The Research on Cross-Language Emotion Recognition Algorithm for Hearing Aid

Xia Shulan*, Wang Jilin

College of Electrical Engineering, Yancheng Institute of Technology, Yancheng 224051, Jiangsu, PR China

Received 1 October 2014, www.cmnt.lv

Abstract

To improve the emotional perception of hearing-impaired people, the speech emotion recognition cross language algorithm is proposed. The mutual features of speech emotions in the Chinese and German corpora are analyzed and a speech emotion recognition algorithm using the Restricted Boltzmann Machine based on Parallel Tempering is proposed. First, the algorithm calculates 375 acoustic features. Then, by using the method of Fisher Discrimination Ratio combined with weighted feature fusion, 20 features are selected to be used in the speech emotion recognition. Afterwards, it sets up the Restricted Boltzmann Machine to recognize 5 types of speech emotions (anger, fear, happiness, neutral, sadness). The experiments show that this algorithm can effectively recognize the cross-language speech emotions except the neutral emotion.

Keywords: hearing aid, speech emotion recognition, Restricted Boltzmann Machine, cross-language

1 Introduction

Emotional communication is an interpersonal regulator. Because of hearing impairment, hearing-impaired patients cannot perceive some important speech components, so there is a significant obstacle in the emotional exchanges among people. With age, more and more people will have different degrees of hearing loss [1], which makes the emotional interaction problem particularly prominent. Although in recent years, hearing aid research in China is deeper and deeper [2-4], few studies focused on emotional problems for hearing-impaired patients have been done, especially for the cross-language emotion recognition.

There are huge difficulties in cross-language emotion recognition for hearing aid due to the differences in emotional expressions and understandings caused by cultural differences. Emotion expression is influenced by many factors like ages, genders, contexts and cultural backgrounds, etc. This leads to the special difficulties in speech emotion recognition compared with other pattern recognition problems. One important element in the cross-language speech emotion recognition is the differences in emotional expressions and understandings caused by cultural differences. In real situations, the cross-language emotion data often come from speakers of different regions and cultures, so the cross-language emotion recognition issue is partly a cross-cultural issue. Although there is a universal commonality in emotion understanding among humans, and people from different language families have approximately same interpretations on emotions as is mentioned in [5, 6], cultural differences will still lead to the diversity of speech emotion expressions. The difficulties in recognition brought by this diversity will be reflected in the following cross-language experiments.

In the area of multi-factor analysis on emotion expression, Tawari [7] analyzed the contextual information in speech emotion recognition and used a gender-specific emotion recognizer. James A. Russell [6] studied the universal recognition on facial expression from the perspective of different cultural backgrounds. Elfenbein and Ambady [5] studied the cultural differences in emotion recognition. Cross-language and cross-cultural researches are always important contents in behavioural science. In the area of emotion recognition research, Smith [8] used to test the emotion recognition abilities of people from different cultures. However, there is still no mature cross-language recognition system in the speech emotion recognition area at present.

In the cross-language speech emotion recognition, the differences among emotion's acoustic features are inevitable problems. The emotional features acquired from the research on one language's database usually perform badly on another language's database. This is caused by the differences in pronunciation characteristics of different languages. We know that in the research and selection of emotional features, main difficulties lie in the exclusion of phonemes' interference. That is, we should choose the acoustic features which are responsive to emotional changes while insensitive to semantic changes. So, the pronouncing characteristics' changes in different languages have a huge influence on emotion's acoustic features, and this will lead to the inconsistency among emotion features acquired from different language's databases which is a great challenge for emotion recognition systems. What we need is an emotion model which is independent of languages types, and we need to find the common points in emotional voice of different languages. And those common points do exist. For example, the rise of speech volume usually means stimulations and might be related to anger. And the rises

* ***Corresponding author's*** e-mail: xslnj@126.com

and falls of pitch traces have much to do with the emotion of happiness. All these characteristics in pronunciations are reported specifically in recent years' researches on speech emotion recognition [9-13].

