Improving Generation Performance of Speech Emotion Recognition by Denoising Autoencoders

Linlin Chao, Jianhua Tao, Minghao Yang, Ya Li

National Laboratory of Pattern Recognition (NLPR), Institute of Automation,
Chinese Academy of Sciences, Beijing
{linlin.chao, jhtao, mhyang, yli}@nlpr.ia.ac.cn

Abstract

A speech emotion recognition algorithm should generalize well when the target person's speech samples and prior knowledge about their emotional content are not included in the training data. In order to achieve this objective, we utilize denoising autoencoders based approach to solve this task. In this study, a relatively small dataset, which contains close to 1500 persons' emotion sentences, is introduced. By unsupervised pre-training with this dataset, denoising autoencoders learn features which contain more emotion-specific information than speaker-specific information in data successfully. Experiment results in CASIA dataset show that this denoising autoencoders based approach can improve the generation performance of speech emotion recognition significantly.

Index Terms: speech emotion recognition, denoising autoencoders

1. Introduction

A central challenge in speech emotion recognition is that models and classifiers do not generalize well when training data and testing data come from different sessions or corpus. Even in the same corpus, there is gap between speaker dependent performance and speaker independent performance. The main reason is that all features commonly utilized in speech based emotion recognition capture both emotion-specific information and speaker-specific information [1]. Thus, removing the influence of speaker-specific information as far as possible is important for speech emotion recognition which is the central topic of this paper.

To reduce the influence of speaker variability, applying lots of persons' emotion data can be a useful approach. However, in speech emotion recognition, the data is very difficult to obtain and the annotation also needs a lot of human effort. On the other hand, there are a lot of different databases. Although these databases come from different design criteria, they provide lots of information for discovery.

Meanwhile, deep neutral network has been successfully in many fields. Deep neutral networks are mainly utilized for extracting more abstract features [2, 3, 4]. The hidden layers of deep neutral network can be considered as increasingly complex feature transformations and the transformation can be shared across different datasets. In speech recognition, multiple languages' dataset are trained simultaneously and benefit each other [2, 3]. These give us the inspiration to utilize deep neutral network to learn more abstract features and train several different databases simultaneously. In this way, features which have better generation performance can be learned.

In this context, this paper tries to learn two datasets simultaneously by denoising autoencoders. One of the dataset (ChongQing database) has relatively small size, but a lot of different persons' emotion samples are contained. Thus, by co-training, the hidden layer learns more abstract features, which contain less speaker-specific information. In this way, the generation performance is improved. The CASIA emotion database is tested in this experiment. Experiment results show that by unsupervised pre-training with ChongQing database, denoising autoencoders based approach improves the classification accuracy 7.13% averagely.

The rest of this paper is organized as follows. In Section 2, several related works are introduced. Section 3 introduces the dataset and audio features. The denoising autoencoders are introduced in Section 4. The experiment setup and experiments results are presented in detail in Section 5. The conclusions are given in Section 6.

2. Related Work

Researchers remove the speaker-specific information mainly by speaker normalization. The features are normalized over each speaker (including training data and testing data) by Z-normalization [5, 6]. This method requires the target person's speech samples and prior knowledge about their emotional content. This method is only suitable for limited condition.

When the target person's speech samples are available, but their emotional contents are not, adaptation based approaches are introduced. These approaches try to transfer the existed models to the target person. Kim [8] proposes an MLLR-based online speaker adaptation technique using accumulated personal data. By selective label refinement, this approach estimates the transformation matrices and constructs personalized emotion models. Rahman [7] propose an iterative feature normalization based approach to establish a personalized emotion recognition system. The feature they utilized is mainly pitch based feature. The feature can be normalized by F0 mean value. These adaptation based approaches shows superior performance compared to models without adaptation. However, these methods also need the target person's emotion sample to finish the adaptation process.

Meanwhile, deep learning techniques have found recent successes in various communities including emotion recognition [9, 10, 11]. Stuhlsatzetal [9] uses generatively pre-trained ANNs to learn discriminative features of low dimension and finds improvement in both weighted and unweighted recall on multiple emotion corpora. Rui Xia [10] uses modified denoising autoencoders to learn more robust features and achieves good performance on speech emotion recognition. Kim [11] introduces the DBNs to learn audio-visual features for emotion recognition. This multimodal deep

978-1-4799-4219-0/14/$31.00 ©2014 IEEE

learning also achieves good performances in audio-visual based emotion recognition.

