PHYSICAL CHARACTERISTICS OF VOCAL FOLDS DURING SPEECH UNDER STRESS

Xiao Yao[1], Takatoshi Jitsuhiro[12], Chiyomi Miyajima[1], Norihide Kitaoka[1], Kazuya Takeda[1]

[1]Graduate School of Information Science, Nagoya University, Aichi, Japan
[2]Department of Media Informatics, Aichi University of Technology, Gamagori-shi, Aichi, Japan

ABSTRACT

We focus on variations in the glottal source of speech production, which is essential for understanding the generation of speech under psychological stress. In this paper, a two-mass vocal fold model is fitted to estimate the stiffness parameters of vocal folds during speech, and the stiffness parameters are then analyzed in order to classify recorded samples into neutral and stressed speech. Mechanisms of vocal folds under stress are derived from the experimental results. We propose using a Muscle Tension Ratio (MTR) to identify speech under stress. Our results show that MTR is more effective than a conventional method of stress measurement.

Index Terms— speech under stress, two-mass model, physical parameters

1. INTRODUCTION

The affect of stress on the speech signals has been the topic of numerous studies. Many factors, such as emotional state, fatigue, physical environment, and workload can cause people to experience stress. It has become increasingly important to study speech under stress in order to improve the performance of speech recognition systems, to recognize when people are in a stressed state, and to understand the context in which the speaker is communicating.

Many scholars have devoted their research to the field of analysis of stress in speech. Stress is a psycho-physiological state characterized by subjective strain or distress, dysfunctional physiological activity, and deterioration of performance [1]. Some external factors (workload, noise, vibration, sleep loss, etc.) and internal factors (emotional state, fatigue, etc.) may induce stress. These stress factors are believed to affect voice quality, and are likely to be detrimental to the performance of communication equipment and systems with vocal interfaces [2].

The performance of speech recognition algorithms are significantly challenged when the speech is produced by people experiencing stress. The influence of the Lombard effect on speech recognition has been examined in [3], and in some special environments, workload task stress has been proven to have an significant impact on the performance of speech recognition systems [4][5]. As a recognition method, Bou-Ghazale and Hansen proposed perturbation models of neutral-to-stressed speech with a hidden Markov model [6]. In 1980, Teager suggested that speech production results from vortex-flow interaction [7], and some nonlinear features based on the Teager energy operator (TEO) have been proposed to detect stress [8][9]. One possible application suggested for stressed speech detection is to catch people committing remittance fraud [10], which is considered to be a serious problem in Asian countries.

Our work mainly concentrates on the analysis of stressed speech. Our proposed features are based on a speech production model rather than on observed speech features. Parameters based on a physical model can represent characteristic of speaking style more clearly than conventional methods. We do this in order to gain a deeper understanding of the working mechanisms of the vocal folds, that are responsible for different speaking styles. We will explore the properties of the underlying physical speech production system, and search for essential factors related to stress. The characteristics of the glottal source of speech can be related to physical parameters, and an explanation of how the two-mass model applies to real speech can be made.

In this paper, we concentrate on the estimation of stiffness parameters of vocal folds from real speech by fitting the two-mass model [11]. Analysis is conducted to represent the corresponding variation in stiffness parameters under neutral and stressed conditions, and to explore a new physical feature, called the muscle tension ratio (MTR), which is used to classify neutral and stressed speech. Muscle tension characteristics can be further studied for better understanding of the behavior of vocal folds. The paper is structured as follows. In Section 2, a fitting method for the two-mass model is proposed to estimate the stiffness parameters. In Section 3, the obtained parameters are analyzed to show the tension characteristics of the vocal folds under different conditions. Finally, in Section 4 we draw our conclusion.

.

2. ESTIMATION OF PHYSICAL PARAMETERS

2.1. Physical model

The two-mass vocal fold model was proposed by Ishizaka and Flanagan to simulate the process of speech production [11].

In the two-mass model, each vocal fold is represented by two mass-spring-damper systems, joined with a coupling stiffness. This can be represented by the following equations:

$$m_1 \frac{d^2 x_1}{dt^2} + r_1 \frac{dx_1}{dt} + s_1(x_1) + k_c(x_1 - x_2) = F_1, \quad (1)$$

$$m_2 \frac{d^2 x_2}{dt^2} + r_2 \frac{dx_2}{dt} + s_2(x_2) + k_c(x_2 - x_1) = F_2, \quad (2)$$

where m_i are the masses, x_i are their horizontal displacements measured from the rest (neutral) position $x_0 > 0$, and k_c is the coupling stiffness. In this equation,

s_i are the equivalent tensions with non-linear relations given by

$$s_i(x_i) = k_i(x_i + \eta x_i^3) \quad \text{i=1,2,} \quad (3)$$

where k_i are stiffness coefficients and η is a coefficient of the nonlinear relations. In this paper, it is mainly the stiffness coefficients which are examined.

