J. Biomedical Science and Engineering, 2016, 9, 37-43
http://www.scirp.org/journal/jbise
ISSN Online: 1937-688X
ISSN Print: 1937-6871

Feature Optimization of Speech Emotion Recognition

Chunxia Yu, Ling Xie, Weiping Hu*

GuangXi Key Lab of Multi-Source Information Mining and Security, GuangXi Normal University, Guilin, China
Email: *huwp@mailbox.gxnu.edu.cn

How to cite this paper: Yu, C.X., Xie, L. and Hu, W.P. (2016) Feature Optimization of Speech Emotion Recognition. *J. Biomedical Science and Engineering*, **9**, 37-43.
http://dx.doi.org/10.4236/jbise.2016.910B005

Received: April 11, 2016
Accepted: September 20, 2016
Published: September 23, 2016

Abstract

Speech emotion is divided into four categories, Fear, Happy, Neutral and Surprise in this paper. Traditional features and their statistics are generally applied to recognize speech emotion. In order to quantify each feature's contribution to emotion recognition, a method based on the Back Propagation (BP) neural network is adopted. Then we can obtain the optimal subset of the features. What's more, two new characteristics of speech emotion, MFCC feature extracted from the fundamental frequency curve (MFCCF0) and amplitude perturbation parameters extracted from the short-time average magnitude curve (APSAM), are added to the selected features. With the Gaussian Mixture Model (GMM), we get the highest average recognition rate of the four emotions 82.25%, and the recognition rate of Neutral 90%.

Keywords

Speech Emotion Recognition, Feature Selection, Feature Extraction, BP Neural Network, GMM

1. Introduction

Speech is one of the important ways for human's communication and it is a convenience and simple way to transmit information. The speech signal contains not only the expressed speech meaning, but also the speaker's emotion information which always be ignored by the traditionally speech processing [1]. But the emotion information plays a very important role in the speech communication. Therefore, in recent years, emotion recognition has become a hot spot. Traditional features, such as energy (E), zero-crossing rate (ZCR), the fundamental frequency (F0), the first formant (FF), Mel Frequency Cepstrum Coefficients (MFCC), linear prediction coefficient (LPC), the short-time average magnitude (SAM) etc. and their statistics, such as maximum (Max), minimum (Min), mean, variance (Var.), first order difference (FOD), rate of change (RC) etc. are generally applied to recognize speech emotion. By combining the above features, we can use it to recognize a speech's emotion. In this paper, speech emotion is divided into four categories: Fear, Happy, Neutral and Surprise.

Due to the redundant and unrelated information between feature parameters, it is indispensable to select the features which remarkably characterize speech emotion information [2]. The paper adopts the Back Propagation (BP) neural network to sequence the features in the saliency measure and selects a set of features.

Based on the feature selection, two new characteristics of speech emotion, MFCC feature extracted from the fundamental frequency curve (MFCCF0) and amplitude perturbation parameters extracted from the short-time average magnitude curve (APSAM), are added to the selected features. The Gaussian Mixture Model (GMM) is used to recognize speech emotion [3]. According to the experiment result, the two new added features can effectively increase the recognition rate.

2. Database

A Chinese emotional database (CASIA) is used in this paper. CASIA was released by the Institute of Automation, Chinese Academy of Sciences [4]. It is composed of 1,200 wave files that represent different emotional states: happy, fear, sad, surprise, neutral and angry. Four actors (two females and two males) read 50 different texts respectively in the six emotions. Four kinds of emotions, Fear, Happy, Neutral and Surprise, 800 utterances in total (half of each emotion data to be trained, the remaining half to be tested) are chosen in the experiment.

3. Feature Extraction

3.1. Traditional Features

Traditional features are generally applied to recognize speech emotion which have proven to be useful [5]. In this section, first of all, we preprocess speech signals, such as pre-emphasis, framing (256 sampling points, 128 points frame-shift), adding-windows. After that we extract the feature parameters such as E, ZCR, F0, FF, MFCC, LPC from each frame signal and figure out their Max, Min, mean, Var., FOD. Moreover, there are two other features, the RC of F0, the RC of FF, 32 traditional features totally.

