Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Characteristics of human auditory model based on compensation of glottal features in speech emotion recognition

Sun Ying *, Zhang Xue-Ying

College of Information Engineering, Taiyuan University of Technology, Shanxi Province Jinzhong CityYuci District College Town, 030600, China

HIGHLIGHTS

- The paper has introduced the glottal features into speech emotion recognition.
- The results show that GCZCMT feature effectively distinguishing emotional state.
- It can be seen that GCZCMT has high practical value.

ARTICLE INFO

Article history:
Received 19 June 2017
Received in revised form 7 September 2017
Accepted 1 October 2017
Available online 11 November 2017

Keywords:
Glottal compensation
Human auditory model
Emotion recognition
Speech signal
Language bank
Features

ABSTRACT

The speech signal carries emotional message during its production. With the analysis on relation between sound production and glottis, the paper has introduced the glottal features into speech emotion recognition, proposed the model where the glottis is used for compensation of glottal features, and extracted the feature of Glottal Compensation to Zero Crossings with Maximal Teager Energy Operator (GCZCMT). Two experiments have been designed, including that: firstly, the single emotional speech databases of TYUT and Berlin are respectively used for experiment (the purpose of such experiment is to research the emotion recognition capability of GCZCMT feature, and the experimental results show that GCZCMT feature is a feature possibly and effectively distinguishing emotional state); secondly, this experiment is one of mixing speech database (the purpose of such experiment is to research the emotion recognition capability of GCZCMT feature on ross-database language, and the experimental results show that the database dependency of GCZCMT feature is the minimum, and such feature is more suitable for actual complex language environment, and has the higher practical value.

1. Introduction

Speech emotion recognition is a technology [1] established based on intensive research and analysis on production mechanism of speech signal, extracting and collecting feature parameters expressing emotion among speech signals, and taking advantage of these parameters for corresponding modeling and recognition so as to confirm speech emotion state. Emotional feature extraction is one of key technologies for speech emotion recognition, and the feature based on human auditory model and features based on speech glottis, such as fundamental frequency, voice rate and power, etc., have been widely concerned in recent years. For example, the typical application in speech emotion recognition of MFCC feature researched by Chandni, et al. of Indian researchers has obtained favorable recognition results [2]. ZCMT feature based on human auditory model proposed by Sun Ying et al. from Taiyuan University of Technology has obtained good results in both isolated words recognition and speech emotion recognition [3]. Juan Pablo Arias of Chilean researcher detected speech emotion type [4] by establishing the fundamental frequency model of neutral emotion speech. Zhao Li and Huang Chengwei, et al. of researchers selected and evaluated the correlation between features (including fundamental frequency, energy and resonance peak, etc.) and emotion dimension and speech emotion feature of words used, such as irritability, and proposed the practical non-judgment speech emotion recognition method [5] for actual application environment. Although researchers have comprehensively and respectively researched glottal feature and auditory feature, the correlation between glottal feature and auditory feature has not been considered. Speech signal carries a large number of emotional messages during its production, and the glottal feature is the specific performance of sound production process. Therefore, the combination between glottal feature and auditory feature will better satisfy the physiological process of speech production, communication and acceptance.

* Corresponding author.
E-mail address: sunyingsys@hotmail.com (S. Ying).

https://doi.org/10.1016/j.future.2017.10.002

Aiming at the relation between nonlinear feature and human auditory model presenting in the sound production process, the paper makes the further research. Firstly, the connection between feature generation of sound production process and vocal tract of glottis is analyzed; next, the human auditory model for compensation of glottal features is proposed; finally, emotional words and sentences in two databases of TYUT database and Berlin voice database are selected and used, and as the comparison of results between independent emotion speech recognition experiment and mixed independent emotion speech recognition experiment with three features (including typical human auditory model MFCC, uncompensated human auditory model ZCMT feature, compensated ZCMT (GCZCMT)), it is verified that GCZCMT has both preferable capability of distinguishing emotional state and the higher recognition rate in cross-database emotion recognition, and the independence of database is good. It has the good application value in the actual language environment.

