations are used as new features for emotion recognition with standard classifiers. Again, we use the same emotional hidden representation for male and female data. This is meant to capture gender independent but emotion specific information. During testing, we do not need to know the gender label for the test instances because features are generated by only passing instances into the emotional projection.

2.2.3. Combination

In the above, we mentioned using different loss functions will result in different models. Here we propose a combination method on the structure level to combine the different losses. Rather than projecting the input to one neutral hidden representation in the modified DAE framework, we project the input to two neutral hidden projections. The parameter set for each projection is associated with the corresponding loss function, \( \theta_{f(m)l_1} \) and \( \theta_{f(m)l_2} \). Given \( \theta_{f(m)l_1} \) and \( \theta_{f(m)l_2} \), two neutral reconstructed inputs \( z_{l_1} \) and \( z_{l_2} \) can be calculated as follows:

\[
z_{l_1} = s(W_{f(m)l_1} f_x + b_{f(m)l_1})W_{f(m)l_1}^T, \quad (14)
\]

\[
z_{l_2} = s(W_{f(m)l_2} f_x + b_{f(m)l_2})W_{f(m)l_2}^T + b_{f(m)l_2}. \quad (15)
\]

Then, we combine \( z_{l_1} \) and \( z_{l_2} \), linearly with equal weights to obtain the neutral reconstructed input \( z_n \) as:

\[
z_n = 0.5 \ast z_{l_1} + 0.5 \ast z_{l_2}. \quad (16)
\]

Parameter training is similar to above, except now in pre-training we use the combined \( z_n \). Estimation of \( z_n \) is the same as before, via pre-training and fine-tuning.

3. EXPERIMENTS

3.1. Features

We use the static features extracted with openSMILE [22] as the input signal in the DAE framework. There are 1,584 features in total, as used in the INTERSPEECH 2010 Paralinguistic Challenge [23]. Since the feature values have very different ranges, we normalized all the features to the range of 0 to 1 before using them as input to the DAE. Details of the features can be found in [23]. Table 1 summarizes these features.

3.2. Data

The interactive Emotional Dyadic Motion Capture (USC IEMOCAP) database [24] is used in this study. This corpus has approximately 12 hours of audiovisual data, including video, speech, motion capture of face, and text transcriptions. It has 10 professional actors (5 male and 5 female) acting in two different scenarios: scripted play and spontaneous dialog, in their dyadic interaction. Each interaction is around 5 minutes long, and is segmented into sentences. These sentences are labeled by at least 3 annotators. We use four emotion categories in this study: angry, happy, sad and neutral. Note that we merged Happy and Excited in the original annotation into one class: happy. Only the utterances with the majority agreement are used in the experiments. There are 5,531 utterances in this database.
3.3. Experimental Setup

We conduct leave-one-speaker-out cross validation for the emotion recognition experiments. Normalization of features is based on all the training set, instead of a speaker-wise manner. To pre-train gender independent $\theta_n$ in the DAE, we use the Wall Street Journal (WSJ) corpus (about 78K instances) to train with the traditional DAE. The learning rate is set to 0.01 and the number of training epochs is 30. Each minibatch contains 1000 instances. The reason of using WSJ corpus here, rather than the IEMOCAP data, is because we want to train a general DAE to represent neutral speech. Then, to train the gender dependent DAE parameter sets $\theta_{(m|f|)l_1}$ and $\theta_{(m|f|)l_2}$, we use 10 iterations with 0.01 as the learning rate. This is done using the training set from the IEMOCAP data. The hyper parameters $\beta$ and $\gamma$ are set as 0.01 and 0.1 respectively. After that, we use $\theta_{(m|f|)l_3}$ and $\theta_{(m|n|)l_4}$ as parameters of the neutral projections in the modified DAE method. In the pre-training stage, the number of training iterations is set to 20 and the learning rate is 0.01. In the fine-tuning stage, we use 12 iterations with 0.05 as the learning rate. The weight combination parameter $\alpha$ is 0.7, which means the emotional reconstruction has more weight than the neutral one. For all the DAE models, a corruption level of 0.1 is used to obtain the corrupted signal from the original features. The number of hidden nodes is 800, the same as that used in previous work [1]. SVMs with radial basis function (RBF) kernels are used as the classifier for the new features from the hidden representation in the proposed DAE method.

