EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Classification of speech under stress based on modeling of the vocal folds and vocal tract

Xiao Yao[1*], Takatoshi Jitsuhiro[1,2], Chiyomi Miyajima[1], Norihide Kitaoka[1] and Kazuya Takeda[1]

Abstract

In this study, we focus on the classification of neutral and stressed speech based on a physical model. In order to represent the characteristics of the vocal folds and vocal tract during the process of speech production and to explore the physical parameters involved, we propose a method using the two-mass model. As feature parameters, we focus on stiffness parameters of the vocal folds, vocal tract length, and cross-sectional areas of the vocal tract. The stiffness parameters and the area of the entrance to the vocal tract are extracted from the two-mass model after we fit the model to real data using our proposed algorithm. These parameters are related to the velocity of glottal airflow and acoustic interaction between the vocal folds and the vocal tract and can precisely represent features of speech under stress because they are affected by the speaker's psychological state during speech production. In our experiments, the physical features generated using the proposed approach are compared with traditionally used features, and the results demonstrate a clear improvement of up to 10% to 15% in average stress classification performance, which shows that our proposed method is more effective than conventional methods.

Keywords: Speech under stress; Stress classification; Physical parameters; Two-mass model; Vocal folds; Vocal tract

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 include workload, background noise, emotions, physical environmental factors (e.g., G-force), and fatigue. These factors are believed to affect voice quality and 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, to recognize when people are in a stressed state and to understand contexts in which speakers are communicating.

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]. The first investigations of emotional speech were conducted in the mid-1980s, using the statistical properties of acoustic features in order to detect emotions from speech [3,4]. It has been found that fundamental frequency (F_0) has different characteristics for each emotion [5] and that respiration patterns and muscle tension also change [6]. The influence of the Lombard effect on speech recognition has also been examined [7,8]. Selected acoustic features have been analyzed, such as amplitude and distribution of spectral energy, and it was found that spectral energy shifted to higher frequencies for consonants in the presence of loud background noise. 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 [9,10]. 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 classification performance [11]. Furthermore, the Teager energy operator (TEO) [12] was proposed to explore variations in the energy of airflow characteristics within the glottis for the purpose of stress classification [13]. However, the features examined in these previous

* Correspondence: xiao.yao@g.sp.m.is.nagoya-u.ac.jp
[1]Graduate School of Information Science, Nagoya University, Nagoya, Aichi, Japan
Full list of author information is available at the end of the article

studies lack a physical basis, and the methods do not consider the whole process of speech production, which is believed to be essential for effective classification of speech under stress.

We propose a stressed speech classification method based on a physical model characterizing the vocal folds (VF) and the vocal tract (VT). This method can represent the process of speech production and model airflow patterns in the vocal folds and the vocal tract, which are essential for stress classification. In this physical model, changes in the physical characteristics of the vocal folds, such as muscle tension, have a modulating effect on the formants, while the shape of the vocal tract can also influence the glottal source because of the interaction between the vocal folds and the vocal tract. It is believed that the presence of stress can result in variations in the physical characteristics of physiological systems and influence the acoustic interaction between the vocal folds and the vocal tract [2]. The parameters of the physical model also help represent the influence of speaking style more directly and clearly. Therefore, a physical model is helpful to estimate the parameters of the physiological system.

An early but still prominent physical model is the source-filter model [14], which models speech as the combination of a glottal source (such as the vocal folds), and a linear acoustic filter representing the vocal tract and its radiation characteristic. An important assumption that is often made in the use of the source-filter model is independence of the source and filter. In such cases, the model should more accurately be referred to as the 'independent source-filter model'. In 1961, Wong proposed a linear model of speech production using a lossless tube model of the vocal tract [15]. In 1979, a linear source tract model was proposed to model the glottal source, the vocal tract, and radiation impedance as linear filters, using covariance analysis [16]. However, the vocal tract and vocal folds do not function independently of each other instead there is some form of interaction between them [17], which results in significant changes in fundamental frequency and formant characteristics.

The two-mass model is a physical model, which attempts to simulate the physical process of vocal fold vibration, characterizing the vocal folds and the vocal tract, and to also model the effect of glottis-vocal tract interaction [18]. Parameters affected under stressed conditions are extracted from the physical model and are used as features to identify speech under stress more precisely. We use the two-mass model as a physical model, and our proposed method estimates the values of parameters included in the model from input speech. To identify speech under stress, we evaluate parameters affected by stress.

In this paper, we propose a method for fitting a physical model to real speech in order to estimate the physical parameters which characterize the vocal folds and the vocal tract. For the physical model, a two-mass model connected to a four-tube model is used to simulate the process of speech production. The physical parameters (stiffness, vocal tract length, and cross-sectional areas of the vocal tract) are estimated by fitting the model to real speech. The estimated parameters can be further analyzed and proposed as features for the classification of neutral and stressed speech. Furthermore, different cost functions are proposed to compare classification performance. As a result, stiffness of the vocal folds and cross-sectional areas of the vocal tract are selected as features for the classification of neutral and stressed speech.

The paper is organized as follows: In Overview, an overview of our method is presented. Physical parameters, related to the vocal folds and the vocal tract, based on the two-mass model are described as features for classification in Physical parameters. This is followed by the presentation of a fitting algorithm for real speech data in Estimation method to help estimate the physical parameters. Classification describes the classification method used for evaluation. In Evaluation, experiments are performed to evaluate the obtained parameters and show their corresponding classification performances when separating neutral and stressed speech. Finally, we draw our conclusions in Conclusion.

2. Overview

An overview of our work is shown in Figure 1. It includes the three steps needed to perform stressed speech classification: proposal of physical parameters, parameter estimation by fitting them to the two-mass model, and the classification of neutral and stressed speech.

Initially, we propose physical parameters considered likely to be useful, which include stiffness parameters of the vocal folds, vocal tract length, and cross-sectional areas of the vocal tract. These parameters characterize the behavior of vocal folds and the shape of the vocal tract. Furthermore, the relationship between the selected physical parameters and acoustic parameters has been shown to represent characteristics of the interaction between the vocal folds and the vocal tract.

The proposed physical parameters are then estimated by fitting the two-mass model to real speech. An algorithm based on the analysis-by-synthesis method is proposed for fitting the model to real speech. The Nelder-Mead simplex method [19] is used as a search strategy in order to find the optimal physical parameters. An iteration method is performed for vocal fold fitting and vocal tract fitting to estimate parameters, because there is interaction between the VF and VT.

For classification, a linear classifier is trained using utterances from each speaker. Currently, a simple linear classifier based on Euclidean distance is used for classification. Also, since we only have speech data for a small number of speakers, we evaluate our proposed method as a speaker-dependent system.