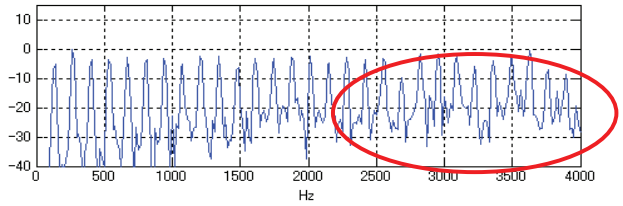
Generally, the stiffness of vocal folds mainly depends on two muscles: the cricothyroid muscle and the thyroarytenoid muscle. The weighted activities of cricothyroid and thyroarytenoid muscles can be denoted as CT and TA. TA is the main factor causing the variation in the fundamental frequency (F_0) in neutral phonation. Thus, in the normal speech phonation modus, an increase in F_0 results from higher TA activity while CT activity is relatively low.

In the two-mass model, the coupling stiffness k_c is relative to the tension in the thyroarytenoid muscle (TA), and stiffness coefficients are represented as k_1=CT+TA and k_2=CT-TA [12]. Therefore, high TA and low CT values during normal phonation lead to high values for k_1, while k_2 is relatively low.

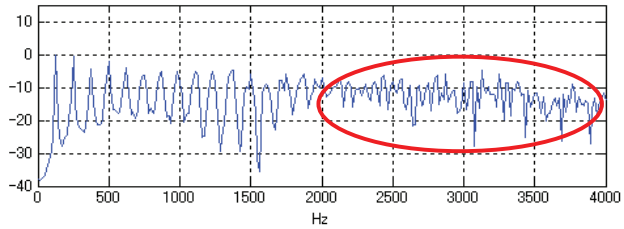
The two-mass model can be represented as a vocal fold model connected to a four-tube model. For the configuration of the vocal fold model, all of the parameters are fixed as constants, using the typical value for males, except the stiffness coefficients. The vocal tract is represented by a standard four-tube configuration for the vowel /a/ [13]. Therefore, the glottal flow will be simulated to analyze the interaction between the vocal folds and vocal tract when saying the vowel /a/.

2.2 Features indicating stress

When a speaker produces speech in a stressed condition, the muscle of the vocal folds can be impacted [9], and the excitation source of speech, which depend on the behavior of the vocal folds, are normally different from the neutral condition. To emphasize the excitation source, real data from a database are analyzed to illustrate the difference between neutral and stressed speech. The speech is preceded by the linear predicative coding (LPC), and the residual signals can be obtained in order to study their spectrum. The results are shown in Figure 1. When stress occurs, the harmonic structure of the spectrum loses clarity in the high frequency band, and the spectrum becomes smooth, which may result from interaction between vocal folds and the vocal tract [10].



(a) Spectrum of residual signal for neutral speech



(b) Spectrum of residual signal for stressed speech

Figure1. Spectrum of residual for neutral and stressed speech

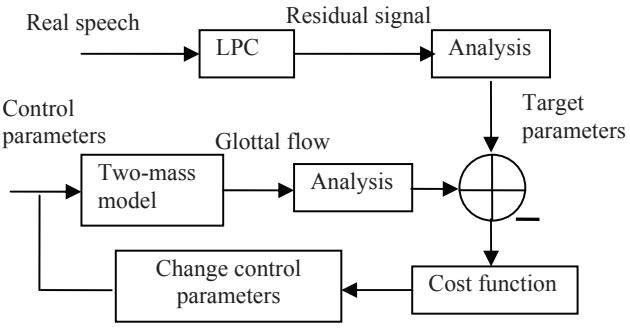


Figure 2 Structure of algorithm

The spectral flatness measure (SFM) can be applied to describe the smoothing level in the power spectrum, which can be a possible indication of stress. Spectral flatness is defined as a ratio of the geometric mean and the arithmetic mean of the power spectrum:

$$SFM = \frac{\sqrt[N]{\prod_{n=0}^{N-1} S(n)}}{\frac{1}{N}\sum_{n=0}^{N-1} S(n)}, \tag{4}$$

where $S(n)$ is the magnitude of the nth bin of the power spectrum.