2. Basic theory of glottal feature

2.1. Model of speech production

In 1980s, with research, Teager found that the nonlinear airflow vortex [6] would be generated before speech production. Later, such discovery was verified by the experiment with fluid in dynamic mechanical model of vocal tract and glottis. The research of Zhuo [7] et al. on emotion classification indicated that: the sound produced with airflow vortex form could be regarded as a part of sound source in the emotional state of being “angry” and “nervous”. Such sound generated with airflow vortex form is very sensitive to emotional state of the speaker.

Fig. 1 is the schematic diagram of nonlinear model under speech production. From Fig. 1, it can be seen that the speech signal consists of plane-wave linear part and nonlinear part of vortex area. Airflow of trachea is differentiated into linear airflow and vortex after running through the glottis. Linear airflow is spread with plane wave in the vocal tract, while vortex airflow firstly has interaction with vocal tract wall and then is spread with plane wave. According to different locations of generation of vortex airflow, the supraglottal and intraglottal vortex airflows could be divided. In early stage of opening vocal cords, the glottis will be shrunk, and the supraglottal vortex is produced at the upward side of vocal cords, In the later state of closing vocal cords, the glottis will be stretched, and the intraglottal vortex is formed in vocal cords. The intraglottal vortex produced by vocal cords with symmetric vibration will result in negative pressure. Such pressure possibly urges the vocal cords to rapidly close. Under the comparison with unsymmetrical vocal cords, the process of rapid closing will change signal energy spectrum in acoustics mechanism. On the other hand, when the supraglottal vortex has strong collision with vocal tract wall or they have interaction with each other, a part of sound source will be generated. Such sound source will finally change signal energy spectrum of speech. From the model under speech production, it can be found that the whole process has close connection with the glottis, and both supraglottal vortex and intraglottal vortex will finally have influence on speech signal. Typical glottal features include fundamental frequency, voice rate and power, etc. [8–10]. Glottal feature used in the paper is the fundamental frequency, and to accurately extract the fundamental frequency, HPS is taken as extraction method of fundamental frequency.

2.2. Extraction method of fundamental frequency

Set $A(f)$ as the range of signal frequency spectrum, f_0 as fundamental frequency and $f_{\max}$ as the maximum frequency of $A(f)$, and

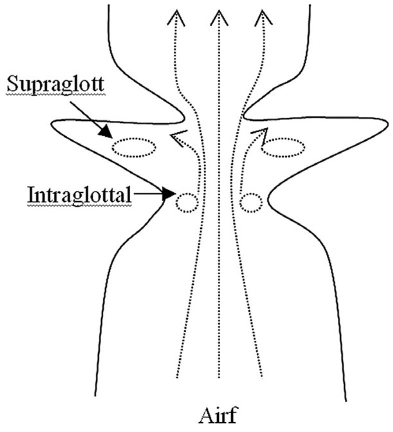

Fig. 1. Nonlinear model for the production of speech.

then define harmonic range as:

$$SH = \sum_{n=1}^{N} A(nf_0) \tag{1}$$

where, N is the maximum harmonic number in frequency spectrum, and $A(f) = 0$

Define subharmonic range SS as:

$$SS = \sum_{n=1}^{N} A\left(\left(n - \frac{1}{2}\right) f_0\right) \tag{2}$$

Then, the ratio of subharmonic to harmonic SHR could be obtained through dividing harmonic range SH by subharmonic range SS:

$$SHR = \frac{SH}{SS} \tag{3}$$

With method in the literature [11], transform linear frequency in the formula (2) into log domain, and it is supposed that $LOGA(\cdot)$ represents logarithmic frequency, and:

$$SH = \sum_{n=1}^{N} LOGA\left(\lg(nf_0)\right) = \sum_{n=1}^{N} LOGA\left(\lg(n) + \lg(f_0)\right) \tag{4}$$

$$SS = \sum_{n=1}^{N} LOGA\left(\lg\left(n - \frac{1}{2}\right) + \lg(f_0)\right) \tag{5}$$

