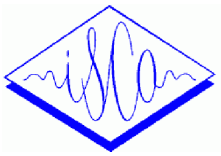

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Classification of Stressed Speech Using Physical Parameters Derived from Two-Mass Model

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

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

xiao.yao@g.sp.m.is.nagoya-u.ac.jp, jitsuhiro@aut.ac.jp,
miyajima@nagoya-u.jp, kitaoka@nagoya-u.jp, kazuya.takeda@nagoya-u.jp

10.21437/Interspeech.2012-377

Abstract

In this study, we investigate physical parameters which can be used to classify speech as either stressed or neutral based on a two-mass vocal fold model. The model attempts to characterize the behavior of the vocal folds and fluid airflow properties when stress is present. The two-mass model is fitted to real speech to estimate the values of physical parameters that represent the stiffness of vocal folds, vocal fold viscosity loss, and subglottal pressure coming from the lungs. The estimated parameters can be used to distinguish stressed speech from neutral speech because these parameters can represent the mechanisms of vocal folds under stress. We propose combinations of physical parameters as features for classification. Experimental results show that our proposed features achieved better classification performance than features derived from traditional methods.

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

1. Introduction

The affect of stress on 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 a speaker is communicating.

Researchers have attempted to probe reliable indicators of stress by analyzing acoustic variables. The first investigations of emotional speech were conducted by Van Bezooijen [1] and Scherer [2] using the statistical properties of acoustic features to recognize emotions from speech around the mid-1980s. Williams and Stevens found that the fundamental frequency (F0) has different characteristics for each emotion [3], and that respiration patterns and muscle tension also change [4]. The influence of the Lombard effect on speech recognition has been examined in [5], which analyzed selected acoustic features, such as amplitude and distribution of spectral energy, and found that spectral energy shifted to higher frequencies for consonants. High workload stress has been proven to have a significant impact on the performance of speech recognition systems, with speech under workload sounding faster, softer, or louder than neutral speech [6]. In 2011, Matsuo, *et al* worked on the frequency domain, and showed how difference in the spectrum of the high frequency band under stressful workload conditions aimed to catch people committing remittance fraud, and their proposed measure achieved better performance [7].

All the features mentioned are based on traditional linear speech production models. In 1980, Teager suggested that speech production is a nonlinear process and proposed a nonlinear model [8] [9]. As a result, some methods based on the Teager energy operator (TEO) [10] have been proposed to detect stress, like TEO-CB-Auto-Env, TEO-Auto-Env, and TEO-FM-Var [11]. But their performances degrade under text-independent conditions, and proposed methods don't consider the airflow patterns in nonlinear model.

It is suggested that the airflow is separated and concomitant vortices are distributed around the false vocal folds, which causes changes in airflow characteristics, thus the variability in interaction has been increased between vocal folds and vocal tract. Therefore, it is likely to be helpful to model airflow patterns in order to characterize speech production. Furthermore, vortex interactions differ markedly between neutral and stressed speech [12]. In physiological systems, it is believed that changes in physical characteristics induced by stressful conditions will affect the vortex-flow interaction patterns [13]. Therefore, a physical model which provides a direct means for representing the speech production is needed to estimate the parameters in the physiological system.

The physical features we proposed are based on a speech production model using fluid flow characteristics. The properties of the underlying physical speech production system are explored in an effort to search for the parameters of the physical model related to stress. Vortex interaction has a modulating effect on both glottal source and the vocal tract, but in this paper the characteristics of the glottal source of speech is chiefly considered, and related to physical parameters to show how it varies under text-independent conditions.

In our previous work [14], we estimated stiffness parameters for classification of stressed speech, which represent the muscle tension based on a physical model. In this paper, we concentrate on proposing a fitting method for the two-mass model to estimate physical parameters including stiffness, viscosity loss in vocal folds, and subglottal pressure from real speech. In section 2 the method for fitting to estimate the physical parameters is represented. In Section 3 our experimental results are analyzed to evaluate the obtained parameters and to show their corresponding classification performance for neutral and stressed speech. Finally, we draw our conclusions in Section 4.

