Preserving actual dynamic trend of emotion in dimensional speech emotion recognition

Wenjing Han, Haifeng Li, Florian Eyben, Lin Ma, Jiayin Sun, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Han, Wenjing, Haifeng Li, Florian Eyben, Lin Ma, Jiayin Sun, and Björn Schuller. 2012. "Preserving actual dynamic trend of emotion in dimensional speech emotion recognition." In *Proceedings of the 14th ACM International Conference on Multimodal Interaction - ICMI '12, October 2012, Santa Monica, CA, USA*, edited by Louis-Philippe Morency, Dan Bohus, Hamid Aghajan, Justine Cassell, Anton Nijholt, and Julien Epps, 523–28. New York, NY: ACM Press. https://doi.org/10.1145/2388676.2388786.

Preserving Actual Dynamic Trend of Emotion in Dimensional Speech Emotion Recognition

Wenjing Han[*]
School of Computer Science
and Technology
Harbin Institute of Technology
150001 Harbin, China
wenjing.han@tum.de

Haifeng Li
School of Computer Science
and Technology
Harbin Institute of Technology
150001 Harbin, China
lihaifeng@hit.edu.cn

Florian Eyben
Institute for Human-Machine
Communication
Technische Universität
München
80290 München, Germany
eyben@tum.de

Lin Ma
School of Computer Science
and Technology
Harbin Institute of Technology
150001 Harbin, China
malin_li@hit.edu.cn

Jiayin Sun
School of Humanities and
Social Sciences
Harbin Institute of Technology
150001 Harbin, China
sunjiayin@hit.edu.cn

Björn Schuller
Institute for Human-Machine
Communication
Technische Universität
München
80290 München, Germany
schuller@tum.de

ABSTRACT

In this paper, we use the concept of *dynamic trend of emotion* to describe how a human's emotion changes over time, which is believed to be important for understanding one's stance toward current topic in interactions. However, the importance of this concept - to our best knowledge - has not been paid enough attention before in the field of speech emotion recognition (SER). Inspired by this, this paper aims to evoke researchers' attention on this concept and makes a primary effort on the research of predicting correct dynamic trend of emotion in the process of SER. Specifically, we propose a novel algorithm named Order Preserving Network (OPNet) to this end. First, as the key issue for OPNet construction, we propose employing a probabilistic method to define an emotion trend-sensitive loss function. Then, a nonlinear neural network is trained using the gradient descent as optimization algorithm to minimize the constructed loss function. We validated the prediction performance of OPNet on the VAM corpus, by mean linear error as well as a rank correlation coefficient γ as measures. Comparing to k-Nearest Neighbor and support vector regression, the proposed OPNet performs better on the preservation of actual dynamic trend of emotion.

Categories and Subject Descriptors

I.6.5 [**Model Development**]: Modeling methodologies; H.5.2 [**User Interfaces**]: Voice I/O

General Terms

Algorithms, Experimentation, Human Factors

Keywords

speech emotion recognition, dynamic trend of emotion, loss function, neural network

1. INTRODUCTION

In the field of dimensional Speech Emotion Recognition (SER), the regression-based approach has been widely used [5, 8, 3] nowadays. This approach views the SER problem as a standard regression task, and employs regression algorithms to predict continuous-valued emotions. In the past few years, considerable amount of works have been done in this area. Take *Activation* (a dimension of emotion evaluating how high or low the subject's physiological reaction is) prediction for example, Grimm et al. [5] achieved a mean linear error (MLE) of 0.15 and a correlation coefficient (CC) of 0.82 using the Support Vector Regression (SVR) on the Vera am Mittag (VAM) corpus [6], and Wöllmer et al. [8] reached a mean squared error (MSE) of 0.08 using Long Short-Term Memory Recurrent Neural Networks on the Belfast Sensitive Artificial Listener (SAL) database [2].