3. Database and Features

There are two databases utilized in this context. One is the public CASIA database. The other database is the recently collected from phone call. These two databases are quite different except they all belong to Chinese.

3.1. CASIA Database

The CASIA database obtains from National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. It's an emotional speech corpus in Mandarin. The corpus is collected from four Chinese native speakers including two men and two women. Everyone expresses 300 sentences in six emotions involving neutral, angry, fear, happy, sad and surprise. The total amount of sentences is 300 x 6 x 4 = 7200. The speech corpuses sampled at 16kHZ frequency and 16 bits resolution with monophonic Windows PCM format. In this experiment, four emotions are utilized, which are happy, angry, sad and neutral.

3.2. ChongQing Dialect Database

The ChongQing dialect database is collected from people's phone call. It's an emotional speech corpus in ChongQing dialect, which is very different from Mandarin. It is recorded in real condition, with background noise recorded. The emotions included are happy, angry, sad and neutral. There are 1745 wave files, which are 507 for angry, 417 for happy, 568 for neutral and 253 for sad. The mean duration of these wave files is 30 seconds. One of the biggest advantages of this database is that these wav files come from different persons. Although this database is relatively small, it has plenty of speakers' information..

3.3. Feature

The features used as the input of the denoising autoencoders are static and dynamic features, which have been successfully applied to emotion recognition task. There are 384 features, composing of low-level descriptors (LLD) related to energy, spectral, voicing, etc, and their statics. These features are used in the INTERSPEECH 2009 Emotion Challenge [12], which are extracted by OpenSMILE toolkit [13]. Table 1 shows details of these features.

All these features are PCA whitened before they are input to the denoising autoencoders.

Table 1 *Features sets: 32 low-level descriptors (LLD) and 12 functions.*

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

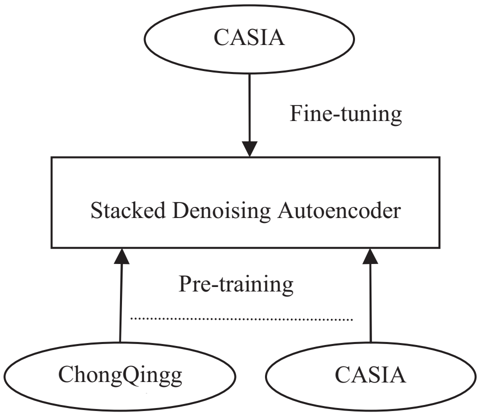

Figure 1 *The outline of the proposed approach*

4. Approach

To improve the generation performance, the speaker-specific information in the feature should be removed as far as possible. As the CASIA database only has four persons' emotion samples and ChongQing dialect database has plenty of persons' samples, these two databases are trained simultaneously to learn features, which have relatively small speaker-specific information for CASIA database. The outline of this approach is showed in figure 1. At first, ChongQing dialect database and CASIA database are utilized to pre-train the denoising autoencoders. Then the fine-tuning stage applies only the CASIA database.

4.1. Denoising Autoencoders

Denoising autoencoders is a stochastic version of autoencoders. It is trained to reconstruct the input from a corrupted version of it. The motivation behind this is to force the hidden layer to discover more robust features. This can be viewed as a regularization method, which improves the generation performance.

Training a denoising autoencoders needs two steps, unsupervised pre-training and supervised fine-tuning. The supervised fine-tuning is an error back propagation process, which updates model parameters based on the labeled training data. The unsupervised pre-training can provide a better initial value than random initialization [4]. Besides, unsupervised pre-training guides the learning towards basins of attraction of minima that support better generalization from the training dataset. The unsupervised pre-training exhibits some properties of a regularizer, which classical regularization techniques cannot achieve [4].

In the pre-training step, corrupted input is generated by adding noise on the clean input. Different noise functions can be utilized to produce the corrupted input. In this context, Gaussian noise is applied. The corrupted input can be expressed as:

$$x_c = x + n$$

(1)

$$n \sim N(0, \sigma^2 I)$$

where σ represents the corruption level.