3.4. Experiment Results

Table 2 shows the emotion classification results based on the unweighted average recall (UAR), a metric that has been used as the standard measurement in the INTERSPEECH Emotion Challenges. This is the average of the results for each emotion category. For a comparison, the first two results in 2 are from systems using the original static features and emotional hidden projection as features in the previous modified DAE method [1]. The following rows show the results with features extracted by using the gender dependent neutral projection learned based on different loss functions, and system combination. The last row shows the result using the new gender dependent neutral projection as in our proposed method; however, for the loss function, we do not include the KL cost, and thus it is just a standard squared loss for each gender separately, without considering their relationship during DAE training.

From Table 2 we can see that features extracted based on our proposed method can yield better results on UAR, 1.5% and 1.2% improved, using the two loss functions respectively, compared to the previous modified DAE method. System combination on the structure level yields further gain. Our method has a significant improvement compared with our previous work (p-value < 0.05 with one tailed z-test) and passes the significance level of 0.01 compared with the baseline static features. When KL cost is not used in the loss function, there is a performance degradation compared to when it is used in our proposed method, indicating the effectiveness of modeling the relationship between the two genders. Table 3 shows the accuracy for each emotion class. We notice that there is more improvement for the 'neutral' and 'sad' classes, and less for 'angry' and 'happy'. This suggests we may need to build models taking into account the valence or arousal dimension. Finally, we extended the above gender dependent framework to speaker dependent ones. Using the same data set, our experimental results show similar performance as when using gender dependent neutral models. This might be because each speaker has very little data, limiting the potential advantage of building speaker dependent models. We will continue to investigate this in the future work.

Table 2. Emotion classification results (in %).

<table>
<thead>
<tr>
<th>System</th>
<th>UAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static features</td>
<td>59.7</td>
</tr>
<tr>
<td>Previous modified DAE [1]</td>
<td>61.4</td>
</tr>
<tr>
<td>New DAE</td>
<td></td>
</tr>
<tr>
<td>$\text{Loss}_1$</td>
<td>62.9</td>
</tr>
<tr>
<td>$\text{Loss}_2$</td>
<td>62.6</td>
</tr>
<tr>
<td>System combination</td>
<td>63.1</td>
</tr>
<tr>
<td>No KL in Loss</td>
<td>62.2</td>
</tr>
</tbody>
</table>

Table 3. Accuracy in % for each emotion category.

<table>
<thead>
<tr>
<th>System</th>
<th>angry</th>
<th>happy</th>
<th>neutral</th>
<th>sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static features</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$66.0$</td>
<td>$52.3$</td>
<td>$53.0$</td>
<td>$67.5$</td>
<td></td>
</tr>
<tr>
<td>Previous modified DAE [1]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$68.3$</td>
<td>$58.2$</td>
<td>$54.0$</td>
<td>$65.2$</td>
<td></td>
</tr>
<tr>
<td>New DAE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{Loss}_1$</td>
<td>$69.4$</td>
<td>$58.4$</td>
<td>$56.3$</td>
<td>$67.3$</td>
</tr>
<tr>
<td>$\text{Loss}_2$</td>
<td>$68.8$</td>
<td>$57.8$</td>
<td>$56.5$</td>
<td>$67.0$</td>
</tr>
<tr>
<td>System combination</td>
<td>$69.3$</td>
<td>$58.8$</td>
<td>$56.8$</td>
<td>$67.3$</td>
</tr>
</tbody>
</table>

4. CONCLUSION

In this paper, we proposed to consider gender information to model neutral projection and better capture emotion specific features in the DAE framework. When training gender dependent models, we use KL-divergence between the hidden representations of the male and female speakers as the additional cost in the objective function to measure correlation between the two genders. Emotional projection is trained under the modified DAE framework with gender dependent DAEs used for neutral projection. Our experiments show that using the new emotional projection as features yielded better system performance, suggesting the benefit of modeling gender variability for emotion recognition.

5. ACKNOWLEDGMENT

This work is supported by U.S. Air Force Award FA9550-10-1-0388, NSF award 1225629, and DARPA contract FA8750-13-2-0041. Any opinions expressed in this work are those of the authors and do not necessarily reflect the views of the funding agencies.
6. REFERENCES


[6] Tauhidur Rahman, Sorooosh Mariooryad, Shalini Keshava-