However, it could be ill-conceived to directly copy these regression solutions to the case of dimensional SER without appropriate adjustment. This is, as unlike a regular regression task aiming only at numerical approximation between predicted and actual values, the dimensional SER task still has an extra demand for preserving the actual dynamic trend of emotion in the prediction process. The so-called *dynamic trend of emotion* is a description of the emotion change during a fixed period of time. In our daily life, it plays

[*] This author is also affiliation with Technische Universität München

ICMI'12, October 22–26, 2012, Santa Monica, California, USA.
Copyright 2012 ACM 978-1-4503-1467-1/12/10 ...$15.00.

an extremely important role for recognizing one's stance toward the current topic in an interaction. For example, an upward trend of emotion on the valence dimension may indicate an increasingly positive evaluation over the product the speaker is talking about, but a downward trend can mean the opposite. Thus, it is important for an advanced human-machine interface to have the ability of predicting the dynamic trend of human's emotion correctly when deciding a machine's next response. However, although the currently used regression algorithms can also contribute preserving actual dynamic trend of emotion by achieving numerical approximation, such effect is, after all, limited.

Driven by this, this paper focuses on developing an improved approach for dimensional SER, which can reach a common benefit between emotions' numerical approximation and dynamic trend approximation. For our case, the latter term *dynamic trend approximation* is used to define the situation that the predicted emotion trend is very close to the actual one for a given set of instances. And what should be noted is that, when we measure how close the predicted and actual trend is, we consider the pairwise trend (i.e., the emotion trend between every pair of instances in the given dataset). This means we say a predicted emotion trend for a set of instances is the same as the actual one, if and only if the trend for every pair of instances from the set is predicted correctly (cf. Section 6.1).

For this purpose, the trend loss and the numerical loss (e.g., MSE) are both considered to define a prediction loss function. The term *trend loss* is used to refer to the trend difference between the actual and predicted emotion sequences (cf. Section3). Then, a Neural Network model is trained using the gradient descent algorithm to minimize above loss function. The major question then becomes how to define the trend loss part of loss function. An ingenious substitution is proposed in this paper: the measure of trend loss is quantified by calculating another measure of ranking order loss. The so-called ranking order loss denotes the order difference between the list of instances ranked by predicted emotion values and the list of instances ranked by annotations. For its calculation, the top-one probabilistic distribution and cross entropy are also involved. We refer to this proposed approach as an Order-Preserving Network (OPNet) algorithm.

Next, to make a comprehensive evaluation of emotion prediction performance, we also propose employing Goodman and Kruskal's γ with generally accepted MLE as our final evaluation measures. Based on these measures, the performance of the proposed OPNet was then compared with two other baseline methods: k-Nearest Neighbor (k-NN) and SVR on the VAM corpus. The results show that the proposed OPNet algorithm can reach a good trade-off between the dynamic trend approximation and numerical approximation.

The rest of the paper is organized as follows. Section 2 gives a general description of dimensional SER. Section 3 introduces the proposed approach for preserving actual dynamic trend of emotion in SER. Database and acoustic feature extraction are described in Section 4 and Section 5, respectively. Section 6 reports our experimental results before concluding.

2. PROBLEM DESCRIPTION

Let us first give a formalized description of the dimensional SER problem. When representing emotions in a k-dimensional emotion space, each emotional state is associated with k emotion values, correspondingly. Moreover, since these dimensions are independent of each other in the psychological view, the dimensional SER task can be naturally separated into k similar and independent subtasks. For simplified depiction, the following description is given in the context of one of the k subtasks.

In training, a set of speech feature vectors $X = \{x_1, x_2, \cdots, x_n\}$ and its corresponding emotion value set $Y = \{y_1, y_2, \cdots, y_n\}$ are given, where n denotes the number of training samples, X is extracted from speech waves, and Y is obtained by human annotation. Each feature vector x_i and its annotated emotion value y_i then form an instance. The training set can be denoted as $I = \{(x_i, y_i)\}_{i=1}^{n}$.