2. Estimation of physical parameters

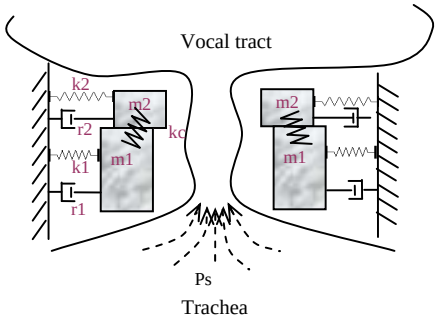

Figure 1: *Two-mass approximation of the vocal folds*

2.1. Two-mass model

The two-mass vocal fold model was proposed by Ishizaka and Flanagan to simulate the process of speech production [15]. Figure 1 shows the structure of the two-mass model. Each vocal fold is represented by two mass-spring-damper systems:

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

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

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

$$s_i(x_i) = k_i(x_i + \eta x_i^3) \quad (3)$$

where k_i are stiffness coefficients and η is a coefficient of the nonlinear relations.

The damping properties of the vocal folds, caused by the viscous resistance of the folds and the larynx tissues, can be represented as:

$$r_i = 2\zeta_i \sqrt{m_i k_i} \quad (4)$$

where ζ_i is a damping ratio.

If the subglottal pressure is represented as P_s, then the pressure is dropped to P_{11} when entering the glottis (at the edge of m_1) according to Bernoulli's equation.

$$P_s - P_{11} = \frac{\rho U_g^2}{2A_1^2} \quad (5)$$

where ρ is the air density, and U_g the volume velocity of glottal airflow, and A_{g1} the cross-sectional lower glottal area, which is represented by $A_{g1} = 2l_g(x_0 + x_1)$, where l_g is the length of the vocal folds, and x_0 is the displacement when the vocal fold is in the rest position. Because of the abrupt contraction in cross-sectional area at the inlet to the glottis, a vena contracta generates, which makes the pressure displays a greater drop. The drop is determined by the flow measurements from van den Berg:

$$P_s - P_{11} = (1.00 + 0.37)\frac{\rho U_g^2}{2A_{g1}^2}, \quad (6)$$

Along the masses m_1 and m_2, pressure drops as a result of air viscosity:

$$P_{i1} - P_{i2} = \frac{12\mu d_i l_g^2 U_g}{A_{gi}^3}, \qquad i = 1,2 \quad (7)$$

where μ is the air viscosity coefficient, and d_1 d_2 is the width of m_1 and m_2. P_{22} is the pressure at the glottal exit.

At the boundary between the two masses, the pressure drop can be calculated by:

$$P_{21} - P_{12} = \frac{\rho U_g^2}{2}\left(\frac{1}{A_{g1}^2} - \frac{1}{A_{g2}^2}\right), \quad (8)$$

Where P_{21} is the air pressure at the lower edge of m_2, and A_{g2} is the cross-sectional lower glottal area.

At the glottal outlet, abrupt expansion causes the pressure to recover because of the relatively large area of the vocal tract. This pressure is given by:

$$P_1 - P_{22} = \tfrac{1}{2}\rho\frac{U_g^2}{A_{g2}^2}[2N(1-N)], \quad (9)$$

where P_1 is the pressure in the inlet of vocal tract. Here the parameter N is defined as $N = A_{g2}/A_1$, with A_1 is the input area to vocal tract.

The Force depends on subglottal pressure, which can be represented by

$$F_i = l_g d_i P_{si} \text{ or } \quad F_i = l_g d_i P_{mi}$$

$$P_{mi} = (P_{i1} + P_{i2})/2 \quad (10)$$

where d_i is the thickness of vocal folds.

