ESTIMATION OF VOCAL TRACT PARAMETERS FOR THE CLASSIFICATION OF 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

In this work, we propose a method for the classification of speech under stress that is based on a physical model. Using this method, the characteristics of the vocal folds and the vocal tract are taken into consideration, based on the process of speech production. In addition to vocal fold parameters, we estimate parameters of the vocal tract representing cross-sectional areas and vocal tract length, by fitting a two-mass model to real speech. Results show that calculation of vocal tract length for each speaker can improve the accuracy of the estimation of other physical parameters. Analysis is performed under vowel-dependent and vowel-independent conditions, showing that the proposed physical features are effective for the classification of neutral and stressed speech.

Index Terms— speech under stress, two-mass model, physical parameters, vocal tract length.

1. INTRODUCTION

Stress is a psycho-physiological state characterized by subjective strain, increased physiological activity, and deterioration of performance [1]. Factors inducing stress on speakers are believed to affect voice quality, and these changes are detrimental to the performance of communication equipment especially automated systems with speech interfaces. Therefore, it has become increasingly important to study speech under stress in order to improve the performance of speech recognition systems.

Researchers have attempted to probe reliable indicators of stress by analyzing acoustic variables. Some external factors (workload, background noise, etc.) and internal factors (emotional state, fatigue, etc.) may induce stress [2]. It has been found that fundamental frequency (F_0) has different characteristics for each emotion [3], and that respiration patterns and muscle tension also change [4]. 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 [5]. Matsuo, *et al.* examined the frequency domain and showed how differences in the spectrum of the high frequency band under stressful workload conditions could be used to catch people committing remittance fraud, and their proposed measure achieved better stressed speech classification performance [6]. Furthermore, the Teager energy operator (TEO) [7] was proposed to explore variations in the energy of airflow characteristics within the glottis for the purpose of stress classification [8]. However, the features examined in these previous works lack a physical basis, and the methods do not consider the whole process of speech production and the airflow pattern in the glottis, which is believed to be essential for effective stress classification.

We therefore propose a speech classification method for identifying speech under stress using parameters estimated from a physical model, based on the working mechanisms of the vocal folds and the vocal tract. The method characterize speech production process and model the airflow pattern in the vocal folds and the vocal tract with the physical model. It is believed that the presence of stress can result in variations in the physical characteristics of physiological systems and influence acoustic interaction between the vocal folds and the vocal tract [9]. The parameters of a physical model can also represent the influence of speaking styles more directly. Therefore a physical model is helpful to estimate the parameters of the physiological system.

In our previous work [10] [11], we estimated only vocal fold parameters, based on a two-mass model, for the classification of stressed speech. The experimental results showed the proposed features for the vocal folds achieved better performance than features derived from traditional methods. However, an assumption was made that the shape of the vocal tract does not change. This is something of an oversimplification. Therefore, in this paper, we concentrate on estimation of vocal tract parameters representing cross-sectional areas and vocal tract length. A fitting method for the two-mass model is proposed to estimate the physical parameters. Furthermore, some dynamic parameters, representing variation in the stiffness of the vocal folds, are also proposed to improve stress classification. In Section 2, the fitting method used to estimate the physical parameters is explained. In Section 3, our experimental results are analyzed to evaluate the obtained parameters and to show their corresponding classification performance for identifying neutral and stressed speech. Finally, we draw our conclusions in Section 4.

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 [12].

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. F_i are the forces acting on the masses.

In this equation, s_i are the equivalent tensions given by

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

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

The viscous loss of the vocal folds can be represented as:

$$r_i = 2\zeta_i\sqrt{m_i k_i} \tag{4}$$

where ζ_i is a damping ratio.

