Modeling of Physical Characteristics of Speech under Stress

Xiao Yao, *Member, IEEE*, Takatoshi Jitsuhiro, *Member, IEEE*, Chiyomi Miyajima, *Member, IEEE*, Norihide Kitaoka, *Member, IEEE*, and Kazuya Takeda, *Member, IEEE*

***Abstract*—This letter presents a method to perform the classification of speech under stress based on physical characteristics. A physical model is proposed to model airflow patterns in the physiological system in order to represent the process of speech production under psychological stress, and physical parameters characterizing airflow variations in the vocal folds, the vocal tract, and laryngeal ventricle are explored. Experimental evaluations show that the physical parameters are effective for the classification of stressed speech.**

***Index Terms*—Airflow pattern, physical characteristics, stress classification, two-mass model.**

I. Introduction

STRESS is a psycho-physiological state caused by environmental factors. It is characterized by subjective strain, dysfunctional physiological activity, and deterioration of performance [1]. Stress results in variations in speaker's pronunciation, making highly reliable speech recognition systems difficult to achieve. Therefore, the classification of speech under stress has become a popular subject of research.

The majority of studies for stress analysis have focused on acoustic features, such as pitch, spectral energy and speaking rate, derived from a linear speech production model [2], [3], [4]. In 1980, Teager proposed a nonlinear theory for speech production [5], [6]. It is believed that the presence of stress results in variability in airflow characteristics due to changes in physiological systems, thereby having modulating effect on speech production [7]. Cairns showed that the variation in airflow patterns differs markedly between neutral and stressed speech, and proposed the Teager energy operator (TEO) for stress classification [8]. However, the features in the previous studies lack a physical basis, and the methods do not consider the process of speech production, which is believed to be essential for classification of speech under stress. Therefore, it is necessary to develop a physical model in order to understand the variation in airflow characteristics caused by stress.

In our study, we mainly concentrate on the classification of stressed speech based on a physical speech production model. The production of stressed speech is characterized by modeling airflow patterns in the physiological system. We propose a classification method to estimate the essential parameters related to stress representing physical characteristics. Compared with acoustic parameters derived from traditional linear speech production theory, physical parameters are more robust and essential at representing the presence of stress. A developed two-mass model is proposed, and an explanation of how the physical model applied to real speech can be made.

II. Physiological System and Model

An assumption of the traditional linear speech production model is that the source and filter function independently of each other. Airflow from the lungs always propagates as a linear plane wave in the glottis and the vocal tract, and the pulsatile flow is the only source of speech production. However, there is increasing evidence suggesting that this assumption may not hold [5], [6]. This is because the airflow coming from glottis is very unstable as it passes the wall of vocal tract. Airflow separation occurs along the walls of the laryngeal ventricle around the false vocal folds, which can cause variability in airflow patterns [9], [10].

The presence of stress will cause speakers change their physiological system to react and adapt himself to the stressed condition, such as the muscle tension of the vocal folds or the shape of the vocal tract. Changes in physiological characteristics can result in variations in aerodynamics in the glottis, the vocal tract, the false vocal folds and the laryngeal ventricle, and then the stressed speech is produced. Therefore the aerodynamics and the physiological characteristics in the vocal system are essential for understanding the process of stressed speech production.

An alternate method is to model vocal airflow in order to characterize speech production. Aerodynamics in the glottis and the vocal tract could be studied by modeling the airflow patterns in physiological systems mathematically, which may allow us to explain the process of speech production more accurately and clearly. Therefore, a physical model is necessary to model the airflow patterns in the physiological system, in order to represent the process of speech production.

Two-mass model simulates the physical process of speech production by characterizing the vocal folds and the vocal tract, proposed by Ishizaka and Flanagan [11]. However, the laryngeal

Manuscript received November 20, 2014; revised April 06, 2015; accepted May 07, 2015. Date of publication May 18, 2015; date of current version June 01, 2015. This work was supported by the Fundamental Research Funds for the Central Universities under Grant 2014B16214. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Frederic Bechet.

X. Yao is with the College of IoT Engineering, Changzhou Key Laboratory of Robotics and Intelligent Technology, Hohai University, Changzhou 213000, China (e-mail: yaox@hhu.edu.cn).

