Research of Speech Biomarkers for Stress Recognition Using Linear and Nonlinear Features

2021 7th International Conference on Computer and Communications (ICCC) | 978-1-6654-0950-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCC54389.2021.9674330

1st Na Li
dept. School of Mathematics and Computer Science
Northwest Minzu University
Lanzhou, China
288142478@xbmu.edu.cn

2nd Nan Li*
dept. Information Center
The 940th Hospital of Joint Logistic Support Force of Chinese People's Liberation Army
Lanzhou, China
317629537@qq.com

3rd Min Guo
dept. School of Mathematics and Computer Science
Northwest Minzu University
Lanzhou, China
guomin0230@163.com

4th Jindong Feng
dept. School of Mathematics and Computer Science
Northwest Minzu University
Lanzhou, China
jindong81@xbmu.edu.cn

***Abstract*—Psychological stress increases the risk of a number of mental disorders such as depression. Stress detection at early time is of great significance for the prevention of mental disorders. This speech-based research is to explore the effective biomarkers for stress detection. Linear and nonlinear features including pitch, short-term energy, sub-band energy ratio, teager energy operator (TEO), formant, mel-frequency cepstral coefficients (MFCC) were extracted to analyze speech under stress. The results showed that pitch, jitter and short-term energy variance can be effective biomarkers for stress recognition.**

Keywords—Speech, Psychological Stress, Linear Features, Nonlinear Features

I. INTRODUCTION

Psychological stress is considered to be one of the important factors affecting human health. Related research has shown that psychological stress increases a risk of a number of mental disorders such as depression [1]. Mild depression affects the normal life of people, but even worse is major depression with high suicide risk. Depression and other mental disorders with low detection rate are difficult to cure. Hence, detecting psychological stress at early time with objective and ubiquitous tools is of great significance for the prevention of mental disorders.

Nowadays the recognition of psychological stress mostly depends on scales, such as Perceived Stress Scale(PSS) [2], Relative Stress Scale(RSS)[3], Psychological Stress Measure(PSM)[4]. These assessment methods bring a risk of subjective bias. In recent years, many researchers have proposed useful tools to detect psychological stress objectively including hormone, functional magnetic resonance imaging (fMRI), electroencephalography (EEG) and so on [5] [6] [7]. Although, fMRI with high spatial resolution and EEG with high time resolution are objective measurements for stress recognition, the devices are expensive. Hormonal methods are not widely available because of their invasiveness. As an inexpensive, objective and non-invasive measurement, speech is an effective tool for recognition of mental disorders. Speech, the most common way of communication, contains both semantic information and emotional information. Relative researches have proved that voice production was influenced by one person's emotion and cognition. [8] [9].

In the field of research on speech and psychological stress, numbers of linear and nonlinear features have been verified correlation with psychological stress. Researchers focused more on changes of pitch under stress, and increased pitch frequency has been verified under stress task in some studies [10]–[12]. In [13]researchers have detected that pitch did carry affect information. Additionally, other features have also been validated biomarkers for stress recognition. Tony W. et al. detected stress can lead to more pause time during speech [14]. John H. L.et al. extracted TEO-based features for stress recognition, achieving a classification accuracy of 95.30% [15]. Sumitra Shukla et al. discovered that both relative formant peak displacement(RFD) and mel-frequency cepstral coefficients(MFCC) can obtain good performance [16].

Existing studies have verified some speech features for stress detection, nevertheless exploring effective and reliable biomarkers for automatic psychological stress detection still needs further research. In this study, we extracted several linear and nonlinear features, including pitch, jitter, short-term energy, sub-band energy ratio, TEO, formant, MFCC after data preprocessing. And then classification and statistical analysis were performed to explore which speech features can be effective biomarkers for psychological stress recognition. It was found that pitch, jitter, energy did play an important role in stress recognition. Pitch, jitter and short-term energy variance showed a significant increase under stress both in male and female speech. Furthermore, relatively high accuracy was obtained in classification between groups.

The structure of this paper was as follows: Section 2 presented the materials and methods including participants, speech recording, data preprocessing, feature extraction, classification and statistical analysis. Section 3 showed the experimental results and discussions. The last section was conclusions.