To obtain SH, leftward shift the frequency spectrum of even number along the horizontal axis of logarithm, and then: $\lg(2)$, $\lg(4)$, ..., $\lg(4N)$. Calculate the sum of frequency spectrum after shifting:

$$SUMA(\lg f)_{even} = \sum_{n=1}^{2N} LOGA\left(\lg f + \lg(2n)\right) \tag{6}$$

For $LOGA(\lg f) = 0$ when $f > f_{\max}$, obtain the following from the formula (4) and the formula (6):

$$SUMA\left(\lg\left(\frac{1}{2} \cdot f_0\right)\right)_{even} = SH \tag{7}$$

$$SUMA\left(\lg\left(\frac{1}{4} \cdot f_0\right)\right)_{even} = SH + SS \tag{8}$$

Similarly, leftward shift $\lg(1), \lg(3), \ldots, \lg(4N-1)$, and obtain:

$$SUMA(\lg f)_{odd} = \sum_{n=1}^{2N} LOGA\left(\lg f + \lg(2n-1)\right) \tag{9}$$

$$SUMA\left(\lg\left(\frac{1}{2}\cdot f_0\right)\right)_{odd} = SH \tag{10}$$

$$SUMA\left(\lg\left(\frac{1}{4}\cdot f_0\right)\right)_{odd} = \Delta \tag{11}$$

Δ represents the sum of value of $\lg(nf_0) + \lg\left(\frac{1}{4}\cdot f_0\right)$

Define the equation

$$DA(\lg f) = SUMA(\lg f)_{even} - SUMA(\lg f)_{odd} \tag{12}$$

From the formulas (7), (8), (10) and (11), obtain:

$$DA\left(\lg\left(\frac{1}{2}\cdot f\right)\right) = SH - SS \tag{13}$$

$$DA\left(\lg\left(\frac{1}{4}\cdot f\right)\right) = SH + SS - \Delta \tag{14}$$

In normal speech, $SS \approx 0$. If the specification of subharmonic is the main part, for $\Delta \approx 0$ and the maximum value of $DA(\cdot)$ is possibly approximate to $\lg\left(\frac{1}{4}\cdot f_0\right)$, the second maximum value is $\lg\left(\frac{1}{2}\cdot f_0\right)$. *SHR* could be approximately calculated with the formula (13) and the formula (14).

$$\frac{DA\left(\lg\left(\frac{1}{4}\cdot f_0\right)\right) - DA\left(\lg\left(\frac{1}{2}\cdot f_0\right)\right)}{DA\left(\lg\left(\frac{1}{4}\cdot f_0\right)\right) + DA\left(\lg\left(\frac{1}{2}\cdot f_0\right)\right)} = \frac{SS - \frac{1}{2}\cdot\Delta}{SH - \frac{1}{2}\cdot\Delta} \approx SHR \tag{15}$$

When the maximum value is being found, firstly define the global maximum value, and record it as $\lg(f_1)$. Then, from such point, find the next maximum value $\lg(f_2)$ in the section of $[\lg(1.9375f_1), \lg(2.0625f_1)]$.

When two peak values are confirmed, *SHR* is determined by the formula (16):

$$SHR = \frac{DA(\lg(f_1)) - DA(\lg(f_2))}{DA(\lg(f_1)) + DA(\lg(f_2))} \tag{16}$$

Express the fundamental frequency as:

$$f_0 = \begin{cases} 2f_2, SHR < Thr \\ 2f_1, SHR \geq Thr \end{cases} \tag{17}$$

From the above formula, it is indicated that when *SHR* is lower than a certain threshold value *Thr*, the subharmonic is weak, and the harmonic shall be concerned. At such moment, the fundamental frequency is $2f_2$; otherwise, the fundamental frequency is $2f_1$. Generally select the threshold value from values in the section of [0.2, 0.4].