The goal of dimensional SER is to create a prediction function $f(\cdot)$; for the list of feature vectors X, it outputs a list of predicted emotion values $Z = \{f(x_1), f(x_2), \cdots, f(x_n)\}$. This predication function is designed to approximate the predicted value set Z to the true value set Y as much as possible. Here we formalize this objective as a minimization of the total losses with respect to the training data. To wit, with a loss function L,

$$L(Z, Y), \tag{1}$$

the prediction function $f(\cdot)$ can be denoted as:

$$f(\cdot) = \arg\min_{f} L(Z, Y). \tag{2}$$

In emotion prediction, when a new speech sample is given, we extract feature vector x' and use the trained prediction function $f(\cdot)$ to assign an emotion value to it per dimension.

In brief, when it comes to build a dimensional SER system, two crucial problems need to be solved: 1) how to define the loss function L, and 2) how to model the predication function f to minimize L.

3. APPROACH

3.1 Definition of the loss function

The definition of loss function is a procedure of measuring the difference between predicted result and reference standard. It quite depends on the final purpose of the task. For example, if we aim only at the numerical approximation as normal regression tasks do, the general terms of loss functions, like squared loss, absolute value loss, etc, could be already appropriate. However, if we additionally desire a preservation of actual dynamic trend of emotion during the prediction, there is no doubt that a more comprehensive loss function is demanded. In this work, a linear combination of the numerical loss L_{num} and trend loss L_{trend} is considered to construct the loss function L:

$$L = (1 - \eta) \cdot L_{num} + \eta \cdot L_{trend}, \tag{3}$$

where $\eta\ in\ [0, 1]$ is an adjustment factor. With η, the proportion of L_{num} and L_{trend} in the whole loss can then be adjusted as appropriate. Furthermore, for the numerical component L_{num} of loss function L, the widely used squared loss function is chosen:

$$L_{num}(Y, Z) = \frac{1}{2n} \cdot \sum_{i=1}^{n} (y_i - f(x_i))^2, \tag{4}$$

where the y_i and $f(x_i)$ are the human label and predicted label of speech instance x_i, respectively.

The remaining problem then becomes the definition of the trend loss part of loss function: L_{trend}.

3.2 From trend loss to order loss

Supposing given a sequence of emotion values named A with upward trend and another sequence of emotion values named B with downward trend, we can say the dynamic trend of A is entirely different from B's. But now the problem is how to quantify such qualitative description of such difference.

Actually, as we get two trends of training samples X according to Y and Z, namely T_Y and T_Z, two sequences of training samples π_Y and π_Z in descending order can be built as well. There is no doubt that any difference between T_Y and T_Z corresponds to a unique difference between π_Y and π_Z. Inspired by this, we propose to use what we call order loss to quantify the trend loss. To calculate the order loss, a probabilistic method is imported in this work.

Specifically, we map the Y and Z to probability distributions P_Y and P_Z, respectively, using the top one probability model (c.f. **Definition 7** in [1]). For training sample x_i, its top one probability according to Y is defined as

$$P_Y(x_i) = \frac{\exp(y_i)}{\sum_{j=1}^{n}\exp(y_j)}, \tag{5}$$

which represents the probability of the sample x_i being ranked on the top of the descending sequence π_Y. Then $P_Y = \{P_Y(x_i)\}_{i=1}^{n}$ forms a probability distribution over training set X. Likewise, x_i's top one probability according to Z is defined as

$$P_Z(x_i) = \frac{\exp(f(x_i))}{\sum_{j=1}^{n} exp(f(x_j))}. \tag{6}$$

$P_Z = \{P_Z(x_i)\}_{i=1}^{n}$ also forms a probability distribution over training set X.