II. MATERIALS AND METHODS

A. Participants

30 students (17 males, 13 females) aged from 18 to 24 were recruited from Lanzhou University and volunteered to take part in this study. All the subjects did not have prior history of related diseases or taking drugs affecting mood, which may have potential negative impact on this study.

B. Materials

Arithmetic task and passage reading were adopted to record speech data. Subjects were required to finish the experiment according to the instructions displayed on a computer screen. Three arithmetic tasks and four passage readings were utilized in an experiment, with each task lasting two minutes. Passage readings were set before and in the middle of each arithmetic task. The arithmetic tasks were the sum operation of 0-9 with increased difficulty. The experiment process was shown in Fig. 1. Participants were required to read the passage clearly, avoiding noise interference caused by physical activities.

Speech data were segmented and labeled in advance. Four recordings for each subject were collected in one experiment. Hence there were 120(30∗4) recordings in total in our experiment.

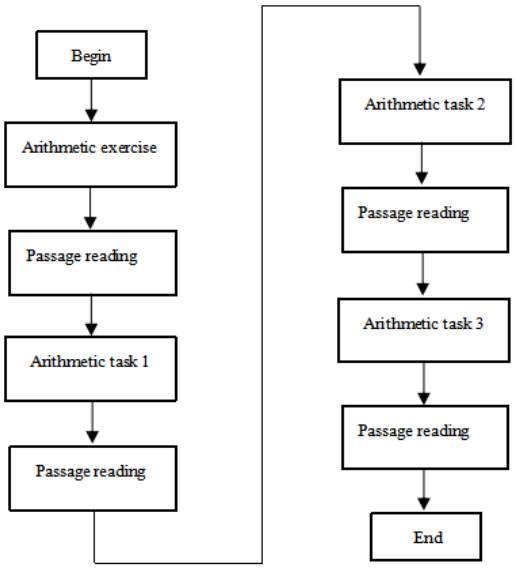


Fig. 1. Experimental flow chart.

C. Speech Recording and Preprocessing

Participants were seated in a quite electromagnetic shielding room. The speech recordings were segmented into 25ms one frame with 50% overlap using hamming window. The 80-4000Hz band-pass filter were adopted to remove power noise on each frame.

D. Feature Extraction

Several speech linear and nonlinear features, including pitch, jitter, energy, TEO, formant, MFCC were extracted for further analysis.

Previous researches have verified pitch to be an effective biomarker for psychological stress recognition[17]. We extracted pitch frequency using cepstrum. The calculation process was shown in Fig. 2.

Jitter reflects the disturbance when the vocal cords vibratethe. The general explanation of disturbance is a deviation from stability or law. Twenty pitch periods were utilized to calculate jitter[18]. The first-order perturbation function can calculate the pitch frequency perturbation. The formula was as follows:

$$\text{Jitter} = \frac{100}{(N-1)\bar{a}} \sum_{i=2}^{N} |a_i - a_{i-1}| \quad (1)$$

N represents the number of pitch period, and a_i indicates pitch.

Teager Energy Operator (TEO) proposed by H. M. Teager[19] can extract the instantaneous energy of signal. The operator was defined as:

$$\psi[x(t)] = \left(\frac{dx(t)}{dt}\right)^2 - x(t)\frac{dx^2(t)}{dt^2} \quad (2)$$

Formant features were calculated employing power spectrum peak detection method based on linear prediction coefficient (LPC)[20]. The power spectrum H(z) was defined as:

$$p(w) = |H(z)|^2_{z=e^{jw}} = \frac{1}{|1-\sum_{k=1}^{p} a_k e^{-jw}|^2} \quad (3)$$

After calculating the power spectrum curve by formula 3, formant center frequency and bandwidth were extracted by peak detection. Fig. 3 displayed an example of formant detection under one frame speech. Compared to Welch method, LPC method can eliminate the influence of pitch, and achieved good performance.