2.3. Features of human auditory model

Human auditory feature used in the paper is: the feature of Zero Crossings with Maximal Teager Energy Operator (ZCMT) based on human auditory feature model. The basic idea of ZCMT feature is to use a group of band-pass filter to simulate the selection function of human ear on frequency, zero-crossing rate for representation of frequency information and speech rate of language [12] and Teager energy operator to calculate Teager energy between two sampling points, and finally combine the above features and obtain final ZCMT feature (see Fig. 2).

Specific experiment steps are shown as follows:

(1) Make frame processing for original emotion speech, and the frame size is divided into 110 sampling points, and frame shift has 55 sampling points.

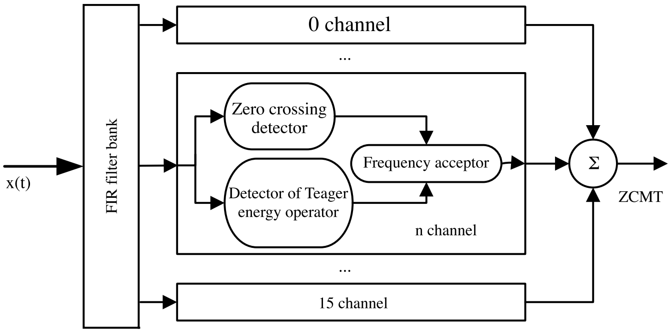

Fig. 2. Diagram of ZCMT feature extraction.

(2) Transmit speech with framing to 16 band-pass filters for filtering.

(3) Frequency feature is extracted with zero crossing detector.

(4) Find the maximum sampling point of absolute value between adjacent rising zero crossing points with Teager energy operator detector, and denote it as $x[n]$. $x[n+1]$ and $x[n-1]$ are the next sampling point value and the previous sampling point value of the maximum sampling point of absolute value. Use the following formula:

$$\psi[x(n)] = x^2(n) - x(n-1)\,x(n+1)$$

To calculate the Teager energy operator of the maximum sampling point of absolute value.

(5) Integrate frequency feature with Teager energy operator of the maximum sampling point of absolute value in the frequency receiver. Set Z_k as the rising zero crossing point belonging to the signal $x_k(t)$ output by the kth filter, and express P_{kl} as the maximum sampling point value of the absolute value between the lth zero crossing point and the $l+1$th zero crossing point of $x_k(t)$; then, the output of frequency receiver at the time of m is shown as the formula (18):

$$y(m, i) = \sum_{k=1}^{N_{ch}} \sum_{l=1}^{Z_k - 1} \delta_{ij_l} \Psi(P_{kl}) \; 1 \leq i \leq N \tag{18}$$

where, N_{ch} is the number of channel of filter bank. δ_{ij_l} is the Kronecker operator, and the sign j_l is the frequency index of frequency receiver of each channel; therefore, the frequency receiver has obtained the corresponding frequency and the maximum range in a certain period of time. The feature $zcmt(m, i)$ of the whole speech signal could be output through combination of frequency receiver of 16 channel.

(6) Regulate the time and range of integrated feature, and regard it as output of complete feature, and such feature is the feature of zero crossing Teager energy operator.

3. Feature of human auditory model on compensation of glottal features

It is assumed that $X(t)$ and $Y(t)$ are respectively human auditory model feature and corresponding glottal feature in a certain speech signal, and ΔT is frame length. Then, $\vec{x}(t)$ and $\vec{y}(t)$ are estimated values respectively of human auditory model feature and glottal feature in the tth frame of speech signal. In the paper, select ZCMT feature as human auditory feature.

Each corresponding time is $\vec{x}(t)$. It assumed that the vector compensated by mean value of glottis is $\vec{x}_{opt}(t)$, then the relation between $\vec{x}(t)$ and $\vec{x}_{opt}(t)$ could be expressed with function of

$\vec{y}\,(t)$, and the human auditory model without compensation is:

$$\vec{x}\,(t) = \Psi\left(\vec{x}_{opt}\,(t)\,,\Theta\left(\vec{y}\,(t)\right)\right) \tag{19}$$

where, function Ψ shows a mathematic relation of two variables, which might be linear relation, sine relation, polynomial relation or more complicated relation, and function Θ shows a transformation form of fundamental frequency.