Therefore, as target parameters in the data, the fundamental frequency (F_0) and the spectral flatness measure (SFM) can be chosen. With these target parameters, the two-mass model can be fitted to the glottal flow coming from real speech in the database to estimate the stiffness parameters.

2.3. Algorithm to fit real data

As control parameters for fitting the model to real speech, the linear stiffness values k_1 and k_2, and coupling stiffness value k_c are selected. Parameters k_1, k_2, and k_c represent the degree of tension occurring in the vocal folds, and they are the main factors controlling fundamental frequency F_0 during vocal fold oscillation.

Fitting the two-mass model to the real data involves two steps. First, the real speech coming from the database is analyzed using linear predictive coding (LPC) to reach the residual signal, which removes the influence of formant and lip radiation. Then, the measured target parameters denoting F_0 and SFM can be determined from the spectrum of the residual signal. In the second step, each set of target parameters is considered separately. Then, the simulation can be conducted using the two-mass model to generate the glottal flow using constant control parameters. F_0 and SFM are calculated from the simulated glottal flow, and are compared with the measured target parameters obtained in the first step to obtain the difference between them. The distinction between simulated target parameters and measured target value can be represented by a cost function. Then the control parameters are varied until the cost function reaches a minimum.

The cost function can be defined as a weighted sum of the squared difference between the simulated parameters and the measured targets from real speech.

$$C = \alpha_1 \left(F_0^* - F_0\right)^2 + \alpha_2 \left(SFM^* - SFM\right)^2, \tag{5}$$

where the asterisk denotes the target value. The weights are given the value $\alpha_1 = 1/\overline{F_0}, \alpha_2 = 1/\overline{SFM}$ to normalize the different target parameters to the same range, where the overbar denotes mean values over the target region. Optimal values of the control parameters were then calculated using a Nelder-Mead simplex method [14], which is implemented to search for the optimal stiffness parameters which will minimize the cost function. The structure of this algorithm is shown in Figure 2.

2.4. Muscle tension ratio (MTR)

We propose a Muscle Tension Ratio (MTR) to estimate speaking style. MTR can be represented as

$$MTR = k_1 / k_c, \quad (6)$$

Since k_c corresponds to the tension of the thyroarytenoid muscle (TA), and k_1 is represented as CT+TA, the MTR is actually the ratio of the tension of the cricothyroid and thyroarytenoid muscles (CT/TA).

When a speaker experiences stress, the contraction of the cricothyroid muscle causes higher tension, while the activity of the thyroarytenoid muscle becomes relatively low, which causes a lack of articulation in the speech produced.

3. EXPERIMENTS

3.1. Database and experimental setup

In the experiment, we used a database collected by the Fujitsu Corporation [10]. This database contains speech samples from eleven subjects, four for male, and seven female. To simulate mental pressure resulting in psychological stress, three different tasks, were introduced. These tasks were performed by the speaker while having a telephone conversation with an operator, in order to simulate a situation involving pressure during a telephone call.

The three tasks involved (A) Concentration; (B) Time pressure; and (C) Risk taking. For each speaker, there are four dialogues with different tasks. In two dialogues, the speaker is asked to finish the tasks within a limited amount of time, and in the other dialogues there is a relaxed chat without any task.

All of the data come from telephone communication, so the sampling frequency is 8 kHz. The segments with the vowel /a/ were cut from the speech, selected as the samples, and analyzed with 12-order LPC. The frame size chosen to perform the experiment was 64ms, with 16ms for frame shift. The frequency band of spectrum is limited to 3KHz-4KHz to calculate the spectral flatness measure.

3.2. Results and analysis

Figure 3 shows the simulation results using real data. The measured target parameters of the real data are plotted in the upper panel, and the lower panel shows simulated target values. As this figure illustrate, a good fit of the model exists, and simulated glottal flow is a good representation of the measured target parameters.

By fitting the model to the real data, the physical stiffness parameters can be estimated. The obtained parameters are then analyzed and the distribution is plotted to explore the

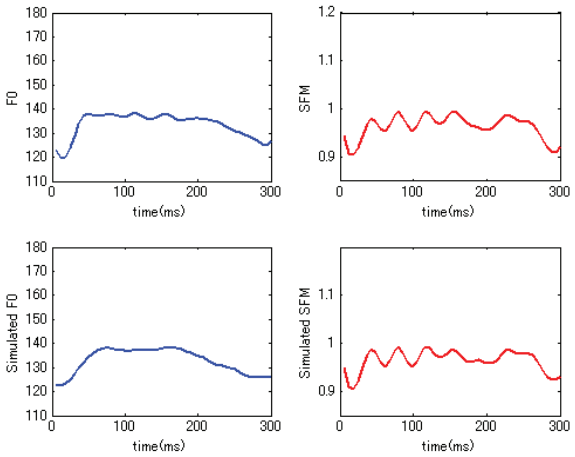


Figure 3. Target parameters in the real data (upper) and simulated speech (lower)