T. Jitsuhiro is with the Department of Media Informatics, Aichi University of Technology, Aichi, Japan (e-mail: jitsuhiro@aut.ac.jp).

C. Miyajima, N. Kitaoka, and K. Takeda are with the Department of Media Science, Graduate School of Information Science, Nagoya University, Nagoya 464-8603, Japan (e-mail: miyajima@nagoya-u.jp; kitaoka@nagoya-u.jp; kazuya.takeda@nagoya-u.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2015.2434732

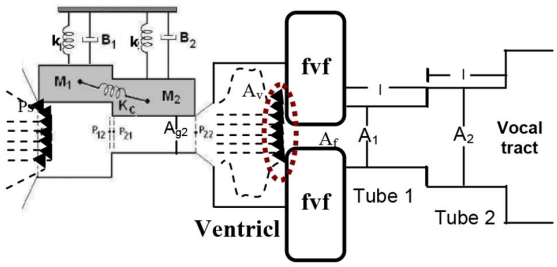

Fig. 1. Sketch of modified two-mass model. The vocal folds are represented by a mass spring-damping system, coupled with a four-tube model. The traditional model is modified by modeling the laryngeal ventricle and false vocal folds.

ventricle and the false vocal folds (fvf) are not considered in the traditional two-mass model. Fig. 1 shows a sketch of our proposed model. The aerodynamics for the vocal folds, the vocal tract, the laryngeal ventricle and false vocal folds are modeled.

III. Physical Characteristics for Stressed Speech

The target of this study is to classify stressed speech from the neutral speech. Comparing with neutral speech, stressed speech makes a significant difference in the variation in airflow patterns, and further in the acoustic interaction between the vocal folds and the vocal tract. Physical characteristics are examined to represent the production of stressed speech.

A. The Vocal Folds

The presence of stress can result in the vibration behavior of the vocal folds, so an increase in the variability of airflow characteristics can be caused due to differences in muscle tension of the vocal folds [7], [12]. The amplitudes of the glottal area and glottal volume velocity decrease gradually with increasing stiffness [13] because variation in the stiffness of the vocal folds influences the time span of the glottal opening and closing phases. During this time span, subglottal airflow is accelerated in the glottis, thus impacting the velocity of glottal airflow as well as the glottal source.

Generally, the stiffness of the vocal folds is considered to depend mainly on two muscles: the cricothyroid muscle (CT) and the thyroarytenoid muscle (TA) [14]. In the two-mass model, coupling stiffness k_{c} is relative to the tension in the TA muscle, so a high k_1 value and a low value for k_{c} represent the contraction of the CT muscle and relaxation of the TA muscle. Therefore, it is our assumption that stiffness parameters, k_1 and k_{c}, can be a potential factor in stress detection.

B. The Laryngeal Ventricle

Laryngeal ventricle is a fusiform fossa, situated between the ventricular and vocal folds on either side, and extending nearly their entire length.

The aerodynamics of the glottis is modeled using equations in the traditional two-mass model. Next, we model airflow patterns around the laryngeal ventricle and false vocal folds using the two-mass model. At the glottal outlet, expansion causes air pressure to recover because of the relatively larger area of the laryngeal ventricle. This pressure rise is represented by:

$$P_{22} - P_v = -\frac{\rho}{2} \cdot \frac{2}{A_{g2} A_E} \left(1 - \frac{A_{g2}}{A_E}\right) {U_{\mathrm{g}}}^2, \qquad (1)$$

where P_{22} is air pressure at the glottal exit. A_E is the area at the entrance to the laryngeal ventricle, and P_v is the pressure at this inlet. In order to simplify our model, we disregard the pressure changes when air enters the laryngeal ventricle. Therefore, we assume airflow is uniform without any expansion $A_{g2} = A_E$.

When air passes the laryngeal ventricle between the true vocal folds and false vocal folds, it is very unstable because of the negative pressure difference. Airflow separation occurs along the wall of laryngeal ventricle. The separation will change the effective area of the laryngeal ventricle into the false vocal folds, causing variability in airflow characteristics, thereby having modulating effect on speech production. Therefore, it is hypothesized that the effective area of the ventricle changes in relation to airflow separation in this area. Here, we use A_V to represent the effective area of the ventricle into the false vocal folds. The pressure drop at the inlet of the false vocal folds is calculated according to Bernoulli's equation:

$$P_v - P_{f1} = \frac{\rho}{2} \left(\frac{1}{A_f^2} - \frac{1}{A_v^2} \right) U_g^2, \qquad (2)$$

where A_f is the area of the false vocal folds. Since the false vocal folds do not vibrate during the process of phonation, A_f can be fixed to a constant.