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 $A_1 \cdots A_n$, and cylinder lengths $l_1 \cdots l_n$. The tube model can be represented by an equivalent circuit. The inductances are $L_n = \rho l_n / 2A_n$, the capacitances are $C_n = l_n \cdot A_n / \rho c^2$, and the resistances $R_n = \left(S_n / A_n^2\right)\sqrt{\rho\mu\omega/2}$, where c is the velocity of sound. Here, the tube model has been limited to four cylindrical sections of equal length, $n = 4$. The model is terminated in a radiation load equal to that of a circular piston in an infinite baffle. $L_n = \left(8\rho/3\pi\right)\sqrt{\pi A_n}$, $R_R = 128\rho c / 9\pi^2 A_n$, where A_n is the area of the mouth. Therefore, the differential equations related to the volume velocities of the system are:

$$\begin{aligned}
&\left(R_{k1}+R_{k2}\right)\left|U_g\right|U_g + \left(R_{v1}+R_{v2}\right)U_g + \left(L_{g1}+L_{g2}\right)\frac{dU_g}{dt} + \\
&L_1\frac{dU_g}{dt} + R_1 U_g + \frac{1}{C_1}\int_0^t \left(U_g - U_1\right)dt - P_s = 0 \\
&\left(L_1+L_2\right)\frac{dU_1}{dt} + \left(R_1+R_2\right)U_1 + \\
&\frac{1}{C_2}\int_0^t \left(U_1 - U_2\right)dt + \frac{1}{C_1}\int_0^t \left(U_1 - U_g\right)dt = 0 \\
&\vdots \\
&L_R\frac{d}{dt}\left(U_R + U_L\right) + R_R \cdot U_R = 0
\end{aligned} \tag{5}$$

2.2. Fitting method

2.2.1. Estimation of vocal tract length

One physical source of the inter-speaker variability is the differences in vocal tract length (VTL). Physical differences in VTL are more marked between male and female speakers. VTL can vary from approximately 13 cm for adult females to over 18 cm for adult males [13, 14], and differences in VTL influence spectral formant frequency. Due to the variation caused by differences in VTL, it is necessary to estimate a speaker's vocal tract length in speaker dependent systems.

Estimation of vocal tract length is the first step to be performed. Since VTL is unique for each speaker, all of the neutral speech data in the database from a given speaker is used to estimate the vocal tract length of that speaker. For VTL estimation, real speech coming from a database is analyzed using linear predictive coding (LPC) to reach the spectral envelope. The

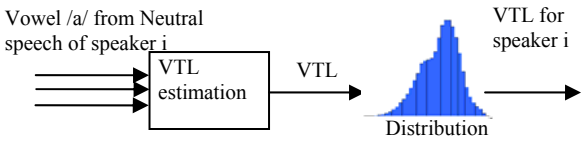

Figure 1 Strategy for VTL estimation

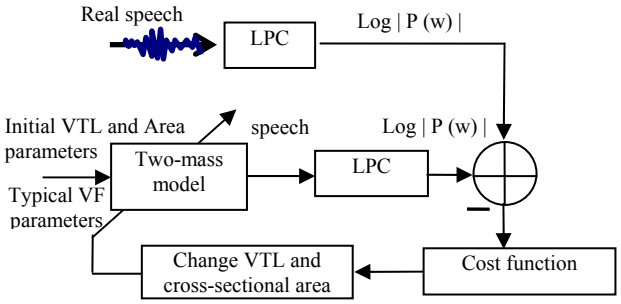

Figure 2 Fitting process for VTL estimation

stiffness parameters are fixed at typical values and are taken as an input. The two-mass model is then fit to the neutral speech of each speaker to estimate the parameters of vocal tract length and cross-sectional areas. For each speaker i, the distribution of occurrence rate $P_i\left(L_{VT}(i,k)\right)$ of VTL $L_{VT}(i,k)$ for all neutral speech is calculated, and we choose the one with the highest occurrence rate as the estimated vocal tract length.

$$L_{VT}^*(i) = arg\,max_{L_{VT}(i,k)}\, P_i\left(L_{VT}(i,k)\right) \tag{6}$$

The algorithm used for VTL estimation is shown in Figure 1. The detailed fitting procedure is shown in Figure 2. The cost function for fitting can be represented as

$$\begin{aligned}
C_1 &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|log\,P(\omega_i) - log\,P^*(\omega_i)\right|^2} \\
P(\omega) &= \frac{1}{\left|A(\omega)\right|^2} = \frac{1}{\left|\sum_{k=0}^{P} a_k e^{-j\omega k}\right|^2}
\end{aligned} \tag{7}$$

where $P(\omega)$ and $P^*(\omega)$ are the LPC spectral envelopes for simulated and real speech respectively.