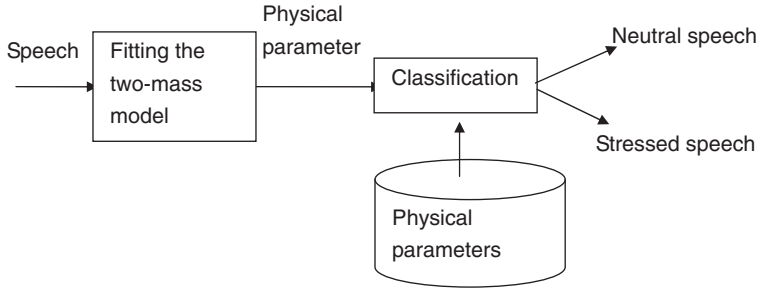

Figure 1 A block diagram of our proposed approach. It includes the three steps necessary to perform classification of stressed speech.

3. Physical parameters

A method which fits the two-mass model to real speech is proposed for classifying speech under stress. Some of the physical parameters characterizing the vocal folds and the vocal tract are estimated. The two-mass vocal fold model was originally proposed by Ishizaka and Flanagan to simulate the process of speech production [18]. We propose three types of feature parameters extracted from the two-mass model: stiffness, vocal tract length, and cross-sectional area of the entrance of the vocal tract. In the following sections, we will define these parameters and describe their characteristics.

3.1 Stiffness

The stiffness parameters are related to muscle tension in the vocal folds. Generally, the stiffness of the vocal folds is considered to depend mainly on two muscles: the cricothyroid muscle (CT) and thyroarytenoid muscle (TA) [16]. 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.

Figure 2 shows a sketch of the model. Each vocal fold is represented by a mass-spring-damper system, joined with a coupling stiffness [18]. It is represented as:

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

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

Tissue elasticity (or 'spring') s_i represents the tension of the vocal folds, which depends on the contraction of different muscles. The equivalent tensions are given by:

$$s_i(x_i) = k_i\left(x_i + \eta x_i^3\right), \quad i = 1, 2, \quad (3)$$

whose notations and variables are documented in Table 1.

Stiffness parameters are the main factors relating to fundamental frequency, and they can also determine the amplitude of the glottal area and glottal volume velocity [20], so source excitation is significantly influenced by the degree of stiffness. During the production of speech, the natural frequency of the vocal folds is determined by both their mass and stiffness. However, in order to

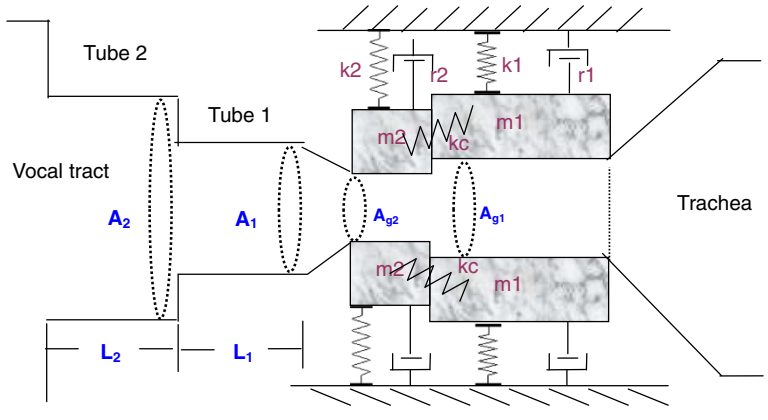

Figure 2 Structure of the two-mass model used to simulate the vocal folds and the vocal tract. The vocal folds are represented by a mass-spring-damping system, coupled with a four-tube model. In this model, m denotes a mass, k_1 and k_2 are linear stiffnesses, k_c is the coupling stiffness connecting the two masses, and r_1 and r_2 are the viscous resistances. L and A represent the length and cross-sectional area of the vocal tract, respectively.

Table 1 Notations and variables in the two-mass model for the vocal folds

Notation/variable	Description
m_i	The masses
x_i	The horizontal displacements measured from the rest (neutral) position x_0
r_i	The equivalent viscous resistances
s_i	The force related to tissue elasticity
F_i	The force of airflow, which is determined by subglottal pressure
k_i	The stiffness coefficients
k_c	The coupling stiffness
η	A coefficient of the nonlinear relations

simplify the estimation algorithm, only the stiffness parameters are estimated, with mass fixed as a constant.

3.2 Vocal tract length and cross-sectional area

The supraglottic area includes the structures that lie above the true vocal folds and below the base of the tongue. The anatomical structures present in this area that are important to speech production lie posterior to the epiglottis. They include the ventricle, false vocal folds, epiglottis, arytenoids, laryngeal aspects of the aryepiglottic folds, and vestibule [21].

The two-mass model is connected to a four-tube model representing the vocal tract [18]. 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 total length of the vocal tract is defined as L_{VT}. The tube model can be represented by an equivalent circuit, as shown in Figure 3. The inductances $L_n = \rho l_n / 2A_n$, the capacitances $C_n = l_n \cdot A_n / \rho c^2$, 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, $n = 4$. In this study, the model is limited to only vowel articulation (as vowels were the subject of the experiments) and modal voice production. These assumptions greatly simplify the modeling of the vocal tract and the glottal source. In this paper, we use a four-tube model to simulate the vocal tract, which followed the original paper [18]. Furthermore, in the following analysis, we propose A_1 as one of our feature parameters because the other areas, A_2, A_3, and A_4 are less effective on classification than A_1. Thus, we currently consider the four-tube model to be sufficient.

The model is terminated in a radiation load equal to that of a circular piston in an infinite baffle. $L_n = (8\rho/3\pi)\sqrt{\pi A_n}$, $R_R = 128\rho c/9\pi^2 A_n$, where A_n is the area of the mouth. The notations and variables are documented in Table 2.

Therefore, the differential equations related to the volume velocities of the system are:

$$\begin{aligned}
&(R_{k1}+R_{k2})\left|U_g\right|U_g + (R_{v1}+R_{v2})U_g + \left(L_{g1}+L_{g2}\right)\frac{dU_g}{dt} + \\
&L_1\frac{dU_g}{dt} + R_1U_g + \frac{1}{C_1}\int_0^t\left(U_g - U_1\right)dt - P_s = 0 \\
&\\
&(L_1+L_2)\frac{dU_1}{dt} + (R_1+R_2)U_1 + \\
&\frac{1}{C_2}\int_0^t(U_1 - U_2)dt + \frac{1}{C_1}\int_0^t\left(U_1 - U_g\right)dt = 0 \\
&\vdots \\
&L_R\frac{d}{dt}(U_R+U_L) + R_R\cdot U_R = 0,
\end{aligned} \tag{4}$$

where $R_{v1} = 12\frac{\mu l_g^2 d_1}{A_{g1}^3}$, $R_{v2} = 12\frac{\mu l_g^2 d_2}{A_{g2}^3}$, $L_{g1} = \frac{\rho d_1}{A_{g1}}$, $L_{g2} = \frac{\rho d_2}{A_{g2}}$, $R_{k1} = \frac{0.19\rho}{A_{g1}^2}$, and $R_{k2} = \frac{\rho\left[0.5-\frac{A_{g2}}{A_1}\left(1-\frac{A_{g2}}{A_1}\right)\right]}{A_{g2}^2}$.