Define $I_k, k = 1, 2, \ldots, N$ as continuous fundamental frequency subspace, N as the number of fundamental frequency subspace, and there is $I_1 \cup I_2 \cup \cdots \cup I_N = [50\text{ Hz}, 550\text{ Hz}]$. Then, the set cover of all subspaces generally refers to the whole scope of fundamental frequency. $E\,(\cdot)$ is the mathematical expectation. For the fundamental frequency is one-dimensional vector, possibly replace $\vec{y}\,(t)$ with $y\,(t)$.

$$\Theta\left(\vec{y}\,(t)\right) = \Theta\,(y\,(t)) = \Theta\left(E\left(\hat{Y}_{I_k}\,(t)\right)\right) \tag{20}$$

$$\hat{Y}_{I_k}(t) = \{y(t)|y(t) \in I_k\} \tag{21}$$

For different fundamental frequencies have different influences on human auditory model and the influence of fundamental frequencies falling in same fundamental frequency section is same as behavior of long-time mean value of frequency frequencies in such section, interpret the function Θ of fundamental frequency as the distance $D\,(\cdot)$ between variable z and mathematical expectation $E\,(z)$, and then:

$$\begin{aligned}\Theta\left(\vec{y}\,(t)\right) &= \Theta\left(E\left(\hat{Y}_{I_k}\,(t)\right)\right)\\ &= D\left(E\left(\hat{Y}_{I_k}(t)\right), E\left(\hat{Y}\,(t)\right)\right)\end{aligned} \tag{22}$$

Substitute the formula (19), and obtain:

$$\begin{aligned}\vec{x}\,(t) &= \Psi\left(\vec{x}_{opt}\,(t)\,,\Theta\left(\vec{y}\,(t)\right)\right)\\ &= \Psi\left(\vec{x}_{opt}(t), D\left(E\left(\hat{Y}_{I_k}(t)\right), E\left(\hat{Y}(t)\right)\right)\right)\end{aligned} \tag{23}$$

Finally, the vector $\vec{x}_{opt}\,(t)$ of emotion feature could be expressed as:

$$\begin{aligned}\vec{x}_{opt}\,(t) &= \Psi^{-1}\left(\vec{x}\,(t)\,,\Theta\left(\vec{y}\,(t)\right)\right)\\ &= \Psi^{-1}\left(\vec{x}(t), D\left(E\left(\hat{Y}_{I_k}(t)\right), E(\hat{Y}(t))\right)\right)\end{aligned} \tag{24}$$

Set $\Psi\,(\cdot)$ as linear equation $y = x + a$ and $D\,(\cdot)$ as the distance. Then, the formula (24) could be expressed as the formula (25), where $\vec{\alpha}$ is vector with same order of that of $\vec{x}\,(t)$, and call it as impact factor.

$$\vec{x}_{opt} = \vec{x}\,(t) - \frac{\vec{\alpha}\left|E\left(\hat{Y}_{I_k}\,(t) - E\left(\hat{Y}\,(t)\right)\right)\right|}{\left|E\left(\hat{Y}\,(t)\right)\right|} \tag{25}$$

4. Experiment and analysis

4.1. Emotion speech database

Emotion speech database is the precondition for analysis on emotion speech and emotion recognition, and provides speech data of training and testing for emotion recognition. To objectively and comprehensively evaluate performance of researched nonlinear feature based on speech chaos feature and simultaneously consider influence of different languages on emotion feature recognition result, the paper has selected and used TYUT speech database and Berlin speech database as database used for experiment.

