Classification of speech under stress by modeling the aerodynamics of the laryngeal ventricle

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-shi, Aichi, Japan
[2]Department of Media Informatics, Aichi University of Technology, Gamagori-shi, Aichi, 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

Abstract

We focus on variations in the aerodynamics of airflow patterns in the laryngeal ventricle and the false vocal folds based on a physical model for the classification of neutral and stressed speech. We modify the two-mass model to include the laryngeal ventricle, and the physical parameters characterizing airflow variations in the laryngeal ventricle under psychological stress are explored. The two-mass model is fitted to real speech by estimating the physical parameters representing stiffness of the vocal folds and effective area of laryngeal ventricle. The estimated parameters can be used to separate stressed speech from neutral speech because these parameters represent the mechanisms of the vocal folds and airflow variation in the glottis under stress. Experimental evaluations show that the area of laryngeal ventricle has a modulating effect on speech production, and is effective for the classification of stressed speech.

Index Terms: physical parameters, two-mass model, stress classification, airflow pattern, laryngeal ventricle.

1. Introduction

The effect 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. By studying speech under stress, we can improve the performance of speech recognition systems, recognize when people are in a stressed state, and better understand the context in which speakers are 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], who used the statistical properties of acoustic features to recognize emotions from speech in 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 due to a speaker's emotional state [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 when speakers increase the volume of their voices. 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.* examined the frequency domain, and showed how differences in the spectrum of the high frequency band of speech of speakers under stressful workload conditions aimed to catch people committing remittance fraud, and the proposed measure achieved better performance [7].

In 1980, Teager suggested that speech production is a nonlinear process and proposed a nonlinear model [8] [9]. Airflow separation occurs along the walls of the laryngeal ventricle around the false vocal folds, causing variability in airflow characteristics, thereby having modulating effect on speech production. So it is helpful to model airflow patterns in order to characterize speech production. So physical models are proposed to simulate the vocal folds and the vocal tract [10] [11]. Furthermore, Cairns showed that the impact of airflow separation differ markedly between neutral and stressed speech [12]. In physiological systems, it is believed that changes in physical characteristics induced by stressful conditions affect airflow separation [13]. Therefore, it is necessary to develop a physical model of the laryngeal ventricle and false vocal folds in order to understand the variation in airflow characteristics caused by stress.

In our previous work [14][15][16], we estimated parameters for the vocal folds and the vocal tract, based on a two-mass model [17], for the classification of stressed speech. However, the laryngeal ventricle and the false vocal folds are not modeled in the two-mass model, and airflow separation in the glottis has not been considered in our previous works. Therefore, in this paper, we expand the two-mass model to include the airflow patterns in the laryngeal ventricle and around the false vocal folds, and estimate the physical parameters representing muscle tension of the vocal folds and effective area of laryngeal ventricle. A fitting method for the two-mass model is proposed to estimate physical parameters.

This paper is organized as follows. In Section 2, we propose a physical model of the airflow dynamics. In Section 3, the fitting method used to estimate the physical parameters is explained. Experiments are performed in Section 4 to evaluate the obtained parameters and show their corresponding classification performance for identifying neutral and stressed speech. Conclusions are drawn in Section 5.

2. Physical modeling

The two-mass model was proposed by Ishizaka and Flanagan to simulate the process of speech production [17]. The laryngeal part can be depicted using the traditional two-mass model to represent the mechanism of the vocal folds.

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

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

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

$$s_i(x_i) = k_i\left(x_i + \eta x_i^{\,3}\right), \quad (3)$$

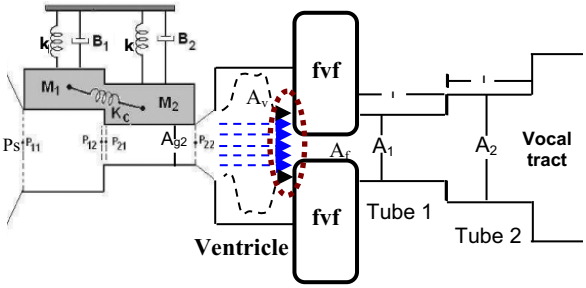

Figure 1 *Sketch of modified two-mass model*

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

The vocal tract is represented by a four-tube model, constructed using a transmission line analogy involving four cylindrical, hard-walled sections. The elemental values of the model are determined by cross-sectional areas $A_1 \cdots A_n$, and cylinder lengths $l_1 \cdots l_n$.