The length of the vocal tract and its cross-sectional areas are the main parameters which determine the

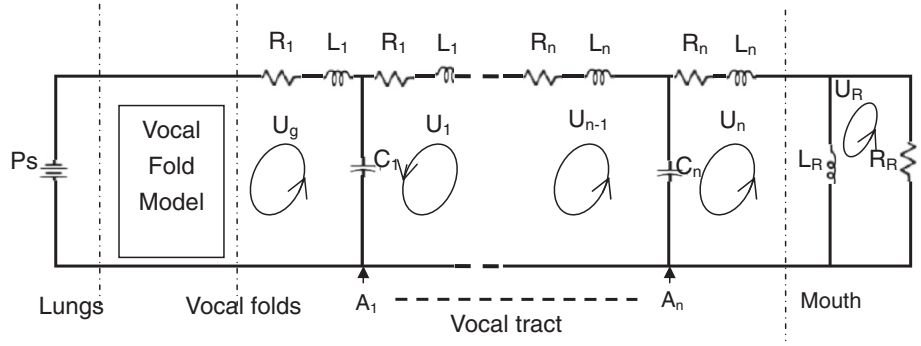

Figure 3 Circuit equivalent of the tube model, representing the production of voiced sound in the vocal tract. Current U_g denotes the velocity of airflow in the glottis. Inductance, capacitance, and resistance depend on the length and area of the vocal tract.

Table 2 Notations and variables in the two-mass model for the vocal tract

Notation/variable	Description
A_i	The cross-sectional areas in the tube model
l_i	The cylinder lengths in the tube model
d_i	The thickness of m_1 and m_2
A_{g1}, A_{g2}	The cross-sectional areas of the glottis
U_g	The average volume velocity across the glottal area
c	The velocity of sound
ρ	The air density
ω	The radian frequency

shape of the vocal tract and have a significant impact on the distribution of formants. Vocal tract length and cross-sectional areas of the tube model are computed from real speech.

3.3 Relationship between physical parameters and acoustic parameters

In this section, we describe experiments which were performed to represent the presence of acoustic interaction and show the relationship between physical and acoustic parameters. Aerodynamics in the glottis is modeled using the two-mass model. In order to clarify the relationship between physical and acoustic parameters, we will first briefly describe the main equations representing the aerodynamics of speech production.

If subglottal pressure is represented as P_s, then air pressure drops to P_{11} when air enters the glottis (at the edge of m_1) according to Bernoulli's equation. The abrupt contraction in the cross-sectional area at the inlet to the glottis causes a phenomenon called vena contracta, which causes the air pressure to undergo an even greater drop. The drop is determined by the flow measurements of van den Berg:

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

where ρ is the air density, U_g is the volume velocity of glottal airflow, and A_{g1} is 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 fold and x_0 is the displacement when the vocal folds are in the rest position.

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

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

where μ is the air viscosity coefficient and d_1 is the width of m_1.

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.

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:

$$P_1 - P_{22} = \frac{1}{2}\rho\frac{U_g^2}{A_{g2}^2}[2N(1-N)], \tag{8}$$

where P_1 is the pressure at the inlet of the vocal tract. Here, the parameter N is defined as $N = A_{g2} / A_1$, where A_1 is the area of the entrance to the vocal tract. N denotes the difference in area between the outlet of the vocal folds and the inlet of the vocal tract, which is significant to the acoustic interaction between the vocal folds and the vocal tract [18]. Since glottal area A_{g2} does not change significantly during the oscillation of the vocal folds, A_1 is the parameter relating to the acoustic interaction.

In Equation 4, it is shown that airflow velocity U_g depends on both the stiffness of the vocal folds and area of the entrance to the vocal tract A_1. Therefore, it is our assumption that parameters k_1, k_2, k_c, and A_1 related to velocity have an impact on acoustic interaction. In this paper, experiments are performed to represent the presence of this interaction by showing the relationship between physical and acoustic parameters. Due to the presence of these interactions, changes in the oscillation of the vocal folds affect the distribution of formants, and different shapes of the vocal tract (length and area) also influence the glottal source. Table 3 lists the physical and acoustic parameters.

We first examine how stiffness parameters impact the distribution of formants. First, we fixed the shape of the vocal tract and examined how variation in the stiffness parameters of the vocal folds affects the shift of formants. The vocal tract model was represented by a standard tube configuration for the vowels /a/ and /e/ [22]. In order to reduce the number of parameters to be estimated and simplify the proposed method, typical values were adopted for the configuration of the tube model. Therefore, as typical values, the length chosen

Table 3 Physical and acoustic parameters

Parameter	Variable
Physical	k_1, k_c, A_1, A_2, A_3, L_{VT}
Acoustic	F_0, F_1, F_2, F_3

for the vocal tract was L_{VT} = 16 cm, with each element l_i = 4 cm, and the cross-sectional area was fixed at A_1 = 0.8 cm^2, A_2 = 0.4 cm^2, A_3 = 3 cm^2, and A_4 = 8 cm^2 for /a/ and A_1 = 1 cm^2, A_2 = 8 cm^2, A_3 = 8 cm^2, and A_4 = 8 cm^2 for /e/. When a specific stiffness is checked, the other stiffness parameters are fixed at typical values. We changed stiffness parameters k_1 (20 to 240 kdyn/cm), k_2 (2 to 40 kdyn/cm), and k_c (2.5 to 70 kdyn/cm) to examine variation in formants. Formant estimation is based on modeling vocal tract frequency response using linear predictive coding (LPC) techniques. It estimates formant frequencies from the all-pole model of the vocal tract transfer function.

Figure 4 shows the relationship between the stiffness parameters and different formants. It shows that k_2 does not significantly influence formants, but that first and second formants will shift their location to a lower frequency with the increase of k_1, although there is no significant change in the third formant (F_3). A similar phenomenon occurs for k_c. When k_c decreases, F_1 also has a tendency to shift to a lower frequency, while F_2 and F_3 are less influenced by the variation of k_c. Therefore, it is shown that stiffness parameters k_1 and k_c can affect the distribution of formants and that the first and second formants are easily affected by acoustic interaction.

Next, we fixed the configuration of the vocal folds and examined how variation of the cross-sectional area of the vocal tract impacts the fundamental frequency (F_0) of speech. Stiffness was fixed at typical values k_1 = 80,000 dyn/cm, k_2 = 8,000 dyn/cm, and k_c = 25,000 dyn/cm to check how the fundamental frequency changes with the area function. When checking the impact of a specific area, other areas and vocal tract length (VTL) were fixed at typical values for /a/ or /e/. When considering VTL, all the cross-sectional areas were fixed at typical values. We then change the cross-sectional area or VTL to examine their impact on F_0. The variation range for VTL was 13 to 19 cm, and for cross-sectional area of VT, the range was 0.1 to 20 cm. The algorithm for estimation of the fundamental frequency of speech is YIN [23]. It is based on the well-known autocorrelation method, with a number of modifications that combine to prevent error.