3.2. New Features

The method of speech emotion recognition based on the traditional features doesn't achieve good results. So we introduce two new features, MFCCF0 and APSAM, to improve the recognition rate.

We extract fundamental frequency parameters from each frame of speech signal by autocorrelation function method and obtain a fundamental frequency curve. Then median filtering of 5 points is adopted to smooth this curve and the points of fundamental frequency off tracking the curve would be deleted. At last we extract 4th order MFCC feature parameters from this processed curve which is our first new feature.

Next we extract SAM of each frame of speech signal and achieve a SAM curve. On the basis of this curve, amplitude perturbation parameters, which is used to describe the jitter level within a certain range, can be figured out. Amplitude perturbation parameters, such as amplitude jitter percentage (Shim), amplitude jitter (ShdB), amplitude perturbation quotient (APQ), their formulas are as follows:

$$Shim = \frac{1/(N-1)\sum_{i=1}^{N-1}\left|A^{i} - A^{(i+1)}\right|}{1/N\sum_{i=1}^{N-1}A^{i}} \tag{1}$$

$$ShdB = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20\log\left(A^{i+1}/A^{i}\right)\right| \tag{2}$$

$$APQ = \frac{1/(N-1)\sum_{i=1}^{N-10}\left|1/10\sum_{r=0}^{10}A^{(i+r)} - A^{(i+2)}\right|}{1/N\sum_{i=1}^{N-1}A^{i}} \tag{3}$$

where A represents short-time average magnitude, N represents numbers of short-time average magnitude, $i = 1, 2, \cdots, N$.

4. Feature Selection Based on BP Neural Network

In order to constitute the best set of features and reduce the dimensions of feature space, Ruck *et al.* come up with the sensitivity of the network outputs to its inputs which is used to rank the input features [6]. In the experiment, the network uses one hidden layer. The activating function of the hidden layer uses Sigmoid function, and the activating function of the output layer is a linear function. In the experiment, we let the number of the hidden nodes be 15. The set of 32 traditional features is used as the network inputs. After the network has been trained, the weights in the network are determined. Then the saliency values for each input were calculated. As each network is started with a different set of random weights, we take 10 trained network's saliency values to an average in order to improve accuracy. After 10 experiments, we obtain the results of ranking the average saliency values, which is showed in **Table 1**. The input with the highest saliency value, is ranked No.1 and the lowest is ranked No.32.

5. Results and Discussion

5.1. Recognition Based on Traditional Features

From the results (**Table 1**) of the saliency sequencing of 32 traditional features, we can see the first six traditional features arranged in descending order: E FOD, F0 mean, MFCC mean, ZCR FOD, LPC mean and E mean. We choose the first four, five, and six feature parameters respectively and recognize them with GMM. Recognition results are showed in **Table 2**, **Table 3**, and **Table 4**. For the different number of Gaussian mixture model (NGMM) results in considerable difference of the recognition rates, we experiment with many different numbers of Gaussian mixture model and show the most representative results as follows.

Table 1. A rank of the 32 traditional features.

No.	Feature	No.	Feature	No.	Feature
1	E FOD	12	F0 Var.	23	E Min
2	F0 mean	13	E Max	24	MFCC Max
3	MFCC mean	14	LPC FOD	25	LPC Max
4	ZCR FOD	15	ZCR Max	26	F0 Min
5	LPC mean	16	F0 RC	27	FF FOD
6	E mean	17	FF Var.	28	FF RC
7	FF Max	18	MFCC Var.	29	E Var.
8	F0 FOD	19	ZCR Min	30	FF Min
9	ZCR Var.	20	MFCC Min	31	MFCC FOD
10	ZCR mean	21	LPC Var.	32	LPC Min
11	FF mean	22	F0 Max		

Table 2. Recognition rates of the first four features (E FOD, F0 mean, MFCC mean, ZCR FOD).