Figure 1 shows a sketch of our proposed model. The laryngeal ventricle and false vocal folds (fvf) are modeled to show the airflow patterns between the vocal folds and the vocal tract.

2.1. Modeling airflow aerodynamics

2.1.1. Pressure drop at the glottis

The aerodynamics of the glottis are modeled using the two-mass model. If subglottal pressure is represented as P_s, then the pressure drops to P_{11} when entering the glottis (at the edge of m_1) according to Bernoulli's equation:

$$P_s - P_{11} = \frac{\rho U_g^2}{2A_1^2}, \tag{4}$$

where ρ is the air density, and U_g is the volume velocity of glottal airflow. A_{g1} is the cross-sectional lower glottal area, which can be 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. Abrupt contractions in the cross-sectional area at the inlet to the glottis cause a vena contracta to occur, which causes an even greater drop in pressure. The drop is determined using the flow measurements from van den Berg:

$$P_s - P_{11} = (1.00 + 0.37)\frac{\rho U_g^2}{2A_{g1}^2}, \tag{5}$$

Along masses m1 and m2, pressure drops as a result of air viscosity:

$$P_{i1} - P_{i2} = \frac{12\mu d_i l_g^2 U_g}{A_{gi}^3}, \qquad i = 1,2 \tag{6}$$

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

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

$$P_{21} - P_{12} = \frac{\rho U_g^2}{2}\left(\frac{1}{A_{g1}^2} - \frac{1}{A_{g2}^2}\right), \tag{7}$$

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

2.1.2. Pressure drop around laryngeal ventricle and false vocal folds

Next, we model airflow patterns around the laryngeal ventricle and false vocal folds. 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:

$$P_{22} - P_v = -\frac{\rho}{2}\cdot\frac{2}{A_{g2}A_E}\left(1 - \frac{A_{g2}}{A_E}\right)U_g^2, \tag{8}$$

where 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. After passing this region, the airflow propagates as a plane wave entering the false vocal folds. Separation causes variations in the effective area of the laryngeal ventricle into the false vocal folds. 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:

$$P_v - P_{f1} = \frac{\rho}{2}\left(\frac{1}{A_f^2} - \frac{1}{A_v^2}\right)U_g^2, \tag{9}$$

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:

$$P_{f1} - P_{f2} = 12\frac{\mu l_f^2 d_f}{A_f^3}U_g, \tag{10}$$

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.

$$P_{f2} - P_1 = -\frac{\rho}{2}\cdot\frac{2}{A_f A_1}\left(1 - \frac{A_f}{A_1}\right)U_g^2, \tag{11}$$

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

The effective area of the ventricle into the false vocal tract A_V can represent the variation in airflow pattern, which has a modulating effect on produced speech. Therefore, it is our assumption that this area parameter can used as an indicator for stress classification.

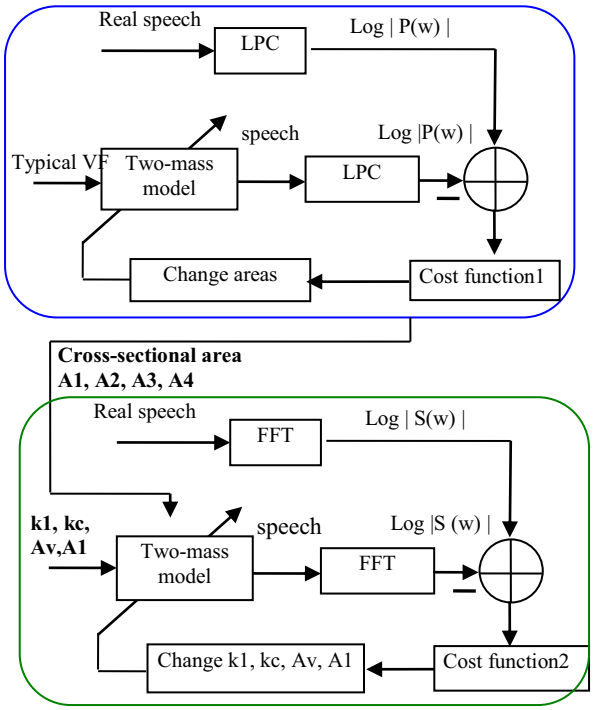

Figure 2 *Method for estimation of physical parameters*

3. Fitting method

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 the glottal source and the vocal tract. A_2, A_3 and A_4, however, are not directly related to the glottal source, and thus have less impact on the interaction, as we showed in [18]. Therefore, parameters k_1, k_c, A_1 and A_V, should be estimated together and selected as feature parameters for stress classification.