Then we can take any metric between probability distributions as an order loss function. In this work the cross entropy is chosen, so the order loss function becomes

$$L_{trend}(Y, Z) = -\sum_{i=1}^{n} P_Y(x_i)\log\big(P_Z(x_i)\big). \tag{7}$$

If and only if $P_Y = P_Z$ (i. e., $\pi_Y = \pi_Z, T_Y = T_Z$), the L_{trend} reaches its minimum value. And, the more differences exist between T_Y and T_Z, the larger L_{trend} is. Finally, the total loss function L becomes

$$L = (1-\eta)\cdot\frac{1}{2n}\cdot\sum_{i=1}^{n}(y_i - f(x_i))^2 - \eta\cdot\sum_{i=1}^{n} P_Y(x_i)\log\big(P_Z(x_i)\big). \tag{8}$$

3.3 Order-Preserving Network

We next propose a learning method for optimizing the loss function defined in Eq. (8), with a neural network as model and gradient descent as optimization algorithm. Considering that this method is designed for order (trend) preservation as well as for numerical approximation, we refer to it as Order-Preserving Network (OPNet).

Table 1: Learning algorithm of OPNet.

```
Initialize: weight matrix $w$ and bias $b$
Input: training instances set $I = \big\{(x_i, y_i)\big\}_{i=1}^{n}$,
number of iterations $T$, adjustment rate $\eta$, learning
rate $\alpha$ and threshold $\varepsilon$
for $t = 1$ to $T$ do
    Calculate $Z = \{f(x_i)\}_{i=1}^{n}$ using Eq. (9)
    Calculate total loss $L$ using Eq. (8)
    if $L < \varepsilon$
        then Stop network training
    end if
    Calculate $\Delta w$ and $\Delta b$ using Eq.(10) and (11)
    Update $w = w - \alpha \cdot \Delta w,\ b = b - \alpha \cdot \Delta b$
end for
Stop network training
```

To guarantee the performance of our method, a three-layer nonlinear neural network model was used in the experiments. However, for the purpose of easy description and space saving, the description presented in this section is based on a two-layer neural network without hidden layer:

$$f(x_i) = \mathrm{tansig}(w^{\mathrm{T}}x_i + b), \tag{9}$$

where w is the weight matrix connecting input layer and output layer, b denotes the bias of output unit, tansig$(\cdot)$ is the transfer function to map network outputs into a range of [-1, 1].

Apparently, the loss function based on Eq. (9) is differentiable, so it can be minimized using gradient descent techniques. The gradient of $L(y, z)$ with respect to parameter $w = \{w_k\}_{k=1}^{m}$ and b can be calculated as follows,

$$\begin{aligned}\Delta w_k = \frac{\partial L(y,z)}{\partial w_k} &= \Big[(1-\eta)\cdot\frac{1}{n}\cdot\sum_{i=1}^{n}\Big(f(x_i) - y_i\Big) + \\ &\eta\cdot\sum_{i=1}^{n}\Big(P_Y(x_i) - P_Z\big(f(x_i)\big)\Big)\Big]\Big(1 - f^2(x_i)\Big)x_i,\end{aligned} \tag{10}$$

$$\begin{aligned}\Delta b = \frac{\partial L(y,z)}{\partial b} &= \Big[(1-\eta)\cdot\frac{1}{n}\cdot\sum_{i=1}^{n}\Big(f(x_i) - y_i\Big) + \\ &\eta\cdot\sum_{i=1}^{n}\Big(P_Y(x_i) - P_Z\big(f(x_i)\big)\Big)\Big]\Big(1 - f^2(x_i)\Big),\end{aligned} \tag{11}$$

where $k = 1, 2, \cdots, m$, and m denotes the number of feature's dimensions. Table 1 shows the learning method of the OPNet. From the learning method we can see that: If the adjustment rate η is set to be 0, the OPNet actually becomes equivalent to a normal feed forward neural network, while when η is set to be 1, the OPNet entirely concentrates on emotion ranking.

4. DATABASE

The database used in our work is the audio part of VAM corpus. The data was collected from a German talk-show named Vera am Mittag and consists of spontaneous, unscripted speech from the talk-show guests. All signals were recorded at 16 kHz with 16 bits per sample. It contains 47 speakers (11m/36f), and a total of 947 utterances. All utterances were annotated on three dimensions: ACTIVATION, DOMINANCE and VALENCE by 6 to 17 raters. The self assessment manikins were used as annotation scheme. Final labels for all utterances were given in the normalized range of [-1, 1] in each dimension by application of evaluator weighted es-

Table 2: Set of 31 low-level descriptors.