Along the false vocal folds, pressure drops from P_{f1} to P_{f2} due to the loss from air viscosity:

$$P_{f1} - P_{f2} = 12 \frac{\mu l_f^2 d_f}{A_f^3} U_{\mathrm{g}}, \qquad (3)$$

where l_f and d_f are the length and thickness of the false vocal folds, respectively.

Since the area of the vocal tract is relatively large compared with the glottal area, an abrupt expansion cause the pressure to recover toward the atmospheric value at the inlet to the vocal tract.

$$P_{f2} - P_1 = -\frac{\rho}{2} \cdot \frac{2}{A_f A_1} \left(1 - \frac{A_f}{A_1}\right) {U_{\mathrm{g}}}^2, \qquad (4)$$

where P_1 is the pressure in the inlet of vocal tract.

C. The Vocal Tract

The vocal tract is defined as the structure bound by soft and hard tissues, which can be shaped by tongue, mouth, teeth, oral cavity, palate, nasal cavity and other articulators. The two-mass model is connected to a four tube model representing the vocal tract. The tube model is constructed using a transmission line analogy involving n cylindrical, hard-walled sections. The elemental values of the model are determined by cross-sectional areas $A_1 \cdots A_n$

In theory, (4) shows that both the velocity of glottal airflow, and the difference between the area of the outlet of the vocal folds and the inlet of the vocal tract, have an impact on the pressure difference inside and outside of the glottis. So the two factors can cause variations in the airflow patterns in the glottis, and thus are likely to be effective to represent the presence of stress.

A_1 in the four-tube model is the area of the entrance to the vocal tract in the supraglottis. Narrowing A_1 facilitates phonation by decreasing the oscillation threshold pressure of the vocal folds [15]. Since glottal area does not change significantly during the oscillation of the vocal folds, A_1 is the main

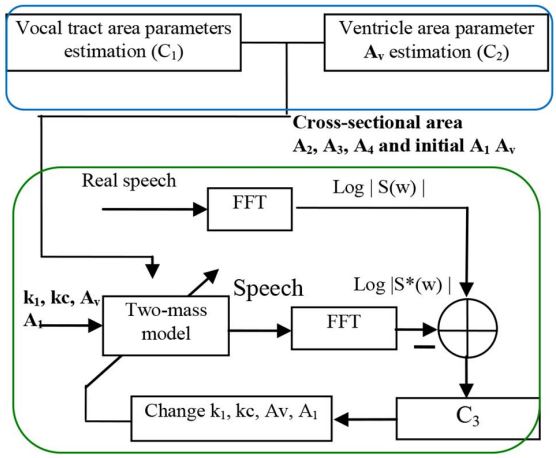

Fig. 2. Method for estimation of physical parameters. The first step is conducted for initialization, and the main parameters are estimated in the second step.

factor determining the pressure difference between the inside and outside of the glottis and has an impact on the acoustic interaction between the glottal source and the vocal tract. So A_1 should be selected as a parameter for representing stress.

IV. Algorithm

Fitting the model to real speech poses a difficulty because the existence of interaction makes it impossible to fit the vocal folds (VF) and vocal tract (VT) separately. Based on the pressure distribution discussed above, it is believed that stiffness parameters k_1, k_c and cross-sectional areas A_V, A_1, determining volume velocity U_g, are related to the acoustic interaction between VF and VT. Therefore, parameters k_1, k_c, A_1 and A_V, should be estimated together and selected as feature parameters for stress classification.

A_2, A_3 and A_4 have less impact on the interaction, as we showed in [16], [17], so they can be estimated separately. A_V dramatically affects irregularity in the harmonic structure of the spectrum in the high frequency band, so these parameters are estimated firstly.