2.2.2. Estimation of stiffness and cross-sectional areas

The goal of stress classification is to determine from speech data if a specific person is under stress when he or she is speaking. So VTL for each speaker is first calculated using the algorithm described in the section 2.2.1. When speech is input to the system, it is split into several frames, and further estimation of the physical parameters is performed for each frame. The main structure is shown in Figure 3.

Fitting the model to real speech poses a difficulty: the existence of interaction makes it impossible to fit VF and VT separately. It is believed that stiffness parameters k_1, k_c and cross-sectional area A_1, affecting both the glottal source and formants, are related to the acoustic interaction between the glottal source and the vocal tract. L_{VT}, A_2, A_3, and A_4, however, do not influence the glottal source, thus having no impact on the interaction [15]. Therefore, parameters k_1, k_c, and A_1 should be estimated together and selected as feature parameters for classification.

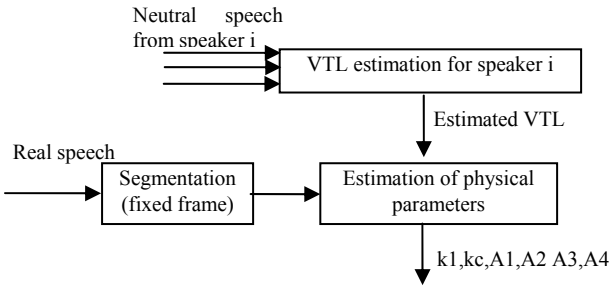

Figure 3 Main structure of the method

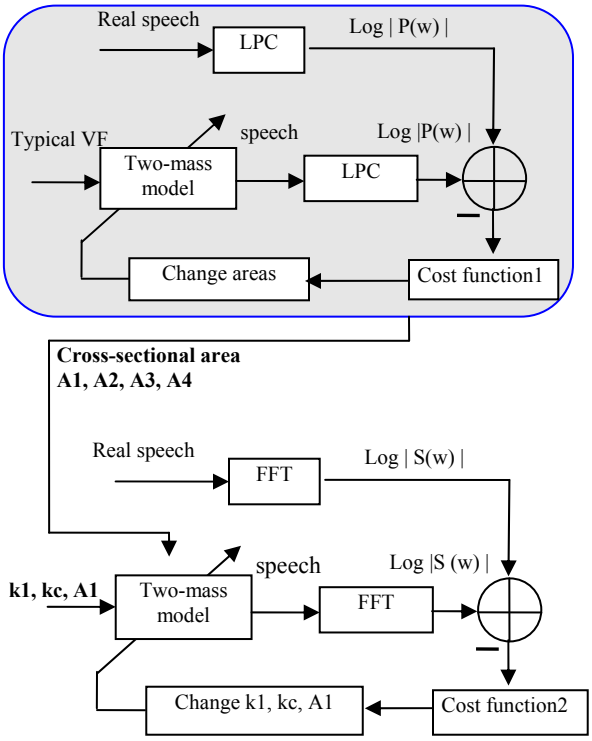

Figure 4 Detail of estimation of physical parameters

The detailed fitting method for estimation of physical parameters is shown in Figure 4. This method includes two steps. First, vocal tract fitting is performed with a typical vocal fold setting. The output of this part of the model is 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 the simulated and the original speech, which is shown in equation (7).

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, and A_1 are selected as control parameters, and cost function 2 is defined as:

$$C_2 = \frac{1}{N}\sum_{i=1}^{N} \left|\log S(\omega_i) - \log S^*(\omega_i)\right|^2 \quad (8)$$

where $S(\omega)$ and $S^*(\omega)$ are the power spectrums of the signals for simulated and real speech after Fourier transform. Optimal values of the physical parameters are estimated using a Nelder-Mead simplex method [16], which is implemented to search for the optimal stiffness parameters resulting in minimizing the cost function.