The detailed fitting method for estimation of the physical parameters is shown in Figure 2. This method includes two steps. First, vocal tract fitting is performed with a typical vocal fold setting. The outputs of this part of the model are the estimated cross-sectional areas of the four-tube model: A_1, A_2, A_3, and A_4. Cost function 1 (C_1) is defined as the root mean square (RMS) distance between the spectral envelope of simulated and original speech.

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

$$P(\omega) = \frac{1}{|A(\omega)|^2} = \frac{1}{\left|\sum_{k=0}^{P} a_k e^{-j\omega k}\right|^2} \qquad (12)$$

In the second step, A_2, A_3, and A_4 are fixed at obtained values, and A_1 is considered as the initial value for the next fitting. In the second fitting, k_1, k_c, A_V and A_1 are selected as control parameters, and cost function 2 is defined as:

$$C_2 = \frac{1}{N}\sum_{i=1}^{N} \left| \log S(\omega_i) - \log S^*(\omega_i) \right|^2 \qquad (13)$$

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 [19], which is implemented to search for the optimal stiffness parameters which minimize the cost function.

4. Experiments

4.1. 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). 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. All of the vowels were mixed for the vowel-independent condition. The experiments were conducted for each speaker, and all of the results were speaker dependent. Here, we used samples from seven subjects (three male, four 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 450-700 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. The samples were analyzed with 12th-order LPC and the frame size chosen to perform the experiment was 64ms, with 16ms for frame shift.

4.2. Results and analysis

4.2.1. Effect of parameter Av

In this section, we describe experiments which were performed to represent the effect of proposed parameter A_V. We selected the formants (F_1, F_2, F_3), the fundamental frequency (F_0) and the spectral flatness measure (SFM) as stress measurements. Formants depend on the shape of the vocal tract, while F_0 and SFM represent characteristics of the glottal source generated from the vocal folds.

Figure 3 shows the relationship between A_V and these acoustic parameters. It is illustrated that A_V does not significantly affect formants and F_0, however, an increase in A_V does have an impact on SFM. SFM is a measurement quantifying the irregularity of the spectrum, which loses clarity in its harmonic structure in the high frequency band when stress occurs. Our results show that variation in A_V dramatically affects irregularity in the harmonic structure of the spectrum in the high frequency band.

In order to further evaluate the influence of A_V on the spectrum, comparison of spectral distortion for real speech

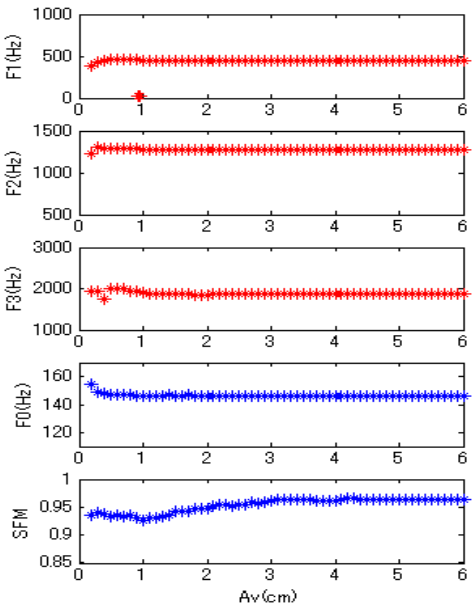

Figure 3 *Impact of* A_V *on acoustic parameters*

and simulated speech with and without estimation of Av, was made. Log-spectral distance (LSD) was used to describe the difference in spectral distortion between real and simulated.

$$\mathrm{LSD} = \sqrt{\frac{1}{f(b)} \sum_{\omega_i \in B(b)} \left(10 log_{10} \left|S^*(\omega_i)\right| - 10 log_{10} \left|S(\omega_i)\right|\right)^2} \quad (14)$$

where $f(b)$ denotes the bandwidth of sub-band b and $B(b)$ consists of a set containing all the discrete frequencies in sub-band b. $S(\omega)$ and $S^*(\omega)$ are the power spectrums of the residual signals for simulated and real speech, respectively. The sub-bands are described in Table 1.