The detailed fitting method for estimation of the physical parameters is shown in Fig. 2. This method includes two steps. First, cross-sectional areas of the four-tube model: A_1, A_2, A_3, and A_4 are estimated by the method Analysis-by-Synthesis (A-b-S). Cost function 1 (C_1) is defined as the root mean square (RMS) distance between the spectral envelope of simulated and original speech.

$$C_1 = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|\log P(\omega_i) - \log P^*(\omega_i)\right|^2}$$

$$P(\omega) = \frac{1}{|A(\omega)|^2} = \frac{1}{\left|\sum_{k=0}^{P} a_k e^{-j\omega k}\right|^2} \quad (5)$$

A_V is also estimated by A-b-S with cost function 2 (C_2)

$$C_2 = \frac{2}{N}\sum_{i=N/2^+1}^{N}\left|\log S(\omega_i) - \log S^*(\omega_i)\right|^2 \quad (6)$$

In the cost function, the power spectrum in the high frequency is used. The detail of the A-b-S fitting method is similar to that in the second step.

In the second step, A_2, A_3, and A_4 are fixed at obtained values, and A_1 and A_V is considered as the initial value for this fitting process. In the fitting, k_1, k_c, A_V and A_1 are selected as control parameters, and cost function 3 (C_3) is defined as:

$$C_3 = \frac{1}{N}\sum_{i=1}^{N}\left|\log S(\omega_i) - \log S^*(\omega_i)\right|^2 \quad (7)$$

where $S(\omega)$ and $S^*(\omega)$ are the power spectrums of the signals for simulated and real speech, respectively. After Fourier transform, optimal values of the physical parameters are estimated using a Nelder-Mead simplex method [18], which is implemented to search for the optimal stiffness parameters which minimize the cost function.

V. Evaluation

A. Database and Experimental Setup

In our experiments, we used a database collected by the Fujitsu Corporation containing speech samples from seven subjects (three male, and four female) [19]. 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 relaxed chat without any task.

The segments with the Japanese vowels /a/, /i/, /u/, /e/, /o/ were cut from the speech and selected as samples. The experiments were conducted for each speaker, and all of the results were speaker dependent. Here, we used samples from eleven subjects (four male, seven female) to show the classification performance for each speaker, respectively, in this speaker-dependent system. The number of samples depended on the speakers, and the total amount is about 700 for each person. In order to increase the significance level of the experimental results, a K-fold cross-validation method was used in the classification experiments, with 60% of samples used for training, and the rest used for testing. K was set to 4. Linear classifiers based on minimum Euclidean distance, which fit a multivariate normal density to each group, with a pooled estimate of covariance, were used to determine classification performance.

B. Evaluation for the Physical Parameters

It would be helpful to evaluate the accuracy of the fitting method to show that the proposed method works well. However, it is difficult to compare the simulated values with the actual values because sensors are not available to measure the actual values of the vocal system for human beings.

In order to describe the accuracy of the fitting method, comparison was made with different traditional synthesis methods, such as formant synthesis and LPC parameter synthesis and original two-mass model, by calculating the spectral distortion for real speech and simulated speech. Log-spectral distance

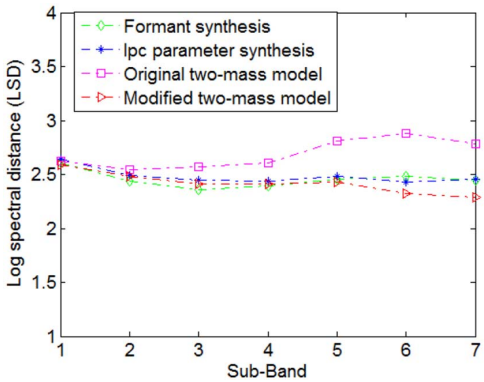

Fig. 3. LSD to evaluate accuracy of the fitting method, comparing with the traditional methods.

(LSD) was used to describe the difference in spectral distortion between real and simulated speech.

$$\mathrm{LSD} = \sqrt{\frac{1}{f(b)} \sum_{\omega_i \in B(b)} \left(10\log_{10}\left|S^*\left(\omega_i\right)\right| - 10\log_{10}\left|S\left(\omega_i\right)\right|\right)^2} \tag{8}$$

where $f(b)$ denotes the bandwidth of sub-band b and $\mathrm{B(b)}$ consists of a set containing all the discrete frequencies in sub-band b. $S(\omega)$ and $S^*(\omega)$ are the power spectrums of simulated and real speech, respectively. Here, $f(b)$ is 1000 Hz, and $B(b)$ consists the discrete frequencies in $[(b-1)*500, b*500+500]$, $b = 1, 2 \ldots 7$.