Energy & spectral low-level descriptors (25)
loudness (auditory model based), zero crossing rate, energy in bands from 250 – 650 Hz, 1 kHz – 4kHz, 25%, 50%, 75%, and 90% spectral roll-off points, spectral flux, entropy, variance, skewness, kurtosis, psychoacoustic sharpness, harmonicity, MFCC 1–10
Voicing related low-level descriptors (6)
F_0 (sub-harmonic summation (SHS) followed by Viterbi smoothing), probability of voicing, jitter, shimmer (local), jitter (delta: 'jitter of jitter'), logarithmic Harmonics-to-Noise Ratio (logHNR)

Table 3: Set of all 42 functionals. [1]Not applied to delta coefficient contours. [2]For delta coefficients the mean of only positive values is applied, otherwise the arithmetic mean is applied. [3]Not applied to voicing related LLD.

Statistical functionals (23)
(positive[2]) arithmetic mean, root quadratic mean, standard deviation, flatness, skewness, kurtosis, quartiles, inter-quartile ranges, 1 %, 99 % percentile, percentile range 1%–99%, percentage of frames contour is above: minimum +25%, 50%, and 90% of the range, percentage of frames contour is rising, maximum, mean, minimum segment length[3], standard deviation of segment length[3]
Regression functionals[1] (4)
linear regression slope, and corresponding approximation error (linear), quadratic regression coefficient a, and approximation error (linear)
Local minima/maxima related functionals[1] (9)
mean and standard deviation of rising and falling slopes (minimum to maximum), mean and standard deviation of inter maxima distances, amplitude mean of maxima, amplitude mean of minima, amplitude range of maxima
Other[1,3] (6)
Linear Prediction (LP) gain, LP Coefficients 1 – 5

timate. For more detailed information about the database, please refer to [6].

5. AUDIO FEATURE EXTRACTION

The audio feature set used is the baseline audio feature set adopted in the AVEC 2011 [7] with 1941 features, extracted by the OpenSMILE [4] toolkit at utterance level. It is composed of 25 energy and spectral related low-level descriptors (LLD) × 42 functionals, 6 voicing related LLD × 32 functionals, 25 delta coefficients of the energy/spectral LLD × 23 functionals, 6 delta coefficients of the voicing related LLD × 19 functionals, and 10 voiced/unvoiced durational features. Details for the LLD and functionals are given in Table 2 and Table 3 respectively.

6. EXPERIMENTS AND RESULTS

In order to verify the validity of the proposed OPNet, we launched a series of experiments on the VAM corpus, and compared its performance with two other baseline regression methods: k-NN and SVR as used in [6]. The VAM corpus was split into ten partitions for stratified cross validation purpose as is the standard procedure on this corpus [6]. In the experiment, our OPNet is set with three layers, with 1 unit in the output layer, 100 units in the hidden layer, and 1941 units (i.e., the size of AVEC 2011 acoustic feature set) in the input layer, respectively.

6.1 Measures of prediction

The goal of our approach is to preserve the dynamic trend of emotion for a given set of instances as much as possible. To fully assess the performance of emotion recognition methods, especially their performance on the pairwise trend preservation, we adopt the nowadays commonly used measures MLE in combination with a rank correlation coefficient (RCC) named Goodman and Kruskal's γ. In statistics, this γ test is often used to measure the degree of similarity between two rankings. Considering the close relationship between dynamic trend and ranking order as discussed in Section 3.2, it is a natural choice to employ a RCC to measure the similarity between two dynamic trends. For T_Y and T_Z and their corresponding ranking lists π_Y and π_Z, γ is defined as

$$\gamma = \frac{N_s - N_d}{N}, \tag{12}$$

where N_s denotes the number of instance pairs ranked in the same order on π_Y and π_Z, N_d denotes the number of pairs of instances ranked differently, and $N = N_s + N_d = n(n-1)/2$ is the total number of all instance pairs for a n size dataset. It is inside the interval [-1, 1] and assumes the the value -1 if one ranking is the inverse of the other, 0 if the rankings are completely independent, and 1 if the two rankings are the same.