NGMM	Speech emotion recognition rate (%)				average
	Fear	Happy	Neutral	Surprise	
3	66	48	81	62	64.25
4	78	70	69	75	73
5	74	67	73	50	66
6	79	71	69	69	72
8	73	63	64	67	66.75
12	69	54	66	67	64

Table 3. Recognition rates of the first five features (E FOD, F0 mean, MFCC mean, ZCR FOD, LPC mean).

NGMM	Speech emotion recognition rate (%)				average
	Fear	Happy	Neutral	Surprise	
3	76	56	72	79	70.75
4	82	73	85	75	78.75
5	82	71	87	79	79.75
6	76	73	88	74	77.75
8	76	77	84	71	77
12	69	65	82	70	71.5

Table 4. Recognition rates of the first six features (E FOD, F0 mean, MFCC mean, ZCR FOD, LPC mean, E mean).

NGMM	Speech emotion recognition rate (%)				average
	Fear	Happy	Neutral	Surprise	
3	76	69	77	64	71.5
4	77	72	86	78	78.25
5	73	76	81	79	77.25
6	69	77	87	80	78.25
8	75	74	83	70	75.5
12	60	59	84	74	69.25

This paper only chooses the first few features used to recognize, because of their greater contribution to the speech emotion recognition. As show in the three tables above, when the first five features (E FOD, F0 mean, MFCC mean, ZCR FOD, LPC mean) are combined, the average recognition rate reaches the highest 79.75%, in which the recognition rate of Fear reaches 82% and the recognition rate of Neutral reaches 87%. If we continue to add emotional features as the inputs, we will find that the rate of single recognition and average recognition all decrease. This proved that the five features involve considerable information to differentiate emotions. With the increase of selected features, redundant and irrelevant between the features increase, and the recognition rates of the speech emotion decrease [7].

5.2. Recognition Based on Traditional and New Features

Based on the feature selection above, two new features of speech emotion, MFCCF0 and APSAM, are added to the selected features [8]. The recognition results of the first five features (E FOD, F0 mean, MFCC mean, ZCR FOD, LPC mean) and MFCCF0 are showed in **Table 5**. The recognition results of the first five features and APSAM are showed in **Table 6**. And the recognition results of the first five features, MFCCF0 and APSAM are showed in **Table 7**.

Table 5. Adding MFCCF0 based on the first five features.

NGMM	Speech emotion recognition rate (%)				average
	Fear	Happy	Neutral	Surprise	
3	75	70	94	75	78.5
4	85	73	90	74	80.5
5	83	76	89	68	79
6	74	73	87	75	77.25
8	74	70	85	71	75
12	73	64	81	76	73.5

Table 6. Adding APSAM based on the first five features.

NGMM	Speech emotion recognition rate (%)				average
	Fear	Happy	Neutral	Surprise	
3	68	73	78	59	69.5
4	82	79	87	75	80.75
5	84	75	86	62	76.75
6	79	76	90	56	75.25
8	76	70	88	76	77.5
12	76	68	83	71	74.5

Table 7. Adding MFCCF0 and APSAM based on the first five features.

NGMM	Speech emotion recognition rate (%)				average
	Fear	Happy	Neutral	Surprise	
3	75	68	72	81	74
4	86	77	90	76	82.25
5	84	75	85	81	81.25
6	79	80	87	72	79.5
8	82	73	86	77	79.5
12	71	80	84	65	75

As show in **Table 5**, after adding MFCCF0, the aver rate reaches 80.5% [9]. It is because of speech emotion is in relation to F0. While reading the different text in different emotions, their fundamental frequency curves are various. Also, emotion has nothing to do with the text. Therefore, feature parameters extracted from the fundamental frequency curve can characterize some emotion information and raise the recognition rate [10]. As show in **Table 6**, after adding APSAM, the aver rate increases by 1.0% and reaches 80.75%. This is because of amplitude perturbation parameters can describe the jitter level within a certain range. Speech in different emotion causes different jitter level, so the feature could characterize emotional information to a certain extent and raise the recognition rates. From **Table 7**, we can see that, after adding the two new features, MFCCF0 and APSAM, the aver rate reaches 82.25%, increases by 2.5%, in which the Neutral gets 90% at its peak and the other three emotions all achieve 76% at least.