The results for the average values of log-spectral distance are illustrated in Fig. 3, which show that there is no difference in the low frequency bands comparing with the traditional methods. However, when the high frequency bands are taken into account, the results achieve an improvement in the accuracy of spectrum simulation when using the modified two-mass model. This indicates that the proposed method provides reliable accuracy for the fitting to real speech.

Evaluation Under Vowel Dependent Condition: Our previous works have showed the proposed physical features achieved better performance than acoustic features derived from traditional methods [16], [20]. In this section, we mainly compared the performance of physical parameter sets, $[k_1, k_\mathrm{c}]$, $[k_1, k_\mathrm{c}, A_1]$, $[k_1, k_\mathrm{c}, A_V]$ $[k_1, k_\mathrm{c}, A_1, A_V]$ from modified model and $[k', k_\mathrm{c}']$ estimated from the classical two-mass model, to evaluate the effectiveness of proposed parameters. Samples of the individual vowels /a/, /i/, /u/, /e/, /o/ were selected respectively for vowel-dependent experiments, and the average classification rate was then calculated. Fig. 4 compares the classification rates of parameter sets $[k_1', k_\mathrm{c}']$, $[k_1, k_\mathrm{c}]$, $[k_1, k_\mathrm{c}, A_1]$, $[k_1, k_\mathrm{c}, A_V]$ and $[k_1, k_\mathrm{c}, A_1, A_V]$. Comparing these results, we can see that parameter sets $[k_1, k_\mathrm{c}]$ estimated from modified model is more effective, and $[k_1, k_\mathrm{c}, A_1]$, $[k_1, k_\mathrm{c}, A_V]$ achieve better performance under the vowel dependent condition, in which individual vowels are considered separately. The performance of $[k_1, k_\mathrm{c}]$ is improved by 5% when A_V is considered because A_V represents the airflow variations in the laryngeal ventricle. A_1 is also effective for stress detection because the shape of the vocal tract does not change significantly when considering individual vowel, so A_1 only represents acoustic interaction, thus improving performance.

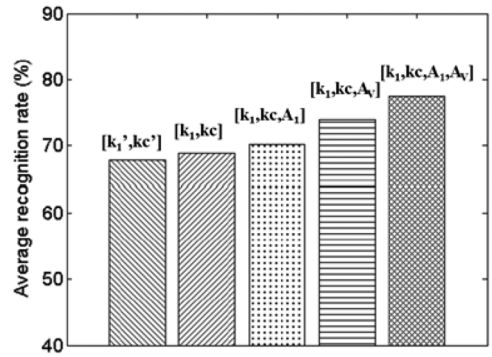

Fig. 4. Evaluation under vowel dependent condition, k_1, k_c, A_1 and A_V are effective for stress detection.

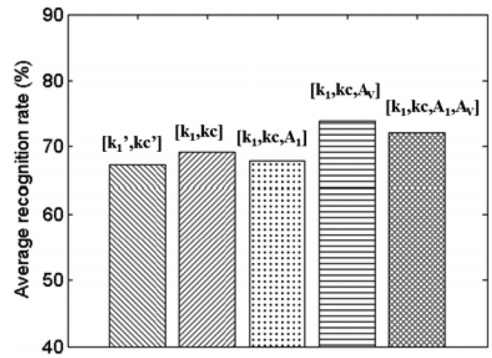

Fig. 5. Evaluation under vowel-independent condition. k_1, k_c and A_V are effective, but A_1 can't show its advantage for stress classification.

Evaluation Under Vowel-Independent Condition: In this evaluation, speech segments with the Japanese vowels.

/a/, /i/, /u/, /e/, /o/ were cut from the speech and selected as samples. All of the vowels were mixed for the vowel-independent condition. Experiments were conducted for each speaker, and all of the results were speaker dependent.