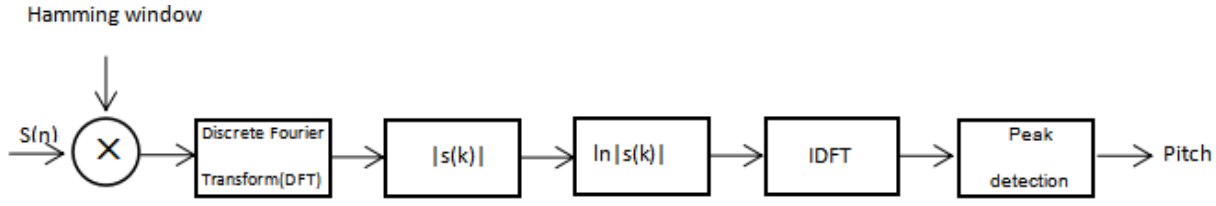
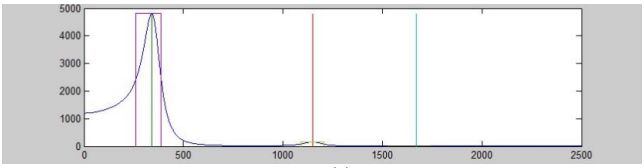
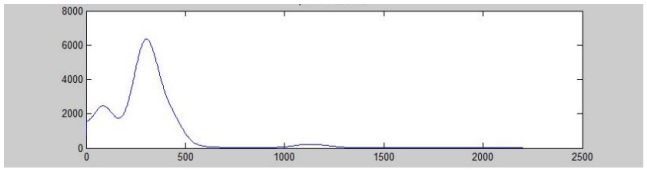


Fig. 2. Pitch extraction using cepstrum.



(a)



(b)

Fig. 3. Formant center frequency and bandwidth chart using LPC and Welch method. (a. LPC method, b. Welch method)

The sub-band energy ratios were features calculated using one frame[21]. Speech recordings were segmented into 51.2 ms one frame to analyze power spectra employing Welch method, and then fast Fourier transform was performed. Four features were obtained by the ratio of energy in four frequency sub-bands including 0-500Hz, 500-1000Hz, 1000-1500Hz and 1500-2000Hz to the energy at 0-2000Hz. Fig. 4 represented the calculation process.

Mel-Frequency Cepstral Coefficients (MFCC)[22] was also extracted following the process in Fig. 5.

E. Feature Selection

High dimensional feature sets may contain redundant information, which is not only computationally expensive, but may also reduce the classification accuracy. Linear Discriminant Analysis (LDA), also known as Fisher Discriminant Analysis was employed in this study to get better performance[23]. The main idea of this method is to project the multi-dimensional model into the best vector space so that the degree of dispersion between classes is as large as possible, and the degree of dispersion within classes is as small as possible to achieve the best classification performance. The calculation process were as follows:

The variable x_i is set, and the mean of a certain type of variable can be obtained according to the following formula:

$$u_i = \frac{1}{n_i}\sum_{x \in class(i)} X \tag{4}$$

n_i means the number of the variables of the type.

The population mean is defined by:

$$\mathrm{u} = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{5}$$

Formula 6 and Formula 7 are the between and within class scatter matrices.

$$S_b = \sum_{i=1}^{c} n_i (u_i - u)(u_i - u)^T \tag{6}$$

$$S_w = \sum_{i=1}^{c} \sum_{x_k \in class(i)} (u_i - u)(u_i - u)^T \tag{7}$$

The Fisher discriminant formula is defined as follow:

$$\mathrm{J}(\varphi) = \frac{\varphi^T S_b \varphi}{\varphi^T S_w \varphi} \tag{8}$$

And then maximize J（φ） to achieve the best performance.

F. Classification and Statistical Analysis

After feature extraction, the one sample t-test was employed to find features that are significantly different between groups. The speech recordings before the arithmetic task were took as control group, while the speech recordings between tasks were treated as stress group. And then the classification between the stress group and control group was performed. BP neural network was adopted for classification. Since the significant gender differences, the speech data were analyzed separately between males and females. In order to measure the precision of the results, 10-fold cross-validation was employed.

III. Experimental Results and Discussion

Sixteen linear and nonlinear features extracted for analysis were shown in Table 1. As mentioned previously, classification accuracy was adopted with 10-fold cross-validation. The precision values were 96.37% for males and 100.00% for females. We showed the average classification accuracy in Table 2.