Figure 5 shows the relationship between the vocal tract parameters (vocal tract length and cross-sectional area) and fundamental frequency. It shows that VTL has less impact on F_0 and only determines the distribution of formants. However, an increase in cross-sectional area A_1 can cause F_0 to change significantly. While cross-sectional areas A_2 and A_3 also have an impact on F_0 to some extent, but their influence is insignificant compared to A_1.

Therefore, it is our conclusion that stiffness of the vocal folds and cross-sectional area A_1 affect both the fundamental frequency and formants and, further, the interaction between the vocal folds and the vocal tract.

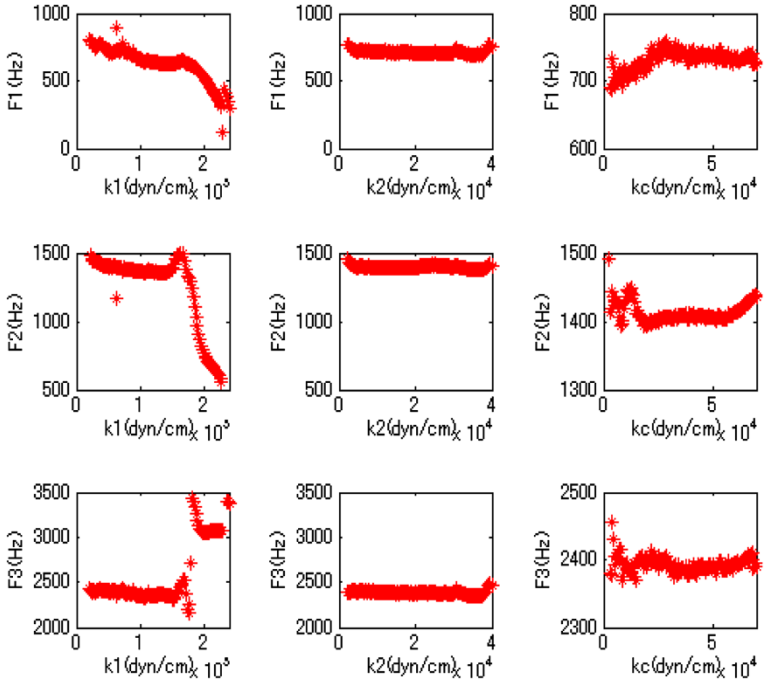

Figure 4 Impact of stiffness parameters in vocal folds on formants.

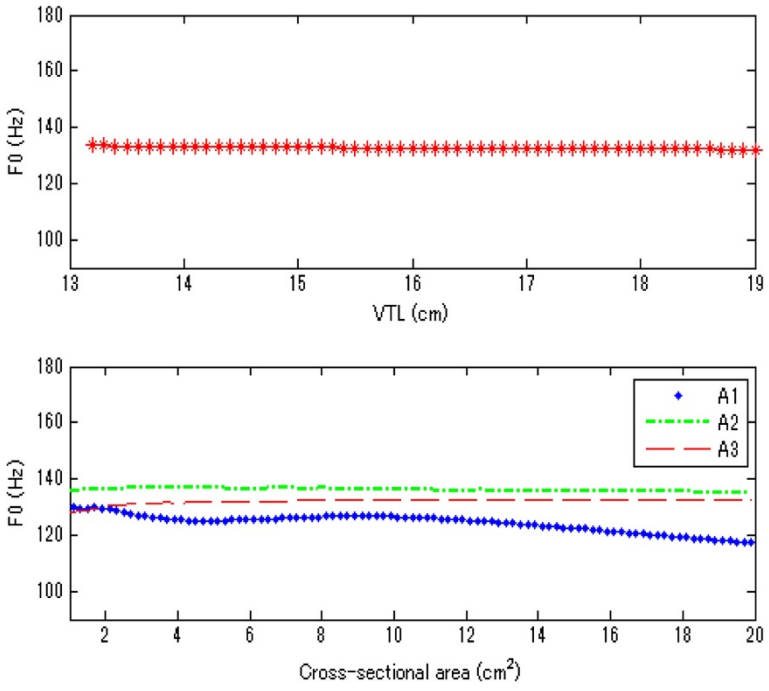

Figure 5 Impact of vocal tract length and cross-sectional area of vocal tract on fundamental frequency.

3.4 Parameters representing stress

In Relationship between physical parameters and acoustic parameters, the experimental results show that stiffness of the vocal folds and cross-sectional area A_1 have an impact on the interaction between the vocal folds and the vocal tract. It is believed that the variations in acoustic interaction differ markedly between neutral and stressed speech [2], so stiffness and A_1 should be selected as parameters for representing stress.

In theory, Equation 8 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. Thus, 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.

Variation in the stiffness of the vocal folds influences the time span of glottal opening and closing phases and causes glottal airflow to accelerate in the glottis, thus impacting the velocity of glottal airflow. Therefore, we can also assume that stiffness parameters can be potential parameters for stress detection.

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 [24]. Since glottal area A_{g2} does not change significantly during the oscillation of the vocal folds, A_1 is the main factor determining the pressure difference between the inside and outside of the glottis and has an impact on the acoustic interaction between VF and VT. Based on these considerations, we also make the assumption that A_1 is an effective parameter for stress classification.

4. Estimation method

4.1 Algorithm for fitting

The goal of stress classification is to determine from speech data if a specific person is under stress when he or she is speaking. 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. VTL for each speaker is first calculated; then, the obtained VTL is input as a known parameter. Then, the two-mass model is fit to each speech sample to simulate the vocal folds and the vocal tract. An outline of our method is shown in Figure 6.

In the first step, estimation of VTL is performed. Since VTL has no impact on the glottal source, it can be estimated separately. Because VTL varies with each speaker, all of the neutral speech data for vowel /a/ from each speaker is used to estimate the vocal tract length of that speaker. Here, we mainly consider the neutral speech for each speaker in the database. During VTL estimation, real speech from a database is analyzed using LPC to obtain the spectral envelope. The 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

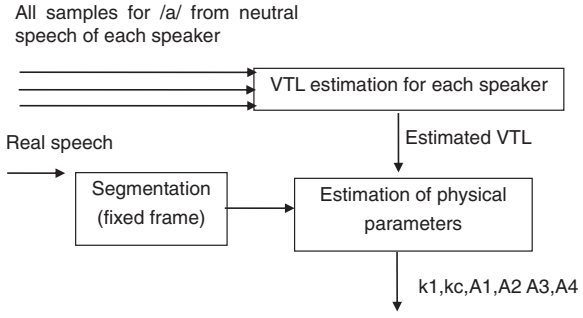

Figure 6 Block diagram showing the outline of our method.

speaker to estimate the parameters of vocal tract length and cross-sectional area. Nelder-Mead simplex method [19] is used to search for the optimal values for fitting. For each speaker i, the probability distribution $P_i(L_{\mathrm{VT}}(i, k))$ of VTL $L_{\mathrm{VT}}(i, k)$ for all neutral speech is calculated, and we choose the one with the highest probability as the estimated vocal tract length.

$$L_{\mathrm{VT}}(i)^* = \arg\max_{L_{\mathrm{VT}(i,k)}} P_i(L_{\mathrm{VT}}(i,k)). \quad (9)$$

The detailed fitting procedure is the same as that used for vocal tract fitting described below, which is shown in Figure 7. Equation 12 is used as the cost function.