Results show that the classification performance decreases when A_1 is considered. A_1 determines the shape of the vocal tract, so it is not effective under the vowel-independent condition. When A_V is taken into account, classification performance is improved. The results are shown in Fig. 5. Since the samples selected in the experiment are mixed data from all the vowels, the results show that A_V can maintain its performance under vowel-independent conditions, because the area of the ventricle has less impact on the vocal tract, and thus does not rely on vowel information. From these results, it is believed that A_V is an essential parameter strongly related to stress. Larger A_V value indicates that the amount of airflow separation is increasing, causing the effective area at the inlet of the false vocal folds to broaden. Variations in the airflow patterns around the false vocal folds are caused, resulting in a stronger modulation effect on the produced speech.

VI. Conclusion

In this letter, the classification of speech under stress was performed based on a physical model. A physical model representing speech production can be used to characterize airflow patterns, and methods were proposed to estimate parameters to provide a better understanding of the physical characteristics for stress production. Results presented provide valuable insights into the classification of stressed speech.

REFERENCES

[1] H. J. M. Steeneken and J. H. L. Hansen, "Speech under stress conditions: Overview of the effect on speech production and on system performance," in *Proc. ICASSP*, Atlanta, Georgia, May 1996, pp. 7–10.

[2] R. Van Bezooijen, *The Characteristics and Recognizability of Vocal Expression of Emotions*. Foris, The Netherlands: de Gruyter, 1984.

[3] C. E. Williams and K. N. Stevens, "Emotions and speech: Some acoustic correlates," *J. Acoust. Soc. Amer.*, vol. 52, no. 4, pp. 1238–1250, 1972.

[4] Z. S. Bond and T. J. Moore, "A note on loud and lombard speech," in *Int. Conf. Speech Language Processing*, 1990, vol. 90, pp. 969–972.

[5] H. M. Teager, "Some observations on oral air flow during phonation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 5, pp. 599–601, 1980.

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

[7] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 201–206, 2001.

[8] D. Cairns and J. H. L. Hansen, "Nonlinear analysis and detection of speech under stressed conditions," *J. Acoust. Soc. Amer.*, vol. 96, no. 6, pp. 3392–3400, 1994.

[9] J. F. Kaiser, "Some observations on vocal tract operation from a fluid flow point of view," *Vocal Fold Physiol.: Biomech., Acoust., Phonatory Contr.*, pp. 358–386, 1983.

[10] S. M. Teager, "Evidence for nonlinear production mechanisms in the vocal tract," *Speech Prod. Speech Model.*, vol. 55, pp. 241–261, 1989.

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

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

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

[14] T. Haji, K. Mori, and K. Omori, "Experimental studies on the viscoelasticity of the vocal fold," *Acta oto-laryngologica*, vol. 112, no. 1, pp. 151–159, 1992.

[15] I. R. Titze and B. H. Story, "Acoustic interactions of the voice source with the lower vocal tract," *J. Acoust. Soc. Amer.*, vol. 101, p. 2234, 1997.

[16] X. Yao, T. Jitsuhiro, C. Miyajima, N. Kitaoka, and K. Takeda, "Classification of speech under stress based on modeling of the vocal folds and vocal tract," *EURASIP J. Audio, Speech, Music Process.*, Jul. 2013.

[17] X. Yao, T. Jitsuhiro, C. Miyajima, N. Kitaoka, and K. Takeda, "Estimation of vocal tract parameters for the classification of speech under stress," in *Proc. IEEE ICASSP'13*, Vancouver, BC, Canada, 2013.

[18] D. Kincaid and W. Cheney, *Numerical Analysis: Mathematics of Scientific Computing*, 3rd ed. Pacific Grove, CA, USA: Brook/Cole, 2002, pp. 722–723.

[19] N. Matsuo, N. Washio, S. Harada, A. Kamano, S. Hayakawa, and K. Takeda, A study of psychological stress detection based on the non-verbal information IEICE, Tech. Rep. IEICE-SP2011-35, 2011, pp. 29–33, (in Japanese).

[20] X. Yao, T. Jitsuhiro, C. Miyajima, N. Kitaoka, and K. Takeda, "Classification of speech under stress by physical modeling," *Acoust. Sci. Technol.*, vol. 34, no. 5, 2013.

This paper previously published in IEEE Signal Processing Letters