2.2.3. Dynamic features of stiffness parameters

The muscle tension of the vocal folds is influenced when a speaker produces speech under a stressed condition [8]. The dynamic changes in muscle tension which occur during stressed speech, representing the behavior of the vocal folds, are normally different from the neutral condition. Therefore, variation in the stiffness parameters can be indicators that represent stress. Here, we propose the parameters of first order differentiation of stiffness as features to represent the dynamic changes in muscle tension:

$$\Delta k_1(t) = \frac{\sum_{i=-T}^{T} i \cdot k_1(t+i)}{\sum_{i=-T}^{T} i^2}, \Delta k_c(t) = \frac{\sum_{i=-T}^{T} i \cdot k_c(t+i)}{\sum_{i=-T}^{T} i^2} \quad (9)$$

3. EXPERIMENTS

3.1. Database and experimental setup

In our experiments, we used a database collected by the Fujitsu Corporation containing speech samples from eleven subjects, four male, and seven 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 a 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 randomly chose seven speakers (three male, four female) from eleven subjects to show the classification performance of each speaker respectively in this speaker-dependent system. The number of samples depends on speakers, and the total amount is about 450-700 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. 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. For the calculation of dynamic parameters, the window size chosen was 5 for the first order differentiation of stiffness, with T set to 2 in equation (9).

In classification, all speech samples in database were labeled as neutral or stressed speech. Two mass model were fit to real speech to estimate physical parameters. A linear classifier was trained using the estimated parameters. For evaluation, the physical parameters from a test sample were extracted and classified into stressed or neutral by the trained linear classifier.

3.2. Results and analysis

In the first evaluation, we estimated the vocal tract length of all of the speakers, and a comparison was made. In this experiment, all of the vowels /a/, /i/, /u/, /e/, /o/ were mixed to form the database.

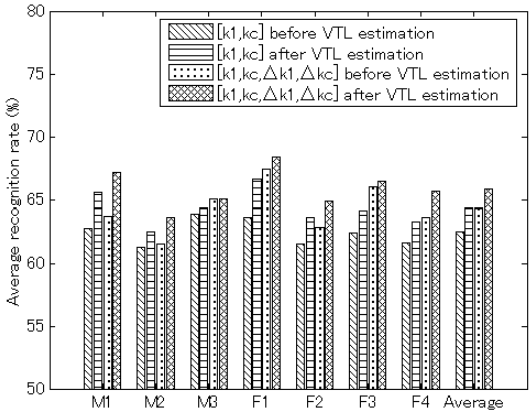

Figure 5 Comparison of performance before and after VTL estimation

The physical parameters were estimated using the proposed fitting method and the estimated parameters were used as features to perform the classification of speech into neutral or stressed speech. The evaluation results for VTL estimation and dynamic parameters are shown in Figure 5. Physical parameters were estimated and Δk_1, Δk_c were calculated. Two features [k_1, k_c] and [k_1, k_c, Δk_1, Δk_c] were compared for their classification performance, both before and after VTL estimation. Our results show that [k_1, k_c, Δk_1, Δk_c] achieve better performance compared to [k_1, k_c] because the proposed first-order differential of physical parameters better represents the dynamic changes in physical parameters for intra-voiced speech. Furthermore, we can see that performance is improved by the estimation of VTL. Since a speaker's vocal tract length is calculated from the neutral speech of that specific speaker, the further estimation of physical parameters will be more accurate, and improvement in classification can be achieved.

In the second experiment, we evaluated the proposed physical parameters. First, 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. The parameters [k_1, k_c, A_1] were estimated from real speech with the obtained VTL for each speaker. Figure 6 compares the classification rates of parameter sets [k_1, k_c], [k_1, k_c, A_1], [k_1, k_c, Δk_1, Δk_c] and [k_1, k_c, Δk_1, Δk_c, A_1]. Comparing the results, we see that [k_1, k_c, A_1] and [k_1, k_c, Δk_1, Δk_c, A_1] achieve better performance under the vowel dependent condition, in which individual vowels are considered separately. A_1 is effective for classification because the shape of the vocal tract doesn't change significantly when considering only one vowel, so A_1 only represents acoustic interaction, thus improving classification performance.