In other words, we are actually making use of a kind of pairwise emotion trend for γ calculation. We consider the trend for every pair of instances in a given set for performance measure. For each pair of instances, its trend can be one of the following three states: upward, downward and unchangeable. By given this, the N_s becomes to the number of instance pairs which are predicted with correct trends, while the N_d becomes to the number of instance pairs which are predicted with incorrect trends. If and only if the trend for every pair of instances from the set is predicted correctly, the γ equals 1.

6.2 Results and discussion

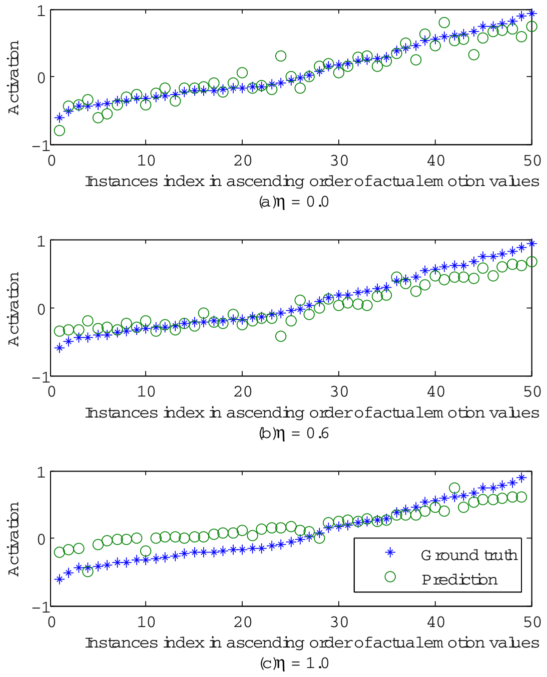
Table 4 summarizes the MLE and γ between the predicted emotions and actual emotions for the three estimators, for each emotion dimension separately. Especially for the OPNet, to investigate the performance impact when η is changing, the measures obtained as η was assigned to .0, .2, .4, .8 and 1.0 are all shown, respectively. And when η is set to .0, the OPNet becomes a normal neural network actually.

It can be observed that the γs of **Act** (activation) and **Dom** (dominance) prediction are obviously higher than those of **Val** (valence) prediction. However, this is a well-known typical behavior as **Act** (activation) and **Dom** are usually well-assessed by acoustic descriptors, whereas **Val** benefits from linguistic information or additional facial expression information [3, 5].

Comparing the measures for SVR and OPNet ($\eta = 0$), we can see that they show similar performance on each dimension, even their MLE are just the same. In this sense, it seems that our proposed OPNet doesn't outperform SVR as

Table 4: Performance comparison of k-NN, SVR and OPNet on VAM corpus.

	Measures	k-NN	SVR	OPNet					
				$\eta = .0$	$\eta = .2$	$\eta = .4$	$\eta = .6$	$\eta = .8$	$\eta = 1.0$
Val	MLE	.13	.13	.13	.13	.13	.13	.14	.42
	γ	.274	.301	.297	.313	.336	.357	.362	.418
Act	MLE	.18	.15	.15	.16	.16	.17	.17	.45
	γ	.539	.625	.622	.634	.657	.681	.704	.715
Dom	MLE	.16	.14	.14	.14	.15	.16	.16	.39
	γ	.532	.595	.589	.603	.634	.651	.670	.726



(a) η = 0.0

(b) η = 0.6

(c) η = 1.0

Figure 1: Emotion prediction results by OPNet when $\eta =$ .0, .6 and 1.0: results in comparison with ground truth.

we looked forward to, especially in the task of preserving the dynamic trend of emotion. Actually, this is because when the η is set to 0, our OPNet just becomes to a normal feed forward network. So we need to modify the value of η to adjust the performance of OPNet. We can realize clearly that when the η increases, OPNet's γ becomes higher as well, which means the capability of OPNet for emotion trend preservation becomes better. It has to be noted that when η is less than .8, the γ increases slowly and steadily, but after that point the rise of γ becomes relatively fast.