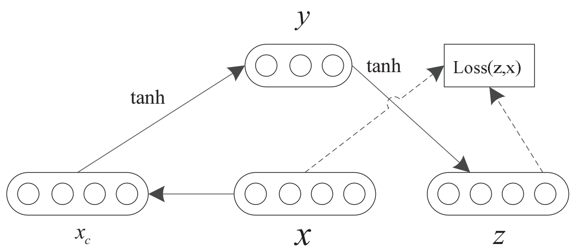

Figure 2 *Architecture of denoising autoencoder*

Then the corrupted input is mapped to a hidden representation through a deterministic mapping:

$$y = s(Wx_c \ + \ b \ ) \qquad (2)$$

In this equation, s is a non-linearity expression such as the sigmoid. As the input to denoising autoencoders belongs to Gaussian distribution, the *tanh* function is utilized as the non-linearity expression. The latent representation y is then mapped back into a reconstruction z of the same shape as x through a similar transformation, e.g.:

$$z = s(W'y \ + \ b') \qquad (3)$$

The weight matrix W' of the reverse mapping can be optionally constrained by $W' = W^T$, which is an instance of tied weights. At last the squared error or cross-entropy function is used as the loss function to train this network. In this context, the squared error function

$$Loss \ = \ \|(z - x)\|^2 \qquad (4)$$

is utilized. The architecture of the denoising autoencoder is showed in figure 2.

5. Experiment

5.1. Experiment Setup

There are four persons' (G1, G2, B1, B2 respectively) speech samples in CASIA database. To better express the generation performance, 400 sentences of G1 are taken as validation dataset. Then G2, B1, B2 are tested respectively. For each testing speaker, the training data is the data from the other speakers.

The baseline results are achieved by SVM [14], RBF kernel is applied. The parameters are fine-tuned by the validation set. There are two baselines. The first one is achieved by speaker normalization. Each speaker's samples are normalized by Z-normalization separately, including training set and testing set. The second baseline is the one without speaker normalization. The training set is normalized by Z-normalization as a whole, and the testing set is normalized by the parameters from training set.

Except this, ChongQing dialect database and CASIA database are also trained simultaneously by SVM in the hope to decrease the influence of speaker-specific information. This experiment can also compare with the proposed approach.

Table 2 *Classification results for each person with different approaches*

	G2	*B1*	*B2*	*Avg.*
Baseline one	58.38	58.25	58.97	58.53
Baseline two	50.94	46.38	42.47	46.60
Proposed approach	55.69	50.50	55.00	53.73

Baseline one is SVM based results with speaker normalization.

Baseline two is SVM based results without speaker normalization.

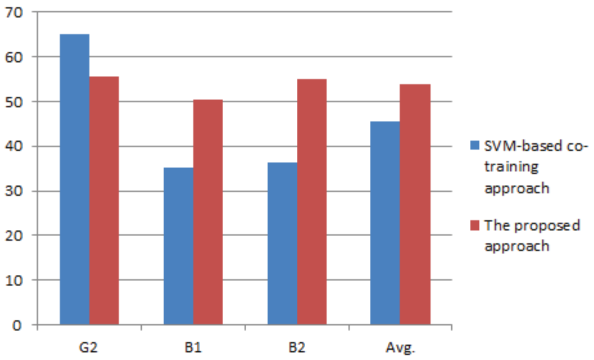

Figure 3 *Comparisions of the SVM-based training ChongQing dialect and CASIA database together and the proposed approach*

For the denoising autoencoders, we use the implementation of theano[1]. The networks have two layers, and each layer has 510 hidden units. The corruption level is 0.1. The validation set is used for early-stopping. The learning rates for G2, B1 and B2 are 0.084, 0.08 and 0.098 separately. All the parameters to train the network are obtained by random search strategy [15].

5.2. Experiment Results

The experiments are shown in table 2 and figure 2. From the table, we can see baseline one get the best classification accuracy, which is 11.93 percent improvement compared with the classification results without speaker normalization. This proves that if the target person's speech samples and their emotional content are known, speaker normalization can remove the speaker-specific information to some extent. However, in real life, this requirement is not easy to satisfy.

The proposed approach faces the 'unseen' data directly, without using the speaker-specific information before testing. This approach gets the second high classification accuracy, and this is the one with the best generation performance. Compared to baseline two, the one the target speaker's prior information is also unused before testing, the proposed approach improves the classification accuracy by 7.13 percent in average.

The figure 3 shows that if the ChongQing dialect database and CASIA database are trained by SVM, the accuracies also not as good as the proposed approach. This proves denoising autoencoders have more strong abilities to utilize different

[1] http://deeplearning.net/tutorial/intro.html

database and learn better features. The SVM based co-training achieves 45.61% in average, which is still lower than baseline two. This suggests the ChongQing dialect database has different distribution with CASIA database. When training the emotion model by SVM, the added ChongQing dialect database deteriorate the performance.