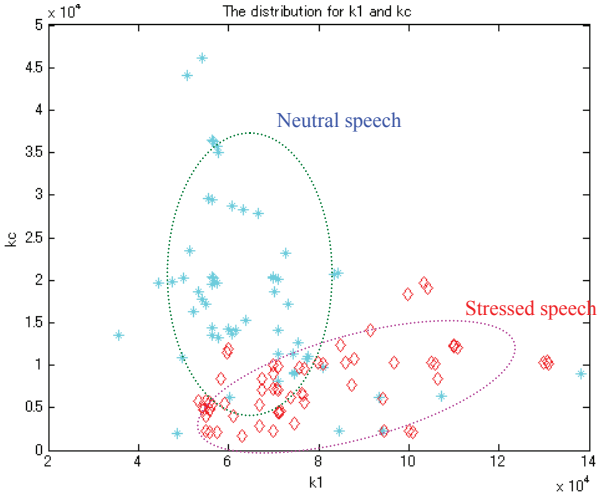


Figure 4. Distribution for k_1 and k_c

characteristics of neutral and stressed speech. In this experiment, the data for three male speakers and three female speakers in the database are used to compute the stiffness values for different conditions. Figure 4 shows the distribution results of k_1 and k_c for one of the male speakers. As this distribution shows, k_1 becomes smaller and k_c becomes larger under neutral conditions. On the contrary, when stress occurs, k_1 becomes higher while k_c is relatively low. Therefore, the MTR for k_1 and k_c is a reasonable means for classification of neutral and stressed speech.

Figure 5 shows the distribution results of SFM (upper panel) and MTR (lower panel) respectively. ROC curves can be used to evaluate and compare the classification performance of these two features. Using the results of the ROC curves in Figure 6, the areas under the curve (AUC) are computed. The AUC for SFM and MTR are 0.72315 and 0.91718 respectively. Therefore, MTR performs better than SFM to classify stressed speech from neutral speech.

4. CONCLUSION

In this paper, the physical stiffness parameters of vocal folds are estimated using a method which fits the two-mass model with the target for F_0 and SFM. The obtained stiffness parameters are

analyzed and a new physical feature, MTR, is proposed for classification of neutral and stressed speech. A physical explanation for the vocal representation of stress is proposed. High stress causes the vocal folds to maintain high tension in the cricothyroid muscle and low tension in the thyroarytenoid muscle. Further improvements in this model are possible if subglottal pressure, which may also have an influence on the production of stressed speech, is taken into account.

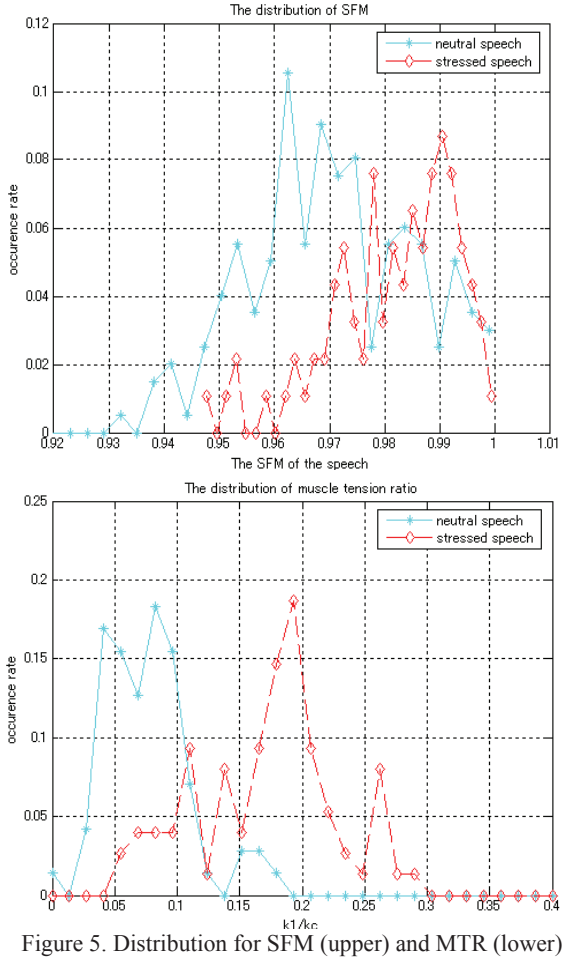


Figure 5. Distribution for SFM (upper) and MTR (lower)