The two-mass model can be represented as a vocal fold model connected to a four-tube model. The vocal tract is represented by a standard four-tube configuration for the vowel /a/ [16]. Therefore, we assume that the shape of the vocal tract doesn't change over the utterance /a/, and the glottal flow is mainly considered to estimate the physical parameters.

2.2. Control parameters

The stiffness parameters, which represent the muscle tension of the vocal folds, are the main factors relating to fundamental frequency, which to some extent have an influence on the vortex-flow interaction in the vocal tract. In addition, viscous loss and subglottal pressure might also be variable parameters for fitting.

The viscosity of vocal fold tissue can determine the amount of energy loss due to internal friction in the folds. During phonation, the vocal folds are lubricated by mucus produced by the lining of the sinuses, and an osmotic gradient is established which induces fluid movement into and out of the vocal folds, thereby causing different hydration effects and presumably changing the viscosity of vocal fold tissue [17]. The damping ratio of viscosity has been estimated by Kaneko and Isshiki [18]. Results show that there is a close correlation between the damping ratio and variation in F_0, which is a stress indicator [19]. Therefore, in this work we assume the damping ratio is a parameter which is varies in a narrow range during phonation under different conditions. Considering the aerodynamics, the increase of viscosity induces a substantial decrease of airflow amount when glottis is closing, which has an impact on vortex-interaction. Because viscosity of the vocal folds depends mainly on the bulk of the cord (m_1 of our model), so here ζ_1 is primarily considered as a potential parameter for fitting.

Subglottal pressure is the pressure of airflow deriving from lungs, and becomes the main factor used by speakers to control phonation when producing speech. It is the other factor to affect F_0 and also has an influence on airflow separation. The separation phenomenon causes energy loss, which is

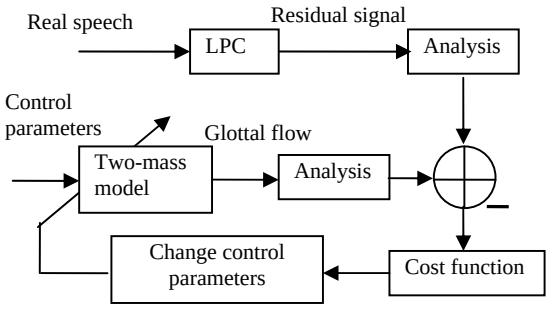

Figure 2: *Structure of fitting algorithm*

proportional to the increase in subglottal pressure. In other words, a higher subglottal pressure increases the velocity of airflow, causing more vortices to generate. That is the other primary reason to contribute the vortex-interaction between the vocal folds and the vocal tract. Based on these considerations, subglottal pressure can be selected as one of the physical parameters for fitting.

Therefore, in order to fit the model to real speech, stiffness k_1, k_2, k_c ,damping ratio ζ_1 , and subglottal pressure P_s are selected as control parameters for estimation.

2.3. Fitting algorithm

As target parameters in the data, the fundamental frequency (F_0) and spectral flatness measure (SFM) can be chosen. It is believed that when stress occurs, the fundamental frequency and spectrum of the glottal source are impacted. The harmonic structure of the spectrum loses clarity in the high frequency band, and the spectrum becomes smooth and irregular. This irregularity can be quantified with a "spectral flatness measure" (SFM):

$$SFM = \frac{\sqrt[N]{\prod_{n=0}^{N-1} S(n)}}{\frac{1}{N}\sum_{n=0}^{N-1} S(n)}, \tag{11}$$

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

Fitting the two-mass model to the real data involves two steps. First, real speech coming from the database is analyzed using linear predictive coding (LPC) to reach the residual signal, which removes the influence of formants 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, simulation can be conducted using the two-mass model to generate 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 the simulated target parameters and the measured target value can be represented by a cost function. The control parameters are then 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:

$$C = \alpha_1\left(F_0^* - F_0\right)^2 + \alpha_2\left(SFM^* - SFM\right)^2,$$
$$\alpha_1 = 1/\overline{F_0}, \alpha_2 = 1/\overline{SFM} \tag{12}$$

where asterisk denotes the target value. The weights are given the value α_1, α_2 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 the Nelder-Mead simplex method [20], 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.