This article analyzes the mutual features of speech emotions in the Chinese and German corpora, and proposes a speech emotion recognition algorithm using the Restricted Boltzmann Machine based on Parallel Tempering for hearing aid. The algorithm selects language features by using the method of Fisher Discrimination Ratio combined with weighted feature fusion, and realizes the speech emotion recognition by setting up the Restricted Boltzmann Machine. The experiments show that this algorithm can effectively recognize the cross-language speech emotions except the neutral emotion.

2 Cross-Language Feature Analysis

This article studies the cross-language speech emotion features based on a Chinese and a German database. The Chinese emotional speech database [14] consists of voice from 6 males and 6 females. It is comprised of 6 types of basic emotions, namely anger, fear, happiness, neutral, sadness and surprise. The widely-used Berlin Database of emotional speech[15] consists of 7 types of emotions from 5 males and 5 females, namely anger, neutral, fear, boredom, happiness, sadness and disgust. This article selects 5 common emotions as the cross-language emotion features, which are anger, fear, happiness, neutral and sadness. The experimental dataset is shown in Table 1.

TABLE 1 The Cross-Language Experiment Dataset

Dataset 1	
Berlin Database of emotional speech	
Type of emotion	Number of samples
anger	127
fear	69
happiness	71
neutral	79
sadness	62

Dataset 2	
Chinese emotional speech database	
Type of emotion	Number of samples
anger	127
fear	69
happiness	71
neutral	79
sadness	62

This article takes 375 features for feature selection and recognition as is shown in Table 2, and uses the Fisher Discrimination Ratio to evaluate every original features. After the evaluation, we use the weighting fusion method to get the evaluation scores on cross-language features

$$FDR = w \times FDR1 + (1-w) \times FDR2 . \quad (1)$$

Here, FDR is the comprehensive feature evaluation result on Chinese and German, FDR1 is the evaluation result on German, FDR2 is on Chinese, and w is the weight in fusion. Table 3 shows the cross-language feature selection results. After determining the emotional features, this article will use the method of Restricted Boltzmann Machine combined with the Support Vector Machine to recognize the emotions.

TABLE 2 List of Acoustic Features

No	**Features**
1-15:	Fundamental tone, averages of its 1st and 2nd order difference, its maximum, minimum, range, and standard deviation.
16-90:	Frequencies of 1st to 5th formant, averages of their 1st and 2nd order differences, their maximums, minimums, ranges, and standard deviations.
91-165:	Bandwidth of 1st to 5th formant, averages of their 1st and 2nd order differences, their maximums, minimums, ranges, and standard deviations.
166-180:	Short-term energy, averages of its 1st and 2nd order difference, its maximum, minimum, range, and standard deviation.
181-375:	Mel-Frequency Cepstral Coefficients (MFCC-0 to MFCC-12), averages of their 1st and 2nd order differences, their maximums, minimums, ranges, and standard deviations.62

TABLE 3 Cross-Language Feature Selection Results

Database and weight	**The best 20 features from selection (features' No.)**
DB1 (German)	3, 1, 213, 183, 184, 211, 256, 345, 214, 330, 215, 260, 340, 185, 360, 241, 315, 300, 233, 286
DB2 (Chinese)	3, 1, 176, 168, 177, 172, 179, 178, 169, 167, 174, 327, 171, 101, 329, 173, 8, 131, 96, 185
Cross-Database evaluation (weight 0.3)	3, 1, 176, 168, 177, 172, 179, 178, 169, 167, 174, 327, 184, 171, 8, 213, 185, 211, 183, 101
Cross-Database evaluation (weight 0.7)	3, 1, 213, 183, 184, 211, 176, 256, 345, 214, 168, 185, 330, 260, 215, 177, 340, 360, 241, 171
Cross-Database evaluation (weight 0.9)	3, 1, 213, 183, 184, 211, 256, 345, 214, 330, 215, 260, 185, 340, 360, 241, 315, 300, 233, 286