In the next step, the estimated VTL of this speaker, which was obtained during the first step, is used, and the two-mass model is fit to the real speech to estimate the other physical parameters. Fitting the model to real speech poses a difficulty: estimation of too many parameters may make the fitting method unstable. The solution to this problem is to split the process into two main parts so that the VF and VT are fit with two different cost functions. However, the existence of interaction between VF and VT makes it impossible to fit VF and VT separately, and changes in the stiffness parameters and in A_1 in the tube model can influence both formants and the glottal source. An alternative is to perform iteration when fitting the vocal folds and the vocal tract. Thus, an iteration method is used for vocal fold and vocal tract fitting, which are accomplished as follows. Figure 8 shows the structure of the fitting algorithm.

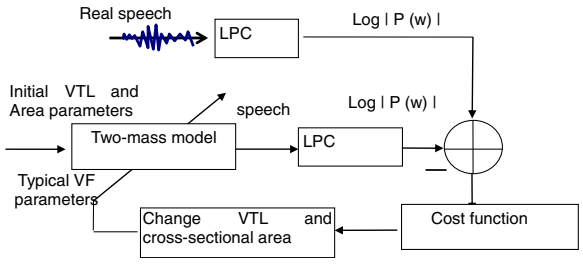

Figure 7 Block diagram showing the details of estimation of vocal tract length.

For vocal tract fitting, stiffness parameters are fixed at typical values and are taken as an input to vocal tract fitting. The parameters for the cross-sectional areas are then estimated. Next, the obtained areas are used as an input for vocal fold fitting, and the two-mass model is fit to estimate the new stiffness parameters. When current stiffness differs significantly from the typical value, the corresponding formants are also affected, and some variations can occur. In such cases, vocal tract fitting needs to be performed again. We take iterations for the two parts until the results reach convergence.

The detailed structure of vocal tract fitting and vocal fold fitting is shown in Figures 9 and 10. Vocal tract fitting includes two steps. First, real speech from a database is analyzed using LPC to calculate the spectral envelope. In the second step, a simulation is performed using the two-mass model to produce speech using an initial area function. The same spectral envelope is calculated from the simulated speech and is compared with the one obtained in the first step to find the difference between them. The difference between the simulated spectrum and the target spectrum is represented by a cost function. The area function is then varied, and glottal flow is simulated until the cost function reaches a minimum. Optimal values of the physical parameters are then estimated using the Nelder-Mead simplex method [19]. Cost function 2 is used in vocal tract fitting. In this paper, we utilize four cost functions in order to compare classification performance, which are described in Cost functions for vocal tract fitting.

The Nelder-Mead algorithm is a simplex method for finding the minimum of a function involving several variables. It is a direct search method and does not require the calculation of a derivative. We use the Nelder-Mead method based on the comparison of the values of the cost function at the $n + 1$ vertices for n-dimensional decision variables to solve our optimization problem. Here, we select A_1, A_2, A_3, and A_4 as variables in vocal tract fitting. Each calculation will generate a new vertex for the simplex. If this new point is better than at least one of the existing vertices, it replaces the worst vertex. The simplex vertices are changed through reflection, expansion, shrinkage, and contraction operations in order to find an improved solution to estimate the parameters. Optimal values of the physical parameters are estimated using the Nelder-Mead simplex method, which is implemented to search for the optimal physical parameters to minimize the cost function.

Vocal fold fitting uses the same process as vocal tract fitting, with the difference that the residual signal is obtained using LPC analysis, and the spectrum of the residual signal is available to construct the cost function 1

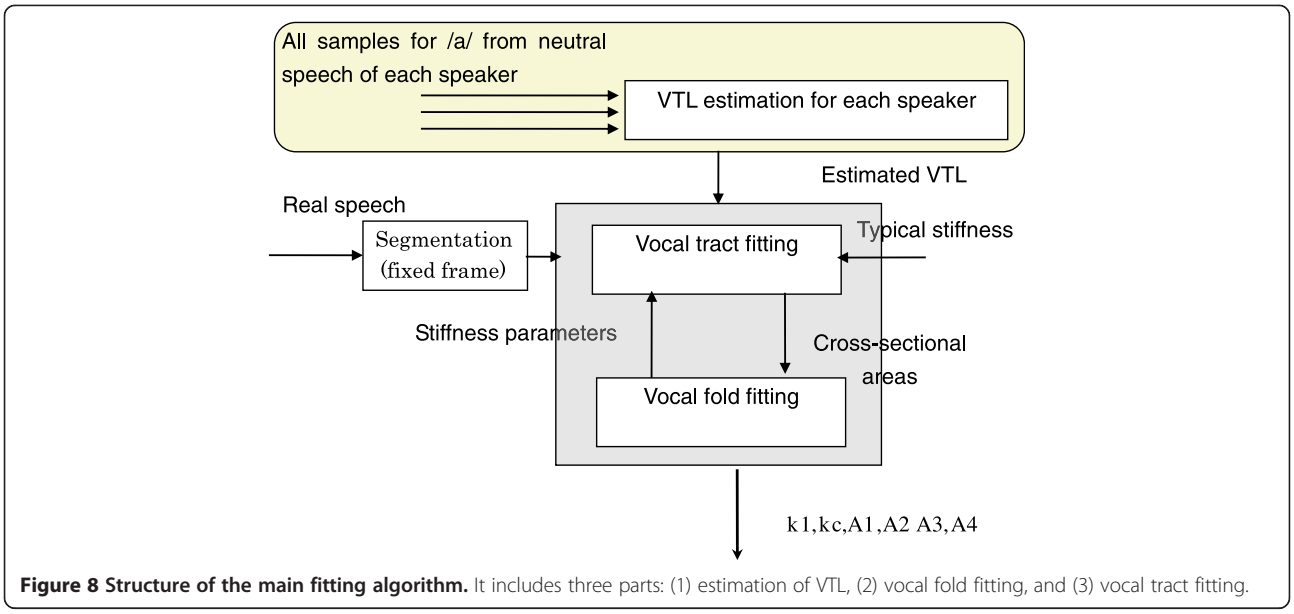

Figure 8 Structure of the main fitting algorithm. It includes three parts: (1) estimation of VTL, (2) vocal fold fitting, and (3) vocal tract fitting.

in Figure 10 for vocal fold fitting. It is used to evaluate the difference in the spectrum of the residual signal in vocal fold fitting, which is described as:

$$C = \frac{\sum_{i=1}^{fs/2} |S^*(\omega_i) - S(\omega_i)|^2}{\sum_{i=1}^{fs/2} |S(\omega_i)|^2}, \quad (10)$$

where $S(\omega)$ and $S^*(\omega)$ are the power spectrums of the residual signal for simulated and real speech, respectively. Here, we select the stiffness parameters k_1, k_2, and k_c as variables for vocal tract fitting.