TYUT emotion speech database includes two languages of English and Chinese mandarin, three emotional states of "being happy", "being angry" and "being neutral" (883 words [13] in total). In view of experimental accuracy and universality, such experiment has randomly selected 91 words from each kind of emotion of TYUT emotion speech database for emotion speech recognition, two thirds of which (61 words) are used for training and one third of which (30 words) are used for testing.

Berlin emotion speech database consists of German, including seven kinds of emotions of "being neutral", "being angry", "being fear", "being happy", "being sad", "being disgusted" and "being agitated" (535 words in total). Berlin emotion speech database has four kinds of emotions not existing in TYUT emotion speech database; therefore, when recognition experiment, these four kinds of emotions shall be eliminated. The experiment selects 71 words from each kind of emotion of Berlin emotion speech database for emotion speech recognition, two thirds of which (48 words) are used for training and one third of which (23 words) are used for testing.

4.2. Experiment steps

The experiment is divided into two steps of training and testing, and the training process is shown as follows:

(1) Extract 1024-dimension ZCMT feature for word for training, and denote it as $\vec{x}\,(t)$ and fundamental frequency feature F_0 as $\vec{y}\,(t)$;

(2) Average divide fundamental frequency space [50 Hz, 550 Hz] into 10 subsections, and the length of each section is 50 Hz;

(3) As for each subspace, make the statistics on ZCMT features of all fundamental frequencies in the scope of such subspace, and calculate the mathematical expectation $E\left(\hat{Y}_{I_k}\,(t)\right)$ of fundamental frequency $\vec{y}\,(t)$ of ZCMT feature $\vec{x}\,(t)$ in such subspace;

(4) Calculate the mth dimension value $\vec{\alpha}\,(m)$ of $\vec{\alpha}$, and given that the value of ZCMT vector of all dimensions outside the mth dimension is not changed, and then:

$\vec{\alpha} = [0, 0, \ldots, \overbrace{-2.0 + 0.1 * k}^{\text{第}m\text{阶}}, 0, 0]$ (the mth order). Where, $k = 1, 2, \ldots, 40$. The value of ZCMT vector of all dimensions outside the mth dimension is not changed. Here, k is the empirical value 16;

(5) As for certain k value, calculate ZCMT feature under compensation with the formula (25), and denote it as human auditory model feature of compensation of glottal feature output;

(6) Train SVM model with human auditory model feature on training of compensation of glottal feature, and obtain the trained SVM model;

(1) This is similar to training process. Respectively extract human auditory model feature on compensation of glottal feature with all words for testing;

(2) Input human auditory model feature extracted on compensation of glottal feature into the trained SVM model, and obtain recognition results;

4.3. Experiment results and analysis

(1) Experiment for one language

This experiment is divided by language categories, and is to recognize three kinds of emotions for two speech databases and three kinds of languages. For example, when the experiment on Chinese word of "being happy" is made, the words for training and testing are both Chinese word of "being happy" in TYUT speech database. Table 1 includes respective recognition results of five features.

From Table 1, it can be known that:

Table 1
Recognition rate of glottal compensation to human auditory model features for single database experiments (%).

Feature	Emotion	TYUT		Berlin	Average
		Mandarin	English	German	
GCZCMT	Happy	90.00	83.33	75.28	82.87
	Angry	90.00	86.67	83.25	86.64
	Neutral	96.67	90.00	86.96	91.21
	Average	92.22	86.67	81.83	86.91
ZCMT	Happy	86.67	76.67	73.91	79.08
	Angry	90.00	86.67	52.17	76.28
	Neutral	96.67	90.00	86.96	91.21
	Average	91.11	84.45	71.01	82.19
MFCC	Happy	83.33	83.33	82.61	83.09
	Angry	90.00	83.33	65.22	79.52
	Neutral	96.67	96.67	91.30	94.88
	Average	90.00	87.78	79.71	85.83

Firstly, regardless of which database (TYUT emotion speech database or Berlin emotion speech database) is used, the recognition performance on emotions of "being happy" and "being angry" is improved after GCZCMT feature is used. Performance on emotion of "being happy" has been increased by 3.79%; that on emotion of "being angry" has been increased by 10.36%; that on emotion of "being neutral" has no change. Especially, under the emotional state of "being angry" with German, the recognition rate has been increased by 31.08%.