Next, all of the vowels /a/, /i/, /u/, /e/, /o/ were mixed for the vowel-independent condition. Figure 7 shows the results for [k_1, k_c], [k_1, k_c, A_1], [k_1, k_c, Δk_1, Δk_c] and [k_1, k_c, Δk_1, Δk_c, A_1]. The results show that the classification rate of [k_1, k_c, A_1] and [k_1, k_c, Δk_1, Δk_c, A_1] under the vowel-independent condition is reduced. This is because the cross-sectional area A_1, not only affect the interaction between VF and VT, but also determines vocal tract shape, and thus relies on vowel information. Furthermore, it is shown that the parameter set for the vocal folds [k_1, k_c] is able to reach the same classification rate as the results obtained under the vowel dependent condition, and can maintain their performance under the vowel-independent condition. Therefore, it is proven that A_1 is effective under the vowel-dependent condition, while

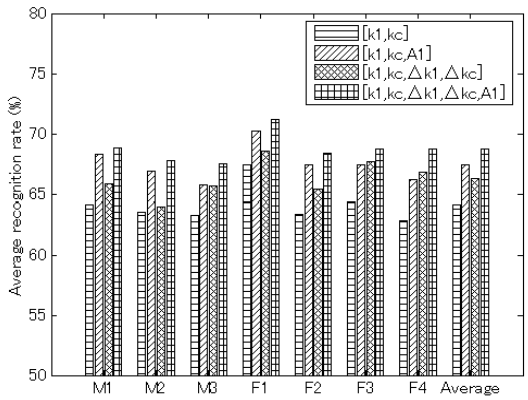

Figure 6 Evaluation under vowel-dependent condition. The parameters are calculated from the real speech after the estimation of vocal tract length for individual vowel.

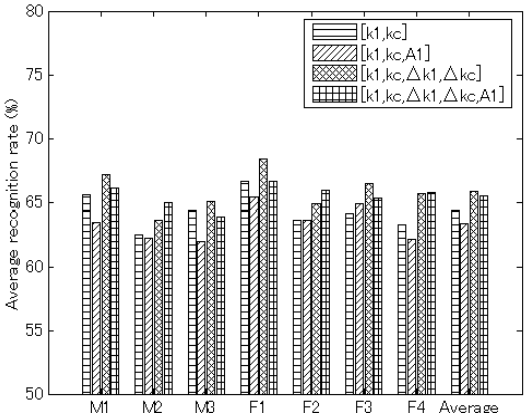

Figure 7 Evaluation under vowel-independent condition. The parameters are estimated from the real speech after the estimation of vocal tract length for all the mixed vowels.

under the vowel-independent condition, [k_1, k_c] is more useful for stress classification.

4. CONCLUSION

In this paper, we proposed a method for stress classification based on a physical model which takes into consideration the characteristics of the vocal folds and the vocal tract. In addition to vocal fold parameters, the physical parameters for the vocal tract, representing cross-sectional areas and vocal tract length, were estimated by fitting the two-mass model to real data. Experiments were performed to show that the calculation of vocal tract length for each speaker improves the estimation accuracy of other physical parameters. Physical parameters were analyzed to show that A_1, cross-sectional area of the vocal tract in the supraglottis, is effective for the classification of neutral and stressed speech under the vowel-dependent condition, while stiffness of the vocal folds achieves better classification performance in the vowel independent condition..

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).

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, 1994.

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

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

[5] J. Whitmore, S. Fisher, "Speech during sustained operations", Speech Commun, vol. 20, pp. 55–70, 1996.

[6] 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).

[7] 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.

[8] 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.

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

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

[11] X. Yao, T. Jitsuhiro, C. Miyajima, N. Kitaoka, K. Takeda, "Classification of stressed speech using physical parameters derived from two-mass model", 13th Annual Conference of the International Speech Communication Association (INTERSPEECH 2012), Portland, Oregon, USA, Sept. 2012.

[12] 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.

[13] L. Lee, R.C. Rose, "Speaker normalization using efficient frequency warping procedures", Proc. IEEE ICASSP96. vol. 1, pp. 353-356, 1996.

[14] L. Lee, R.C. Rose, "A Frequency Warping Approach to Speaker Normalization", IEEE transactions on speech and audio processing, vol. 6, no. 1, pp. 49-60, 1998.

[15] X. Yao, T. Jitsuhiro, C. Miyajima, N. Kitaoka, K. Takeda, "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, pp.269-272, Sept. 2012

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