6. Conclusions and Future Work

This article discusses how to improve the generation performance of speech emotion recognition. The main idea is try to reduce the influence of speaker-specific information in emotional features by denoising autoencoders.

The analysis provided in this article indicates that speaker-specific information does exist in the feature extracted. Speaker normalization can solve this problem to some extent. However, this approach needs prior information about the target person. The proposed approach solves this problem by training CASIA database and ChongQing dialect database simultaneously by denoising autoencoders. As ChongQing dialect has plenty of persons' samples, the denoising autoencoders learn features which are not so related to the persons' information in CASIA database. Experiment results prove the effectiveness of the proposed approach.

Improving the generation performance of speech emotion recognition, more persons' data are needed. In this way, speaker-specific information is weakened. At present, there are a lot of emotion databases, which provide diversity speakers' speech samples. Thus we will put our effort in utilizing these existing emotion databases to improve the generation performance of speech emotion recognition.

7. Acknowledgements

This work is supported by the National Social Science Foundation of China (13&ZD189), the National Nature Science Foundation of China (NSFC) (No.61273288, No.61233009, No.61203258, No.61305003, No. 61332017, and No.61375027).

8. References

[1] Sethu, V., Epps, J., and Ambikairajah, E. (2013, May). Speaker variability in speech based emotion models-Analysis and normalisation. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (pp. 7522-7526). IEEE.

[2] Huang, J. T., Li, J., Yu, D., Deng, L., & Gong, Y. (2013, May). Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (pp. 7304-7308). IEEE.

[3] Heigold, G., Vanhoucke, V., Senior, A., Nguyen, P., Ranzato, M., Devin, M., & Dean, J. (2013, May). Multilingual acoustic models using distributed deep neural networks. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (pp. 8619-8623). IEEE.

[4] Erhan, D., Bengio, Y., Courville, A., Manzagol, P. A., Vincent, P., and Bengio, S. (2010). Why does unsupervised pre-training help deep learning. The Journal of Machine Learning Research, 11, 625-660.

[5] Mower, Emily, Maja J. Mataric, and Shrikanth Narayanan. "A framework for automatic human emotion classification using emotion profiles." Audio, Speech, and Language Processing, IEEE Transactions on 19, no. 5 (2011): 1057-1070.

[6] Nicolle, Jérémie, Vincent Rapp, Kévin Bailly, Lionel Prevost, and Mohamed Chetouani. "Robust continuous prediction of human emotions using multiscale dynamic cues." In Proceedings of the 14th ACM international conference on Multimodal interaction, pp. 501-508. ACM, 2012.

[7] Rahman, T., & Busso, C. (2012, March). A personalized emotion recognition system using an unsupervised feature adaptation scheme. In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on (pp. 5117-5120). IEEE.

[8] Kim, J. B., Park, J. S., & Oh, Y. H. (2011, May). On-line speaker adaptation based emotion recognition using incremental emotional information. InAcoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on (pp. 4948-4951). IEEE.

[9] Stuhlsatz, A., Meyer, C., Eyben, F., ZieIke, T., Meier, G., & Schuller, B. (2011, May). Deep neural networks for acoustic emotion recognition: raising the benchmarks. In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on (pp. 5688-5691). IEEE.

[10] Xia, R., and Liu, Y. (2013). Using Denoising Autoencoder for Emotion Recognition. In Proc. of INTERSPEECH.

[11] Kim, Y., Lee, H., and Provost, E. M. (2013, May). Deep learning for robust feature generation in audiovisual emotion recognition. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (pp. 3687-3691). IEEE.

[12] Schuller, B., Steidl, S., and Batliner, A. (2009, September). The INTERSPEECH 2009 emotion challenge. In INTERSPEECH (pp. 312-315).

[13] Eyben, F., Wöllmer, M., & Schuller, B. (2010, October). Opensmile: the munich versatile and fast open-source audio feature extractor. In Proceedings of the international conference on Multimedia (pp. 1459-1462). ACM.

[14] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[15] Bengio, Yoshua. "Deep Learning of Representations for Unsupervised and Transfer Learning." Journal of Machine Learning Research-Proceedings Track27 (2012): 17-36.