Secondly, under comparison with typical MFCC feature, the average recognition rate after compensation is higher than that of MFCC feature (85.83%). Therefore, it can be considered that the compensation algorithm of such glottis on human auditory model feature is a relatively reasonable algorithm. Such algorithm both maintains advantages of modeling based on human auditory model of ZCMT feature, and has merits of glottal feature introduced. This feature could better grasp acoustic essence of different emotion speeches; therefore, good recognition results will be obtained.

(2) Experiment on mixed emotion speech database

This experiment is to mix words with same emption of TYUT speech database and Berlin speech database, and then recognize three kinds of emotions for three languages. For example, when the experiment on Chinese is made, the word for training refers to all training words on emotion of "being happy" of two speech databases, which means that the number of word for training is $61 + 61 + 48 = 170$ words, while the word for testing refers to all testing words on emotion of "being happy" of TYUT speech database, which means that the number of it is 30 words. To make the comparison with algorithm before compensation, words for testing and training are same. Table 2 includes respective recognition results of five features.

From Table 2, it can be seen that: after use of GCZCMT algorithm, the average recognition rate of mixed database experiment has been reduced by 2.06% than that of independent database experiment, and the recognition rates of "being angry" and "being neutral" of mixed database experiment has respectively been reduced by 2.18% and 3.97% than these of independent database experiment; in addition, the emotion of "being happy" has no change. However, under same experiment conditions, the average recognition rate of MFCC has been reduced by 3.5%, while that of ZCMT has been only reduced by 1.45%. From cross-database experiment, we can see that the database independence of GCZCMT algorithm is superior to that of typical MFCC feature, but inferior to that of ZCMT feature. For the paper has only made the preliminary research on compensation algorithm, the fundamental frequency is taken as glottal feature, and the linear equation is regarded as compensation algorithm. This is main reason why the result of mixed database experiment is inferior to ZCMT feature.

Summarily, the advantage of GCZCMT feature is more obvious than that of ZCMT feature and MFCC feature in independent emotion database, and although the recognition rate of it is inferior to that of ZCMT feature in mixed emotion database, the difference of both is not great. Therefore, it is verified that ZCMT is the effective feature for distinguishing emotions. The method helping glottal feature for compensation on human auditory model and unifying sound production, spreading and receiving process is feasible. In addition, the complexity of actual language environment makes cross-database research more practical. Therefore, the emphasis for the next research will include finding out more appropriate glottal feature and compensation algorithm.

5. Conclusion

With two experiments, it can be found that the human auditory model feature based on compensation of glottal feature is a relatively ideal emotion speech recognition feature. Such feature has shown favorable recognition performance in independent emotion speech recognition experiment, and also presented preferable database independence in mixed speech database experiment. Here, the paper describes the relation between glottal feature and human auditory model feature only with the simplest linear function, and in such manner, the perfect effect has also been obtained. If more functions could be used for respective experiment and analysis on experiment results, the system is possibly further optimized.

Acknowledgment

Authors acknowledge National Natural Science Foundation of China (No. 61371193) for the support.

Table 2
Recognition rate of glottal compensation to human auditory model features for merged databases experiments (%).

Feature	Emotion	TYUT		Berlin	Average
		Mandarin	English	German	
ZCMT after compensation	Happy	90.00	83.33	75.28	82.87
	Angry	90.00	83.33	80.04	84.46
	Neutral	90.00	90.00	81.71	87.24
	Average	90.00	85.55	79.01	84.85
ZCMT	Happy	86.67	80.00	73.91	80.19
	Angry	90.00	83.33	39.13	70.82
	Neutral	93.33	93.33	86.96	91.21
	Average	90.00	85.55	66.67	80.74
MFCC	Happy	83.33	73.33	47.83	68.16
	Angry	90.00	76.67	93.91	86.86
	Neutral	93.33	96.67	85.96	91.99
	Average	88.89	82.22	75.90	82.33

References

[1] A.I. Iliev, M.S. Scordilis, J. Papa, et al., Spoken emotion recognition through optimum-path forest classification using glottal features, Comput. Speech Lang. 24 (3) (2010) 445–460.