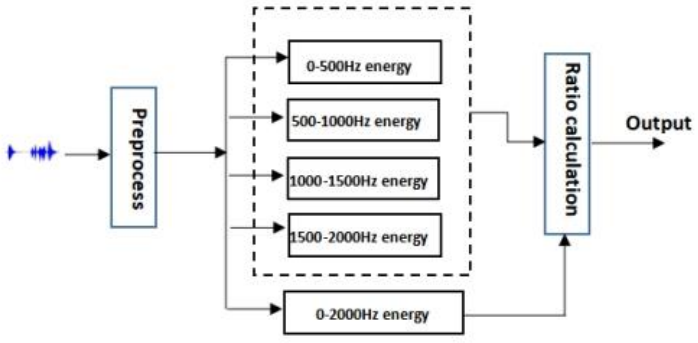


Fig. 4. The calculation process of sub-band energy ratio.

Fig. 5. The process of MFCC calculation.

Linear and nonlinear features mentioned above were used for analysis. The speech recordings before the arithmetic task were took as control group, while the speech recordings between tasks were treated as stress group. The one sample t-test for normality evaluation at 0.05 significance level was employed on each feature. The features with p value less than 0.05 were considered significant. The mean difference between groups were significant for pitch, jitter, short-term energy variance. Pitch, jitter and short-term energy variance were found a significant increase in stress group compared to control group both in male and female data. Vocal cords were strained under stress, resulting in the increase of pitch, jitter and short-term energy. It was also found that pitch, jitter and short-term energy increased first and then decreased as the difficulty of the task increased. The possible explanation was that as the difficulty of the task increased, the stress felt on the human body also enhanced, and then lead to the alteration of speech indicators. Furthermore, the features first increased and then decreased, which may be explained as the result of adaptability to stress.

TABLE I. LINEAR AND NONLINEAR FEATURES TABLE TYPE STYLES

Features	Features
Pitch	Short-term energy variance
Jitter	Mean short-term energy
Pitch fluctuation range	Short-term energy fluctuation range
Maximum pitch	First formant variance
Minimum pitch	Mean of the first formant
Median pitch	Maximum value of first formant
MFCC	Minimum value of first formant
TEO	Sub-band energy ratio

TABLE II. ACCURACY OF THE SPEECH FEATURES FOR MALES AND FEMALES

Sex	Male	Female
Accuracy	96.37%	100.00%

IV. CONCLUSION

In this study, we extracted different linear and nonlinear features including pitch, short-term energy, sub-band energy ratio, teager energy operator (TEO), formant, mel-frequency cepstral coefficients (MFCC) to analyze speech under psychological stress. It was found that pitch, jitter, energy did play an important role in stress recognition. Pitch, jitter and short-term energy variance showed a significant increase under stress both in male and female speech. Furthermore, average classification accuracy of 96.37% for males and 100.00% for females were obtained. In summary, this study may provide effective and reliable biomarkers for stress detection.

ACKNOWLEDGEMENT

This work was supported in part by the Fundamental Research Funds for the Central Universities (Grant No.31920160062), in part by the National Natural Science Foundation of China (Grant No. 71861030).

REFERENCES

[1] S. Cohen, D. Janicki-Deverts, and G. E. Miller, “Psychological Stress and Disease,” JAMA, vol. 298, no. 14, pp. 1685–1687, Oct. 2007, doi: 10.1001/jama.298.14.1685.

[2] S. F. Chan and A. M. La Greca, “Perceived Stress Scale (PSS),” Encycl. Behav. Med., pp. 1646–1648, 2020, doi: 10.1007/978-3-030-39903-0_773.

[3] I. Ulstein, T. B. Wyller, and K. Engedal, “High score on the Relative Stress Scale, a marker of possible psychiatric disorder in family carers of patients with dementia,” Int. J. Geriatr. Psychiatry A J. psychiatry late life allied Sci., vol. 22, no. 3, pp. 195–202, 2007.

[4] L. Lemyre and R. Tessier, “Measuring psychological stress. Concept, model, and measurement instrument in primary care research.,” Can. Fam. Physician, vol. 49, p. 1159, 2003.

[5] H. M. Burke, M. C. Davis, C. Otte, and D. C. Mohr, “Depression and cortisol responses to psychological stress: A meta-analysis,” Psychoneuroendocrinology, vol. 30, no. 9, pp. 846–856, 2005, doi: https://doi.org/10.1016/j.psyneuen.2005.02.010.