After fitting, the physical parameters can be estimated using the two-mass model. Each parameter $\{P_s\}$, $\{[k_1, k_c]\}$, $\{\zeta_1\}$and parameter sets $\{[P_s, k_1, k_c]\}, \{[k_1, k_c, \zeta_1]\}, \{[P_s, k_1, k_c, \zeta_1]\}$ are analyzed to show and compare their classification performance.

3. Experiments

3.1. Database and experimental setup

In the experiments, we used a database collected by the Fujitsu Corporation containing speech samples from eleven subjects, four male, and seven female [9]. To simulate mental pressure resulting in psychological stress, three different tasks were introduced, which were performed by the speakers while having telephone conversations 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 calls, so the sampling frequency is 8 kHz. We chose 6 speakers, 3 male and 3 female). The segments with the vowel /a/ were cut from the speech, selected as samples. The number of samples depends on speakers, and the total amount is about 60-110 for each person. In order to increase the significance level of experimental results, a K-fold cross-validation method was used in experiments of classification, with 60% of samples for training, and the rest for testing. The samples are analyzed with 12th-order LPC, and frame size chosen to perform the experiment was 64ms, with 16ms for frame shift. The frequency band of the spectrum was limited to 3000Hz-4000Hz for calculating the spectral flatness measure. Linear classifiers were used based on minimum Euclidean distance to perform classification.

3.2. Results and analysis

By fitting the model to the real data, the physical parameters can be estimated. The obtained parameters are used as features to perform the classification into neutral and stressed speech. First, we focus on each parameter individually, and fix the other parameters at typical values. Then each parameter's impact on stress recognition is examined respectively. The results are shown in Figure 3. For these physical parameters, the results show that the stiffness achieves the best classification performance, which means it is strongly linked to stress. The other two parameters vary in performance depending on the speakers. For males, damping ratio plays a more important role, while for females, subglottal pressure, which determines the fundamental frequency, is a better indicator of stress.

F_0 is dependent on stiffness and subglottal pressure, while the viscosity of vocal folds is completely determined by stiffness and damping ratio, therefore the parameter sets