[2] S. Ramakrishnan, I.M.M.E. Emary, Speech emotion recognition approaches in human computer interaction, Telecommun. Syst. 52 (3) (2013) 1467–1478.

[3] L. He, M. Lech, J. Zhang, et al., Study of wavelet packet energy entropy for emotion classification in speech and glottal signals, in: International Conference on Digital Image Processing, Vol. 2013, International Society for Optics and Photonics, 2013, pp. 2714–2739.

[4] D. Ververidis, C. Kotropoulos, Emotional speech recognition: Resources, features, and methods, Speech Commun. 48 (9) (2006) 1162–1181.

[5] H. Muthusamy, K. Polat, S. Yaacob, Improved emotion recognition using Gaussian mixture model and extreme learning machine in speech and glottal signals, Math. Probl. Eng. 2015 (6) (2015) 1–13.

[6] A.I. Iliev, M.S. Scordilis, Spoken emotion recognition using glottal symmetry, EURASIP J. Adv. Signal Process. 2011 (1) (2011) 1–11.

[7] R. Sun, E. Moore, Investigating glottal parameters and teager energy operators in emotion recognition, in: Affective Computing and Intelligent Interaction - Fourth International Conference, Acii 2011, Memphis, Tn, Usa, October 9-12, 2011, Proceedings, Vol. 2011, DBLP, 2011, pp. 425–434.

[8] K. Mohan Kudiri, A. Md Said, M.Y. Nayan, Emotion detection using relative amplitude-based features through speech, in: IEEE International Conference on Control System, Computing and Engineering, Vol. 2012, IEEE, 2012, pp. 115–118.

[9] S.G. Koolagudi, K.S. Rao, Emotion recognition from speech using source, system, and prosodic features, Int. J. Speech Technol. 15 (2) (2012) 265–289.

[10] J. Pohjalainen, T. Raitio, S. Yrttiaho, et al., Detection of shouted speech in noise: human and machine, J. Acoust. Soc. Am. 133 (4) (2013) 2377-2389;
R. Sun, E. Moore, J.F. Torres, Investigating glottal parameters for differentiating emotional categories with similar prosodics, in: IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 2009, IEEE Computer Society, 2009, p. 4509-4512.

[11] C. Gobl, Ailbhe N. Chasaide, The role of voice quality in communicating emotion, mood and attitude, Speech Commun. 40 (1–2) (2003) 189–212.

[12] S.J.L. Mozziconacci, Modeling emotion and attitude in speech by means of perceptually based parameter values, User Model. User-Adapt. Interact. 11 (4) (2001) 297–326.

[13] C.K. Yogesh, M. Hariharan, R. Ngadiran, et. al, A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal, Expert Syst. Appl. 69 (2016) 149–158.

Sun Ying received her Ph.D. degree in engineering from Taiyuan University of Technology in Shanxi, China. She is currently a lecturer in Taiyuan University of Technology. Her research interest is mainly in the area of Speech Signal Processing and Speech Emotion Recognition. She has published more than 10 papers in periodicals of national level and international conferences. About 4 papers were indexed by SCI, EI or ISTP. Two of province research projects have finished by her supervising.

Zhang Xueying received her Ph.D. degree in underwater acoustic engineering from Harbin Engineering University in 1997. She is currently a professor in Taiyuan University of Technology. Her research interest is mainly in the area of auditory model and robustness speech recognition, emotional speech recognition, low-speed and broadband speech coding, embedded system. She has published more than 100 papers in periodicals of national level and international conferences. About 30 papers were indexed by SCI, EI or ISTP. A lot of national, province or ministry research projects have finished by her supervising.