However, we observe that for the improvement of γ, we pay a cost of deterioration of MLE. Take the **Act** prediction as an example, with an increasing η from .0 to 1.0, the γ is improved from .622 to .715, whereas MLE is increased from .15 to .45. When η equals 1.0, its MLE are even worse than k-NN. But this is reasonable, since OPNet's focus is shifting to trend preservation more and more along with η's increase (cf. Eq. (3)). Moreover, from the table we can see that when η is set to be .4, the OPNet achieves similar MLE with k-NN, and better γ than both the k-NN and SVR.

As a summary of the results, we randomly select 50 instances from the test set, and draw their prediction results in Fig.1. In this figure, the emotion values are arranged in ascending order for clear comparison. We can see clearly that with the growth of η the emotion trend of these 50 instances becomes better.

7. CONCLUSION

In this paper, we believe that a strong capability for predicting correct dynamic trend of emotion can assist a human-computer interface to make correct decisions. Driven by this, our work then mainly focused on the modeling of a dimensional SER approach considering the preservation of actual dynamic trend of emotion in the prediction process. Specifically, the key issue of dimensional SER was formalized to be a minimization task of a loss function. Moreover, numerical loss and trend loss were linearly combined to construct the total loss function in this paper. For trend loss definition, a probability method was used: firstly the sequence of predicted emotion values and the sequence of actual emotion values were mapped to two top one probability distributions, then a metric between probability distributions named cross entropy was calculated as trend loss. Next, the optimization of loss function was implemented by the proposed OPNet algorithm. The core of the OPNet involved a neural network model and a gradient descent optimization. For the experimental verification, a rank correlation coefficient named Goodman and Kruskal's γ was imported to measure the degree of trend similarity between two emotion sequences. With γ, the SER system's capability of preserving the dynamic trend of emotion can be reflected intuitively. The experiment results show that, comparing to normal regression methods, the proposed OPNet expresses better capability on preserving the dynamic trend of emotion with a suitable adjustment rata η.

As to future work, we intend to investigate on the dynamic trend of emotion for each person during a meaningful speech segment instead the whole range of a database. We also plan to try different measures to evaluate the degree of trend similarity between two emotion sequences.

8. REFERENCES

[1] Z. Cao, T. Qin, and T. Liu. Learning to rank: from pairwise approach to listwise approach. In *ICML*, pages 129–136, Corvallis, USA, 2007.

[2] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis. The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. *Lecture Notes in Computer Science*, 4738:488–500, 2007.

[3] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie. On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. *Journal on Multimodal User Interfaces*, 3:7–19, 2010.

[4] F. Eyben, M. Wöllmer, and B. Schuller. openSMILE - the Munich versatile and fast open-source audio feature extractor. In *ACM Multimedia*, pages 1459–1462, Firenze, Italy, 2010.

[5] M. Grimm, K. Kroschel, and S. Narayanan. Support vector regression for automatic recognition of spontaneous emotions in speech. In *ICASSP*, volume IV, pages 1085–1088, Honolulu, USA, 2007. IEEE.

[6] M. Grimm, K. Kroschel, and S. Narayanan. The Vera am Mittag German audio-visual emotional speech database. In *ICME*, pages 865–868, Hannover, Germany, 2008.

[7] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. AVEC 2011 - the first international audio/visual emotion challenge. *Lecture Notes in Computer Science*, 6975:415–424, 2011.

[8] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie. Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *INTERSPEECH*, pages 597–600, Brisbane, 2008.