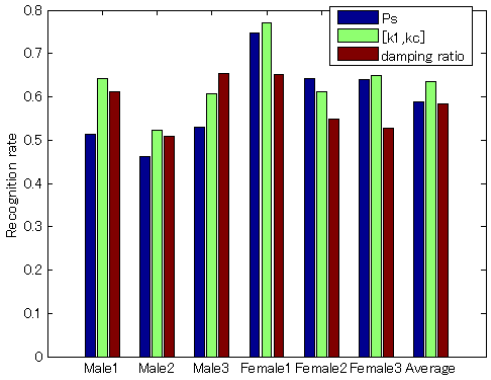

Figure 3: *Performance for each parameter*

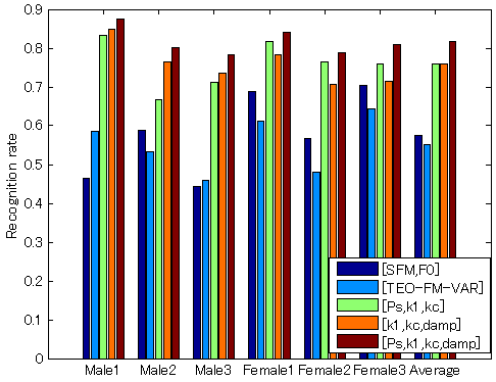

Figure 4: *Classification performance*

$\{[P_s, k_1, k_c]\}, \{[k_1, k_c, \zeta_1]\}, \{[P_s, k_1, k_c, \zeta_1]\}$ were proposed. We checked their performance and made a comparison with traditionally proposed features, and the results are shown in Figure 4. The results show that the proposed physical parameters perform better than the traditional features used for stress detection, which suggests that parameters estimated from a physical model are more effective at representing stress during phonation. Of the proposed sets, the stress classification rate of $\{[P_s, k_1, k_c]\}$ is higher than $\{[k_1, k_c, \zeta_1]\}$ with female data. This suggests that females are more likely to exhibit stress vocally through variation in F_0 than male speakers, which agrees with the results above. Furthermore, results show that $\{[P_s, k_1, k_c, \zeta_1]\}$ has the best stress recognition performance of the physical parameter sets. This illustrates that stiffness, damping ratio of the vocal folds, and subglottal pressure below the trachea are the factors to be impacted when a speaker is under stress.

4. Conclusions

In this paper, a physical model characterizing the fluid airflow properties was used to simulate speech production. The physical parameters, stiffness, damping ratio, and subglottal pressure, were estimated using a method that fits the two-mass model to real data using F_0 and SFM as targets. The obtained parameters were used as physical features for the classification of neutral and stressed speech under text-independent conditions. The conclusion drawn is that subglottal pressure from lungs, muscle tension, and viscosity of the vocal folds are key indicators of stress during phonation.

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] Van Bezooijen, R., "The Characteristics and Recognizability of Vocal Expression of Emotions", Foris, The Netherlands, 1984.

[2] Tolkmitt, F.J., Scherer, K.R., "Effect of experimentally induced stress on vocal parameters", J. Exp. Psychol. [Hum. Percept.] 12 (3): 302-313, 1986.

[3] Williams.C.E and Stevens, K,N., "Emotions and speech: Some acoustic Correlates", J. Acoust. Soc. Am. 52(4): 1238-1250, 1972.

[4] Bou-Ghazale S. E. and Hansen, J. H. L., "Generating stressed speech from neutral speech using a modified CELP vocoder", Speech Commun.,vol. 20:93–110, Nov. 1996.

[5] Bond Z. S. and Moore T. J., "A note on loud and lombard speech", in Int. Conf. Speech Language Processing '90: 969-972, 1990.

[6] Baber, C., Mellor, B., Graham R., Noyes J. M., and Tunley C., "Workload and the use of automatic speech recognition: The effects of time andresource demands", Speech Commun., 20(12): 37-54, 1996.

[7] Kamano, A., Washio, N., Harada, S., Matsuo, N., "A study of psychological suppression detection based on non-verbal information", IEICE Technical Report, IEICE-SP2010-64:107-110, 2010 (in Japanese)

[8] Teager, H. M., "Some observations on oral air flow during phonation", IEEE Trans. Acoustics, Speech, Signal Processing, 28(5): 599-601, 1980.

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

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

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

[12] Cairns, D., Hansen, J.H.L., "Nonlinear analysis and detection of speech under stressed conditions", J. Acoust. Soc. Am. 96(6): 3392-3400, 1994.

[13] Krane M., Barry M., and Wei T., "Unsteady behavior of flow in a scaled-up vocal folds model", J. Acoust. Soc. Am. 122: 3659-3670 2007.

[14] Yao, X., Jitsuhiro, T., Miyajima, C., Kitaoka, N., Takeda, K., "Physical characteristics of vocal folds during speech under stress", Proc. IEEE ICASSP'12, Kyoto, 4609-4612, 2012.

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

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

[17] Finkelhor, B.K, Titze, I.R., Durham, P.L., "The effect of viscosity change in the vocal folds on the range of oscillation", J Voive, 1: 320-325, 1988

[18] Isshiki, N., "Functional Surgery of the Larynx" ~Kyoto University, Kyoto, Japan, 62–67, 1977.

[19] Fung, Y. C.,. "Biomechanics. Mechanical Properties of Living Tissues", 2nd ed. Springer, New York, 23–65; 242–320, 1993.

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