Here, we use the residual signal from LPC analysis to estimate the parameters of the vocal folds. The LPC model is based on a mathematical approximation of the vocal tract. We use it to remove the effect of the vocal tract and obtain the residual signal to estimate the stiffness parameters with generated cost functions. In order to make a comparison with the spectrum of the residual signal from real speech, an LPC inverse filter is used for the simulated speech to obtain the residual signal. Our target here is to evaluate the similarity of the spectrums of residual signals both from real and simulated speech instead of representing the source wave. The aim of this paper is to classify speech under stress. It is believed that the main differences between neutral and stressed speech are focused on the harmonic structure of the spectrum of residual signal [11]. Thus, in this study, obtaining the residual signal using LPC can work well for showing the harmonic structure of the spectrum.

4.2 Cost functions for vocal tract fitting

As for the definition of cost function 2, we utilized four different cost functions in order to compare their classification performance.

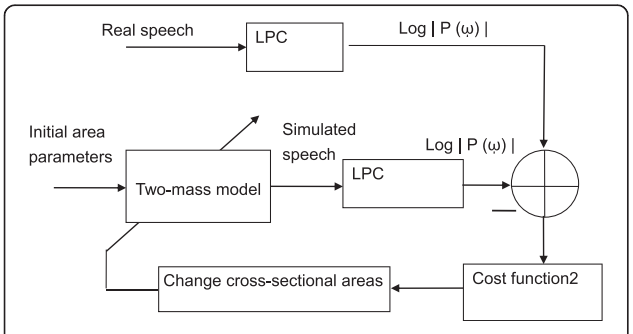

Figure 9 Block diagram showing the detailed structure of our vocal tract fitting method.

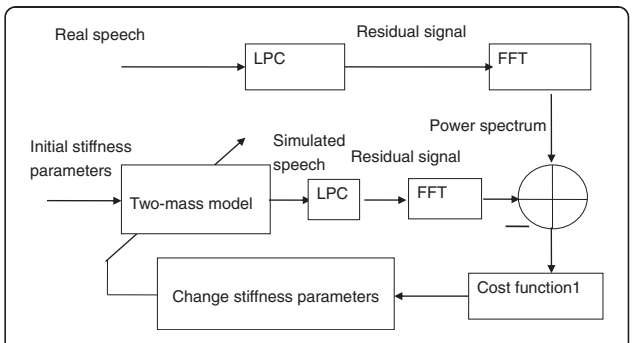

Figure 10 Block diagram showing the detailed structure of our vocal fold fitting method.

4.2.1 Formant ($C^{F_1-F_2}$)

The presence of stress causes an increase in the variability of airflow characteristics due to differences in the muscle tension of the vocal folds. This should cause changes in acoustic interaction around the false vocal folds, thus having an impact on the first and second formants (F_1 and F_2). Thus, F_1 and F_2 are calculated from the spectral envelope to define a cost function:

$$C^{F_1-F_2} = \alpha_1\left(F_1^*-F_1\right)^2 + \alpha_2\left(F_2{}^*-F_2\right)^2,$$
$$\alpha_1 = \frac{1}{\bar{F}_1}, \alpha_2 = \frac{1}{\bar{F}_2}, \tag{11}$$

where the asterisk denotes the target value for real speech. The weights are given the values α_1 and α_2 to normalize the different target parameters to the same range, and the overbar denotes mean values over the target region.

4.2.2 RMS distance of spectral envelope (C_{rms})

$C_{F1\text{-}F2}$ only focuses on the frequency of the first two formants, which is not accurate enough to describe the spectrum. Thus, we find a set of all-pole model coefficients, the cost function of which can be defined as the root mean square (RMS) distance between the spectral envelope of simulated speech and the original speech:

$$C_{\text{rms}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|\log P(\omega_i) - \log P^*(\omega_i)\right|^2}$$
$$P(\omega) = \frac{1}{|A(\omega)|^2} = \frac{1}{\left|\sum_{k=0}^{P} a_k e^{-j\omega k}\right|^2}. \tag{12}$$

4.2.3 Itakura-Saito distance of spectral envelope ($C_{I\text{-}S}$)

The Itakura-Saito distance is a measure of the perceptual difference between an original spectrum and an approximation of that spectrum. It was proposed by Fumitada Itakura and Shuzo Saito in the 1970s and can be described as:

$$C_{\text{I-S}} = \frac{1}{N}\sum_{i=1}^{N}\frac{P(\omega_i)}{P^*(\omega_i)} - \log\frac{P(\omega_i)}{P^*(\omega_i)} - 1. \tag{13}$$

4.2.4 Envelope and formant ($C_{E\text{-}F}$)

The cost functions C_{rms} and $C_{\text{I-S}}$ catch the difference between the rough shapes of the spectral envelopes, but they neglect local information when locating the formant. Since only the first two formants are affected by the oscillation of the vocal folds, the characteristics of F_1 and F_2 should be the chief focus. We propose matching the spectral envelope initially in the first iteration, and then, in the next iteration, the characteristics of the formant are fully considered:

$$C_{E\text{-}F}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\left|\log P(\omega_i) - \log P^*(\omega_i)\right|^2 \qquad n = 1,$$
$$C_{E\text{-}F}^{(n)} = \alpha_1\left(F_1^*-F_1\right)^2 + \alpha_2\left(F_2^*-F_2\right)^2 + w_1\left(H_1^*-H_1\right)^2 + w_2\left(H_2^*-H_2\right)^2 \qquad n\geq 2, \tag{14}$$

where F_1, F_2, H_1, and H_2 refer to the frequency and amplitude of the first and second formants and n is the iteration number.

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 for human beings. In this paper, we calculate the error in acoustic features between real and simulated speech to describe the accuracy of the fitting method.

Using the fitting method described above, the optimal simulated speech corresponding to the inputted real speech can be obtained. Some acoustic features like F_0, F_1, F_2, F_3, and F_4 can also be estimated from the simulated speech. In order to describe the accuracy of the fitting method, we calculate the error in F_0, F_1, F_2, F_3, and F_4 between real and simulated speech. Here, cost function $C_{\text{E-F}}$ is used.

$$\begin{aligned}\text{Err}_{F_0} &= \left(F_0-F_0^*\right)\\ \text{Err}_{F_1} &= \left(F_1-F_1^*\right)\\ \text{Err}_{F_2} &= \left(F_2-F_2^*\right)\\ \text{Err}_{F_3} &= \left(F_3-F_3^*\right)\\ \text{Err}_{F_4} &= \left(F_4-F_4^*\right),\end{aligned} \tag{15}$$

where the asterisk denotes the target value for real speech.

