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
0167-739X/© 2017 Elsevier B.V. All rights reserved.

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

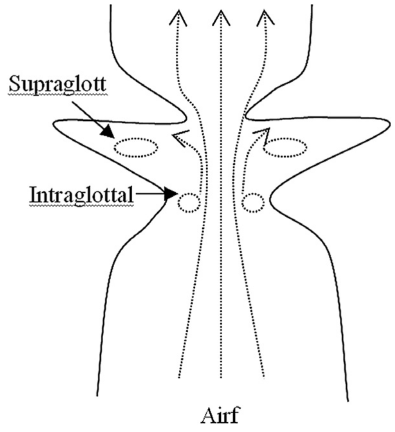

Fig. 1. Nonlinear model for the production of speech.

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:

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

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

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:

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

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

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

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