3 Restricted Boltzmann Machine(RBM)

The learning algorithm of Deep Belief Network (DBN) is proposed by Hinton [16]. In essence, this method learns from more useful features by setting up a machine learning model with many hidden layers and taking massive amount of training data, and finally increases the accuracy in classification or prediction. DBN can be viewed as a complicated neural network consisting of many layers of Restricted Boltzmann Machine (RBM). RBM has a visible layer and several hidden layers, and uses the Contrastive Divergence algorithm to train its weights which has good effects in binary neuron models. Sampling plays an important role in constructing neural networks. In order to improve the sampling accuracy, some researches use the Parallel Tempering (PT) method to take samples [17]. In the PT Monte Carlo algorithm, many non-contact temperature replicas do sampling with each other. After many times of sampling, comparing and swapping, it gets the sampling values which satisfy the model's joint distribution. The more similarity the sampling values have with the model's joint probability distribution, the less errors they will have, and the better the model will be. The Contrastive Divergence method based on PT has received good test results in RBM [18].

Restricted Boltzmann Machine (RBM) model is a special form of Boltzmann Machine (BM). RBM consists of visible layers *v* and hidden layers *h*. There is no connection between nodes in the same layers, that is, no connection exists between nodes in hidden layers and those in visible layers, and the nodes in the same layer are conditional independent of each other.

When inputting *v*, we will get the hidden layer *h* from $p(h|v)$. Then, we will also get the visible layer from $p(h|v)$. RBM is an energy-based model, in which the energy of the joint configuration of visible variables v and hidden variables *h* is

$$E(v,h) = -\sum_{i \in visible} a_i v_i - \sum_{j \in hidden} b_j h_j - \sum_{i,j} v_i h_j w_{ij} . \quad (2)$$

Usually, RBM's parameters can be denoted as

$$\theta = \{W, a, b\},$$

where *W* is the weight of the connections between visible units and hidden units, and *a* and *b* are the bias vectors of visible units and hidden units respecttively. The joint probability of visible layer *v* and hidden layer *h* can be calculated from the joint configuration energy of *v* and *h*

$$P_\theta(v,h) = \frac{1}{Z(\theta)} \exp(-E(v,h;\theta)) . \quad (3)$$

The model can achieve the optimal performance by training RBM's parameters [19]. The training methods include Contrastive Divergence, Maximum Random Likelihood, etc. By training the binary neurons in RBM using the Contrastive Divergence method, we can get the input probabilities in active state

$$\begin{cases} P(h_j = 1|v) = sigmoid\left(\sum_i w_{ij} v_i + a_j\right) \\ P(v_i = 1|h) = sigmoid\left(\sum_j w_{ij} h_j + b_i\right) \end{cases}, \quad (4)$$

where $sigmoid(x) = \frac{1}{1+e(-x)}$.

4 RBM based on Parallel Tempering (RBM-PT)

Parallel Tempering (PT) sampling is an effective way for RBM training. In the RBM-PT training process, every temperature corresponds to a Gibbs chain and is sampled by the PT method. Every Gibbs chain corresponds to the temperature t_i, where $1 = t_1 < \ldots\ldots < t_i < \ldots\ldots < t_{M-1} < t_M$. Whether or not to swap sampling values between different Gibbs chains is determined according to specific conditions.

According to equation (3), the joint probabilities of RBM-PT under different temperatures are

$$P_r(v,h) = \frac{1}{Z(t_i)} \exp\left(-\frac{1}{t_i} E(v,h;\theta)\right). \quad (5)$$

The PT Monte Carlo algorithm has two steps:

STEP 1. Metropolis-Hastings sampling step [20]: According to the existing sampling values, calculate the next sampling point under the current temperature. The basic sampling function is

$$x^{i+1} | x^i = Metropolis - Hastings\left(x^i + N\left(0, \frac{\sigma_i^2}{t_k}\right)\right), \quad (6)$$

where $N\left(0, \frac{\sigma_i^2}{t_k}\right)$ represents a normal distribution function with mean 0 and variance $\frac{\sigma_i^2}{t_k}$, and t_k represents the temperatures.