We calculate the errors from simulated and real speech for all the samples for vowels /a/ and /e/ and show the distributions of the errors as shown in Figure 11. Simulated results using these four cost functions are shown in Figure 12. The errors, as shown in Figure 11, are smaller in F_0, F_1, and F_2 (±3 Hz) to obtain higher accuracy. However, the errors in F_3 and F_4 may be increasing, because the cost function chosen places more emphasis on the first and second formants, which are believed to be more important for stress classification. Thus, based on the distributions of errors, it is shown that the proposed method provides reliable accuracy for the fitting to real speech.

5. Classification

Evaluation of the physical parameters is speaker dependent. The structure of the classification method is shown in Figure 13.

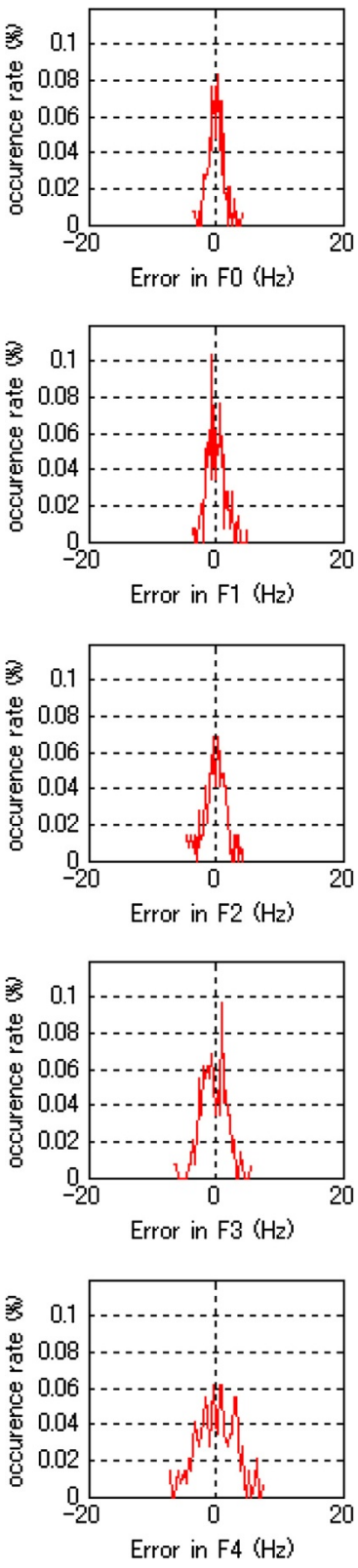

Figure 11 Error distributions of F_0, F_1, F_2, F_3, and F_4 between real and simulated speech. The cost function used is $C_{E\text{-}F}$.

During the training process, all of the speech samples from a specific speaker are labeled as neutral or stressed speech. The labeled speech is segmented into fixed frames, and all of the frames are fit using the two-mass model to estimate the proposed parameters. A linear classifier based on minimum Euclidean distance is trained for the classification, using the estimated physical parameters from all of the frames.

During testing, test speech is input into the system and split into frames, and the trained linear classifier then separates them into neutral or stressed speech. We use Euclidean distance to make a final decision for speech data with several frames. For a test sample with K frames, the feature vector of the ith frame is V_i. We calculate its Euclidean distance $d_i(V_i, a_N)$ $d_i(V_i, a_S)$ to the neutral and stressed classes, respectively, where a_N and a_S are the average vectors of classes for neutral and stressed speech. The final decision is made for the test sample using the following equation:

$$j = \arg\max\left(\sum_{i=1}^{K} d_1(V_i, a_N), \sum_{i=1}^{K} d_i(V_i, a_S)\right) \quad j = N \text{ or } S. \tag{16}$$

A K-fold cross-validation method was used in the training and testing process, and K was set to 4. Using this method, the data set was divided into four subsets, and for each classification, one of the subsets was used as a test set and the other three subsets were combined to form a training set. The final result was obtained by calculating the average classification rate across four trials.

6. Evaluation

6.1 Database and experimental setup

In the experiments, we used a database collected by the Fujitsu Corporation containing speech samples from eleven subjects (four males and seven females) [24]. To simulate mental pressure resulting in psychological stress, the speakers performed three different tasks 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 was asked to finish the tasks within a limited amount of time, and in the other dialogues, there is relaxed chat without any task.

All of the data comes from telephone calls, so the sampling frequency was 8 kHz. Segments with the vowels /a/ and /e/ were cut from the speech and selected as samples. The experiments were conducted for each speaker, and all of the results were speaker

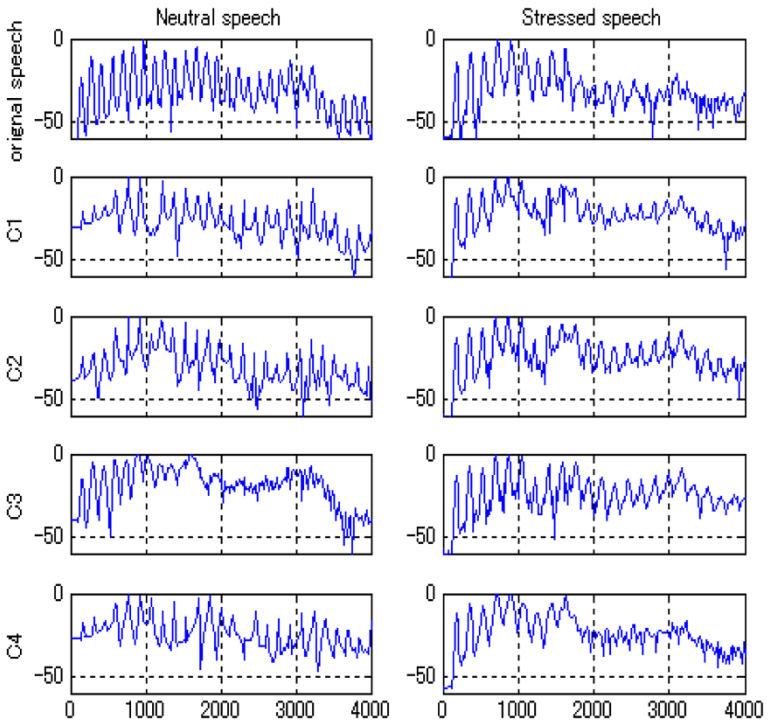

Figure 12 Simulation results of fitting for neutral and stressed speech. Spectrums for original speech (top) and simulated speech with four cost functions ($C^{F_1\text{-}F_2}$, C_{rms}, $C_{\text{I-S}}$, and $C_{\text{E-F}}$) under neutral (left column) and stressed (right column) conditions. In this figure, $C_1 = C^{F_1\text{-}F_2}$, $C_2 = C_{\text{rms}}$, $C_3 = C_{\text{I-S}}$, and $C_4 = C_{\text{E-F}}$.

dependent. The number of samples was different for each speaker. The range of the total number of samples is from 100 to 250 for each vowel from each person. We randomly chose six speakers (three males and three females) from eleven subjects to test classification performance. A K-fold cross-validation method was used in the classification experiments, in which K was set to 4. Using this method, the data set was divided evenly into four subsets, and for each classification, one of the subsets was used as a test set and the other three subsets were combined to form a training set. The final result was obtained by calculating the average classification rate across four trials. The samples were analyzed with 12-order LPC, and the frame size chosen to perform the experiment was 64 ms, with 16 ms for frame shift.