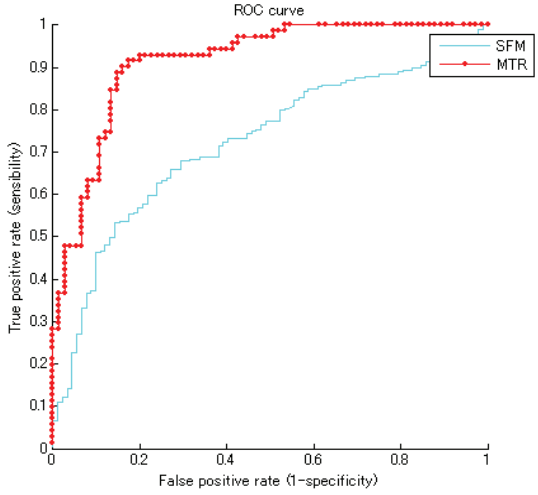


Figure 6. ROC curve for SFM and MTR

5. ACKNOWLEDGEMENTS

This work has been partially supported by the "Core Research for Evolutional Science and Technology" (CREST) project of the Japan Science and Technology Agency (JST). We are very grateful to Mr. Matsuo of the Fujitsu Corporation for the use of their database and for his valuable suggestions.

6. REFERENCES

[1] H.J.M. Steeneken, J.H.L. Hansen, "Speech Under Stress Conditions: Overview of the Effect on Speech Production and on System Performance", in Proc. ICASSP, vol. 4, pp. 2079-2082, 1999.

[2] D. Cairns, J.H.L. Hansen, "Nonlinear Analysis and Detection of Speech Under Stressed Conditions", The Journal of the Acoustical Society of America, vol. 96, no. 6, pp. 3392-3400, December 1994.

[3] J. C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," J. Acoust. Soc. Amer., vol. 1, pp. 510–524, 1993.

[4] E. G. Bard, C. Sotillo, A. H. Anderson, H. S. Thompson, and M. M. Taylor, "The DCIEM map task corpus: Spontaneous dialogue under sleep deprivation and drug treatment", Speech Commun., vol. 20, pp.71–84, Nov. 1996.

[5] C. Baber, B. Mellor, R. Graham, J. M. Noyes, and C. Tunley, "Workload and the use of automatic speech recognition: The effects of time and resource demands", Speech Commun., vol. 20, no. 12, pp. 37–54, Nov. 1996.

[6] S.E. Bou-Ghazale, J.H.L. Hansen, "Stress Perturbation of Neutral Speech for Synthesis based on Hidden Markov Models", IEEE Transactions on Speech & Audio Processing, vol. 6, pp. 201–216, May 1998.

[7] H. M. Teager and S. M. Teager, "A phenomenological model for vowel production in the vocal tract", Speech Science: Recent Advances, pp. 73–109, 1983.

[8] J. F. Kaiser, "On Teager's Energy Algorithm and Its Generalization to Continuous Signals", in Proc. 4th IEEE Digital Signal Processing Workshop. New Paltz, NY, Sept. 1990.

[9] G Zhou, J H L Hansen, J F Kaiser. "Nonlinear Feature based Classification of Speech under Stress", IEEE Trans. On Speech and Audio Processing, Vol. 3, pp: 201-206, Sept, 2001.

[10] N. Matsuo, N. Washio, S. Harada, A. Kamano, S. Hayakawa, K. Takeda. "A study of psychological stress detection based on the non-verbal information", IEICE Technical Report, IEICE-SP2011-35, pp 29-33, 2011. (In Japanese)

[11] K. Ishizaka, J.L. Flanagan. "Synthesis of voiced sounds from a two-mass model of the vocal cords", Bell.Syst.Tech. Journal, Vol. 51, pp. 1233-1268, 1972.

[12] C. Lucero, "Chest- and falsetto-like oscillations in a two-mass model of vocal folds", J.Acoust.Soc.Am. pp. 3355-3399, 1996.

[13] J. L. Flanagan, "Speech Analysis, Synthesis, and Perception", Springer-Verlag, New York, 1972.

[14] D. Kincaid, W. Cheney. "Numerical Analysis: Mathematics of Scientific Computing", 3rd ed. (Brook/Cole, Pacific Grove, CA), pp. 722-723, 2002.