Table 1 *Sub-bands for the spectrum*

	Sub-Band						
	1	2	3	4	5	6	7
Frequency band(Hz)	0-1000 (Hz)	500-1500 (Hz)	1000-2000 (Hz)	1500-2500 (Hz)	2000-3000 (Hz)	2500-3500 (Hz)	3000-4000 (Hz)

We obtained the spectrums of the residual signals of simulated speech by fitting the two-mass model to all of the real speech. The average values for LSD were calculated for all of the speech data. The results for log-spectral distance are illustrated in Figure 4, which shows that there is no difference in the low frequency bands. However, when the high frequency bands are taken into account, the results achieve an improvement in the accuracy of spectrum simulation when using the estimation of A_V, spectral distortion decreases significantly. This indicates that the estimation of A_V can improve simulation accuracy in the high frequency bands.

4.2.2. Evaluation of physical parameters

In this section, we compared the performance of two physical parameter sets, [k_1, k_c] and [k_1, k_c, A_V], to evaluate the effectiveness of proposed parameter A_V. The results are shown in Figure 5. When A_V is taken into account, classification performance is improved. Since the samples selected in the

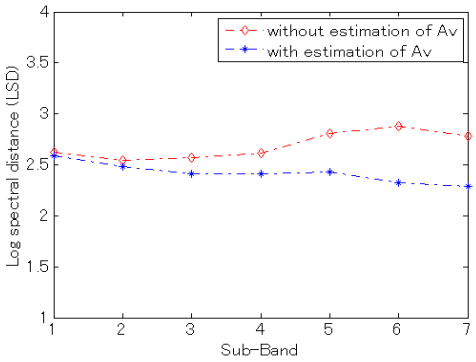

Figure 4 *LSD to evaluate impact of* A_V *on spectrum simulation*

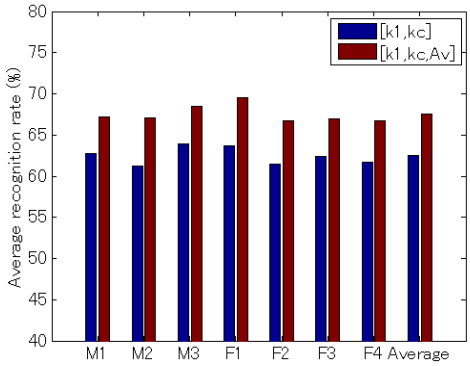

Figure 5 *Classification rates for physical parameters*

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 values indicate that the amount of airflow separation is increasing, causing the effective area at the inlet of the false vocal folds to broaden. This results in variations in the airflow pattern around the false vocal folds, causing a stronger modulation effect on the speech produced.

5. Conclusions

In this paper, we considered the aerodynamics of airflow patterns in the laryngeal ventricle and false vocal folds, and modeled the airflow patterns for the purpose of improving stress classification. A physical parameter representing the effective area of the laryngeal ventricle into the false vocal folds, is explored, which characterizes airflow separation during speech production. An estimation of the physical parameters is performed by fitting the modified two-mass model to real speech. Results show that the proposed physical parameters can lead to improvements in the classification of speech under stress by physically modeling the modulating effect of stress-induced changes in airflow pattern on speech.

6. Acknowledgements

This work was partially supported by the Core Research for Evolutional Science and Technology (CREST) Project of the Japan Science and Technology Agency (JST).

7. 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. and 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] Benkrid, A., Benallal A., and Benkrid K., "Real-time vocal tract modelling", Modelling and Simulation in Engineering 4:1, 2007.

[11] Mathur, S., Story, B. H., & Rodriguez, J. J., "Vocal-tract modeling: Fractional elongation of segment lengths in a waveguide model with half-sample delays", Audio, Speech, and Language Processing, IEEE Transactions on 14.5: 1754-1762, 2006.

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

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

[15] Yao, X., Jitsuhiro, T., Miyajima, C., Kitaoka, N. and Takeda, K., "Classification of stressed speech using physical parameters derived from two-mass model", (INTERSPEECH 2012), Portland, Oregon, USA, Sept. 2012.

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

[17] 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.

[18] Yao, X., Jitsuhiro, T., Miyajima, C., Kitaoka, N. and Takeda, K., "Evaluation for vowel-independent classification of speech under stress based on interaction between the vocal folds and the vocal tract", 2012 Autumn Meeting, Acoustic Society of Japan (ASJ), Shinshu University, Nagano, 1-2-19, 269-272, Sept. 2012.

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