6. Conclusion

The method of feature selection based on BP neural network is not only convenient to choose the most effective ones in various traditional features, but also reduces the dimension of feature space [11]. The average rate reaches 79.75% while a set of 5 traditional features (E FOD, F0 mean, MFCC mean, ZCR FOD, LPC mean) is used to recognize speech emotion. Based on the feature selection, two new characteristics of speech emotion, MFCCF0 and APSAM, are added to the selected features (the first five features). With GMM, we get the highest average recognition rate of the four emotions 82.25%, and the recognition rate of Neutral 90%. According to the experiment result, the two new features can improve the recognition rate of speech emotion, because they can characterize some new emotion information.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61062011, No. 61362003) and GuangXi Key Lab of Multi-source Information Mining & Security, Electronic Engineering College, Guangxi Normal University.

References

[1] Ayadi, M.E., Kamel, M.S. and Karray, F. (2011) Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases. *Pattern Recog.*, **44**, 572-587. http://dx.doi.org/10.1016/j.patcog.2010.09.020

[2] Sultana, S., Shahnaz, C., Fattah, S.A., *et al.* (2014) Speech Emotion Recognition Based on Entropy of Enhanced Wavelet Coefficients. 2014 *IEEE International Symposium on Circuits and Systems* (*ISCAS*), Melbourne VIC, 1-5 June 2014, 137-140. http://dx.doi.org/10.1109/ISCAS.2014.6865084

[3] Zheng, W.Q., Yu, J.S. and Zou, Y.X. (2015) An Experimental Study of Speech Emotion Recognition Based on Deep Convolutional Neural Networks. 2015 *International Conference on Affective Computing and Intelligent Interaction* (*ACII*), Xi'an, 21-24 September 2015, 827-831. http://dx.doi.org/10.1109/ACII.2015.7344669

[4] Wang, K.X., An, N., Li, B.N., Zhang, Y.Y. and Li, L. (2015) Speech Emotion Recognition Using Fourier Parameters. *IEEE Transactions on Affective Computing*, **6**, 69-75. http://dx.doi.org/10.1109/TAFFC.2015.2392101

[5] Bouwmans, T., Baf, F.E. and Vachon, B. (2008) Background Modeling Using Mixture of Gaussians for Foreground Detection. *Recent Patents on Computer Science*, 219-237. http://dx.doi.org/10.2174/2213275910801030219

[6] Ruck, D.W., Rogers, S.K. and Kabrisky, M. (1990) Feature Selection Using a Multilayer Perceptron. *Journal of Neural Network Computing*, **2**, 40-48.

[7] Devijver, P.A. and Kittler, J. (1982) Pattern Recognition: A Statistical Approach. Prentice Hall International, London.

[8] Vlasenko, B., Schuller, B., Wendemuth, A. and Rigoll, G. (2007) Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing. *Proceedings of the* 2*nd International Conference on Affective Computing and Intelligent Interaction (ACII'*07), 139-147.

[9] Yuan, Y.J., Zhao, P.H. and Zhou, Q. (2010) Research of Speaker Recognition Based on Combination of LPCC and MFCC. *Proc. IEEE Int. Conf. Intell. Comput. Intell. Syst.*, **3**, 765-767. http://dx.doi.org/10.1109/icicisys.2010.5658337

[10] Li, Y. and Zhao, Y.X. (1998) Recognizing Emotions in Speech Using Short-Term and Long-Tern Features. *Porc. ICSLP*, Sydney, Australian, 2255-2258.

[11] Dellaert, F., Polzin, T. and Waibel, A. (1996) Recognizing Emotion in Speech. *Proceedings of ICSLP*, 1970-1973. http://dx.doi.org/10.1109/icslp.1996.608022