STEP 2. Swapping step: After sampling, find out whether or not the visible and hidden nodes under the neighbouring temperatures t_r and t_{r-1} in the temperature set, that is (v_r, h_r) and (v_{r-1}, h_{r-1}), satisfy the swapping requirements. The swapping requirement in RBM-PT is

$$\min\left\{1, \exp\left(\left(\frac{1}{t_r} - \frac{1}{t_{r-1}}\right) * \left(E(v_r, h_r) - E(v_{r-1}, h_{r-1})\right)\right)\right\}. \quad (7)$$

If this condition is satisfied, swap the sampling points in the neighbouring temperature chains, or make no swap otherwise. After many cycles of sampling and swapping, finally, the sampling value in temperature $t_1 = 1$ is used as the parameter θ in the RBM pre-training model. The target sampling value obtained from PT enables the RBM training to achieve good results.

Multiply the connection weight value *W* between visible and hidden nodes from the RBM parameter $\theta_{RBM} = \{W, a, b\}$ from equation (1) by the temperature β,

then the whole model will become $\theta_{RBM-PT} = \{\beta W, a, b\}$ while the bias weights a and b will not change. Now, the PT algorithm can be implemented into the RBM to improve the training efficiency.

5 Simulation

In the emotion recognition experiment based on RBM, this article does tests on every sets of cross-language features in Tab. 3. In order to study the cross-language characteristics, training set and testing set use different languages. That is, the algorithm trains in German and tests in Chinese, or trains in Chinese and tests in German. Due to the insufficiency of data, we use the same dataset for feature selection and recognition test.

5.1 RESULTS OF NEURAL NETWORK'S TRAINING

In order to improve the recognition effect, the neural network is trained first. This article uses 300 training samples and 100 test samples. Input the randomly selected training samples into the neural network for training, and set the maximum training times as 5000 and target error as 10-4. The convergence curve for training is shown in Figure 1. We can see that after 612 times of training, the mean squared error of the neural network converges to the expected error limits. Use the test samples to verify the trained neural network, and the speech emotion recognition network is established.

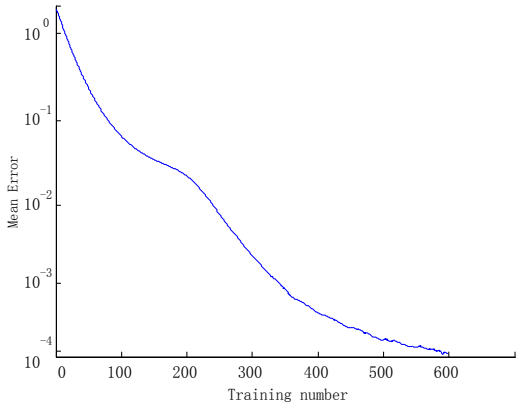
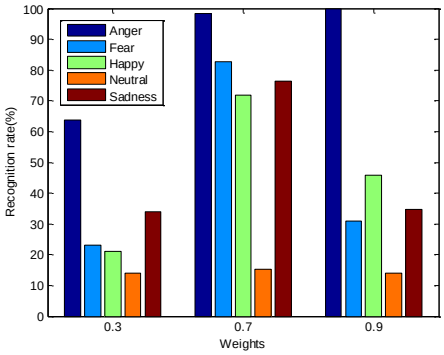


FIGURE 1 The Iterative Curve of DBN

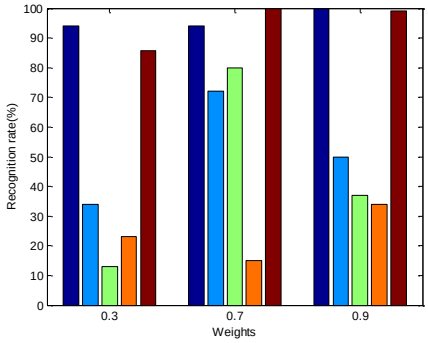
5.2 SPEECH EMOTION RECOGNITION EXPERIMENTS

The average recognition rate increases as the fusion weight increases. From the aspect of average recognition rate, the system performs best when the fusion weight is 0.7.

Figure 2 shows that all emotion's recognition rate can exceed 70% except the neutral emotion. Every emotion recognition rate exceeds 70% when the fusion weight is 0.7. Even though the false recognition rates between emotions is relatively high when the fusion weight is 0.3 or 0.9, the high recognition rates on one or two emotion targets still demonstrate that these acoustic models have the same emotional patterns on German and Chinese.



a)



b)

FIGURE 2 the Cross-Language Training and Testing Results:
a) Train in Chinese and test in German,
b) Train in German and test in Chinese.