[6] S. C. Mueller et al., “Early-life stress is associated with impairment in cognitive control in adolescence: An fMRI study,” Neuropsychologia, vol. 48, no. 10, pp. 3037–3044, 2010, doi: https://doi.org/10.1016/j.neuropsychologia.2010.06.013.

[7] F. M. Al-shargie, T. B. Tang, N. Badruddin, and M. Kiguchi, “Mental Stress Quantification Using EEG Signals,” in International Conference for Innovation in Biomedical Engineering and Life Sciences, 2016, pp. 15–19.

[8] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes, “Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk,” Ieee Trans. Biomed. Eng., vol. 51, no. 9, pp. 1530–1540, 2004.

[9] C. L. Giddens, K. W. Barron, J. Byrd-Craven, K. F. Clark, and A. S. Winter, “Vocal indices of stress: A review,” J. Voice, vol. 27, no. 3, pp. 390.e21-390.e29, 2013, doi: 10.1016/j.jvoice.2012.12.010.

[10] T. Johnstone, C. M. Van Reekum, T. Bänziger, K. Hird, K. Kirsner, and K. R. Scherer, “The effects of difficulty and gain versus loss on vocal physiology and acoustics,” Psychophysiology, vol. 44, no. 5, pp. 827–837, 2007.

[11] E. Mendoza and G. Carballo, “Acoustic analysis of induced vocal stressby means of cognitive workload tasks,” J. Voice, vol. 12, no. 3, pp. 263–273, 1998.

[12] S. Sondhi, M. Khan, R. Vijay, and A. K. Salhan, “Vocal Indicators of Emotional Stress,” Int. J. Comput. Appl., vol. 122, no. 15, pp. 38–43, 2015, doi: 10.5120/21780-5056.

[13] K. R. Scherer, D. R. Ladd, and K. E. A. Silverman, “Vocal cues to speaker affect: Testing two models,” J. Acoust. Soc. Am., vol. 76, no. 5, pp. 1346–1356, 1984.

[14] T. W. Buchanan, J. S. Laures-Gore, and M. C. Duff, “Acute stress reduces speech fluency,” Biol. Psychol., vol. 97, no. 1, pp. 60–66, 2014, doi: 10.1016/j.biopsycho.2014.02.005.

[15] J. H. L. Hansen, W. Kim, M. Rahurkar, E. Ruzanski, and J. Meyerhoff, “Robust emotional stressed speech detection using weighted frequency subbands,” EURASIP J. Adv. Signal Process., vol. 2011, 2011, doi: 10.1155/2011/906789.

[16] S. Shukla, S. Dandapat, and S. R. M. Prasanna, “Spectral slope based analysis and classification of stressed speech,” Int. J. Speech Technol., vol. 14, no. 3, pp. 245–258, 2011, doi: 10.1007/s10772-011-9100-x.

[17] J. W. Heisse, “Audio stress analysis–a validation and reliability study of the psychological stress evaluator (PSE),” in Proceedings of the Carnahan Conference on Crime Countermeasures, 1976, pp. 5–18.

[18] I. R. Titze and H. Liang, “Comparison of Fo extraction methods for high-precision voice perturbation measurements,” J. Speech, Lang. Hear. Res., vol. 36, no. 6, pp. 1120–1133, 1993.

[19] H. M. Teager and S. M. Teager, “Evidence for nonlinear sound production mechanisms in the vocal tract,” in Speech production and speech modelling, Springer, 1990, pp. 241–261.

[20] S. McCandless, “An algorithm for automatic formant extraction using linear prediction spectra,” IEEE Trans. Acoust., vol. 22, no. 2, pp. 135–141, 1974.

[21] T. Yingthawornsuk, H. K. Keskinpala, D. M. Wilkes, R. G. Shiavi, and R. M. Salomon, “Direct acoustic feature using iterative EM algorithm and spectral energy for classifying suicidal speech,” 2007.

[22] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans. Acoust., vol. 28, no. 4, pp. 357–366, 1980.

[23] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. R. Mullers, “Fisher discriminant analysis with kernels,” in Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No.98TH8468), 1999, pp. 41–48, doi: 10.1109/NNSP.1999.788121.