6 Conclusions

In this paper, a speech emotion recognition algorithm for hearing aid based on using the Restricted Boltzmann Machine based on Parallel Tempering is proposed. Aiming at the problem of cross-language speech emotion recognition, this article studied and experimented on the speech emotion recognition between Chinese and German. This article selected the cross-language features by using the method of Fisher Discrimination Ratio combined with weighted feature fusion, and established emotional models based on the Restricted Boltzmann Machine. The experiments showed that there exists a generic acoustic model between Chinese and German to describe the same speech emotion behaviours. But from experiments, the algorithm should be greatly improved. In addition, although the method has large training time, but during the actual identification, the algorithm takes less time and has high efficiency. So, in conclusion, this algorithm is suitable for the hearing aid.

Acknowledgments

The work was supported by the National Natural Science Foundation of China under Grant No. 61301219, No. 61375028 and No. 61301295, the Natural Science Foundation of Jiangsu Province under Grant No. BK20130241.

References

[1] Abrams H B 2012 *Am. J. Audiol*. 21(2), 329-30.
[2] Liang R, Zou C, Zhao L, Wang Q and Xi J 2012 *Shengxue Xuebao/Acta Acustica*, 37(5), 527-33.
[3] Ruiyu Liang J X, Jian Zhou, Cairong Zou and Li Zhao 2013 *Appl. Acoust*. 74(1): 71-78.
[4] Liang Rui-Yu, Xi Ji, Zhao Li, Zou Cairong and Huang Chengwei 2012 *Acat. Phys. Sin-ch ED*, 61(13), 134305(1-11).
[5] Elfenbein H A, Ambady N 2002 *Psychol. Bull*. 128(2), 203-35.
[6] Russell J A 1994 *Psychol. Bull*. 115(1), 102-41.
[7] Tawari A, Trivedi M M 2010 *IEEE T. Multimedia*, 12(6), 502-9.
[8] Edmond M B, Granberg E, Simons R, Lei M K. 2014. Journal of Adult Development. 21(1): 13-29
[9] Xuecheng Jin. The research on affect recognition based on speech signals. 2007, University of Science and Technology of China: Hefei.
[10] Sethu V, Ambikairajah E, Epps J. 2013. EURASIP Journal on Audio, Speech, and Music Processing, 2013(1): 1-14
[11] Gobl C, Chasaide A N 2003 *Speech Commun*. 40(1), 189-212.
[12] Navas E, HernáEz I, Luengo I. 2006. IEEE T AUDIO SPEECH. 14(4): 1117-1127
[13] Yeping Wang 2004 The feature analysis and recognition of vocal affect signals, Southeast University: Nanjing.
[14] Lili Cai 2005 The vocal affect analysis and recognition based on information fusion, Southeast University: Nanjing.
[15] Hassan A, Damper R, Niranjan M. 2013. IEEE T AUDIO SPEECH. 21(7): 1458-1468.
[16] Hinton G E, Salakhutdinov R R 2006 *Science* 313(5786), 504-7.
[17] Earl D J, Deem M W 2005 *Phys. Chem. Chem. Phys.* 7(23), 3910-16.
[18] Xu J, Li H, Zhou S. 2014. Neurocomputing. 139: 328-335
[19] Hinton G E 2002 *Neural comput.* 14(8), 1771-1800.
[20] Ping Chen, Ruoxi Xu 2008 *J. Sys. Sci. & Info*. 28(1), 100-8.

Authors

Xia Shulan, August 8, 1969, Nanjing, Jiangsu Province

Current position, grades: Master, the Associate Professor of Yancheng Institute of Technology.
Scientific interest: Signal processing, speech signal processing.
Experience: An expert in the field of signal processing.

Wang Jilin, August 11, 1966, Yancheng, Jiangsu Province

Current position, grades: Master, the Associate Professor of Yancheng Institute of Technology
Scientific interest: Signal processing, speech signal processing
Experience: An expert in the field of signal processing