For configuration of the two-mass model, the following values were adopted, using typical values for males: $m_{1\text{M}} = 1.25 \times 10^{-4}$ kg, $m_{2\text{M}} = 2.5 \times 10^{-5}$ kg, $l_{\text{gM}} = 0.014$ m, $d_{1\text{M}} = 0.0025$ m, $d_{2\text{M}} = 5 \times 10^{-4}$ m, $\zeta_{1\text{M}} = 0.1$, $\zeta_{2\text{M}} = 0.6$, $x_0 = 2 \times 10^{-4}$ m, and $P_{\text{s}} = 500$ Pa. The vocal tract model was represented by a tube model, and the number of elements was limited to four cylindrical sections of equal length. Typical values used for configuration for females were as follows: $m_{1\text{F}} = 4.56 \times 10^{-5}$ kg, $m_{2\text{F}} = 9.1 \times 10^{-6}$ kg, $l_{\text{gF}} = 0.01$ m, $d_{1\text{F}} = 1.79 \times 10^{-3}$ m, $d_{2\text{F}} = 3.6 \times 10^{-4}$ m, $\zeta_{1\text{F}} = 0.1$, $\zeta_{2\text{F}} = 0.6$, $x_0 = 2 \times 10^{-4}$ m, and $P_{\text{s}} = 500$ Pa. Furthermore, the ranges for the control parameters were $k_1 = 10$ to 140 kdyn/cm, $k_2 = 2$ to 14 kdyn/cm, $k_{\text{c}} = 4$ to 45 kdyn/cm, VTL = 13 to 19 cm, and A_1, A_2, A_3, $A_4 = 0.2$ to 20 cm.

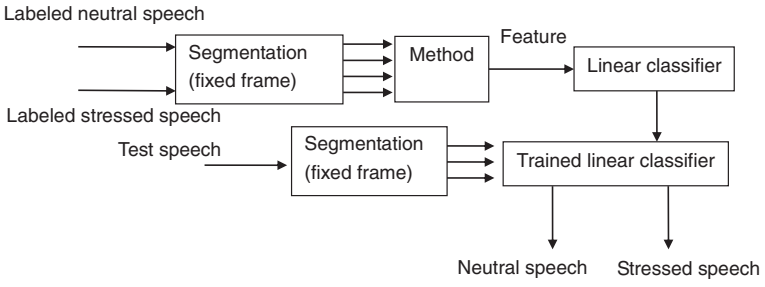

Figure 13 Block diagram of our classification method. A linear classifier is used for the training and testing process.

6.2 Results for cost functions

In the first evaluation, we estimated the vocal tract length of all of the speakers, and two comparisons were made. First, we estimated the cross-sectional area function using the vocal tract fitting method with the four proposed cost functions and then the shape of the vocal tract was fixed at the obtained values (length and area). We used $[k_1, k_c]$ to check classification performance for neutral and stressed speech using only the cost function for the vocal folds in Equation 10. In the second comparison, we estimated stiffness parameters $[k_1, k_c]$ with varied vocal tract, so cost functions both for VF and VT were used to perform the fitting, and iteration was performed. Here, varied VT denotes that the parameters for cross-sectional area are also estimated by fitting the two-mass model instead of being fixed as constants. Finally, the performance of cost functions $C^{F_1\text{-}F_2}$, C_{rms}, $C_{\text{I-S}}$, and $C_{\text{E-F}}$ was evaluated using the classification rate of $[k_1, k_c]$. We used a linear classifier for classification, and the average classification rate for all of the speakers was calculated. The results are shown in Figure 14.

The results illustrate that classification performance is improved when vocal tract values are variable. In this case, the cost functions for the vocal tract are used, and formants are also considered, which results in more information about the frequency domain of the speech being available, making the estimated results more reliable. Furthermore, we compared the performance of different cost functions. Our results show that the stress classification rate for $C_{\text{E-F}}$ is higher than for the other cost functions. Since $C_{\text{E-F}}$ can match the rough shape of the spectral envelope and also effectively catch the characteristics of F_1 and F_2, which have been proven to be sensitive to the interaction between the VF and VT, the classification of stressed speech is improved.

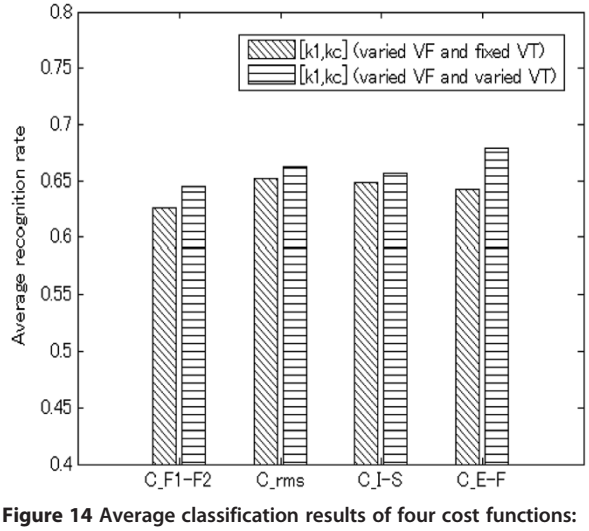

Figure 14 Average classification results of four cost functions: $C^{F_1\text{-}F_2}$, C_{rms}, $C_{\text{I-S}}$, and $C_{\text{E-F}}$. The results for varied VF and fixed VT are the classification rate when the stiffness parameters are estimated with fixed VTL and cross-sectional area. Varied VF and varied VT denote that the parameters for stiffness and cross-sectional area are estimated by fitting the two-mass model to real speech.

6.3 Results for physical parameters

In the second evaluation, VTL was first estimated for each speaker, and further evaluations were based on the obtained vocal tract length. Here, we selected cost function $C_{\text{E-F}}$, which achieved the best performance in classification during the first evaluation. The purpose of this evaluation was to verify which parameters in the stiffness and area functions are related to stress and then check the classification performance of these parameters in comparison to traditionally used features.

6.3.1 Evaluation of vocal tract length estimation

A comparison was first made to evaluate the vocal tract length estimation for each speaker. In this experiment, segments with the vowels /a/ and /e/ were selected as samples. However, the samples for /a/ and /e/ were not mixed together. The two vowels were first used for evaluation separatelyand then the average recognition rate for the two vowels was calculated to show the experimental results. The physical parameters were estimated using the proposed fitting method, and the estimated parameters were used as features to perform the stress classification. The evaluation results for VTL estimation are shown in Figure 15. Features of physical parameters $[k_1, k_c]$ were compared for their classification performance before and after VTL estimation. Our results show that the performance of $[k_1, k_c]$ is improved by the estimation of VTL. Since a speaker's vocal tract

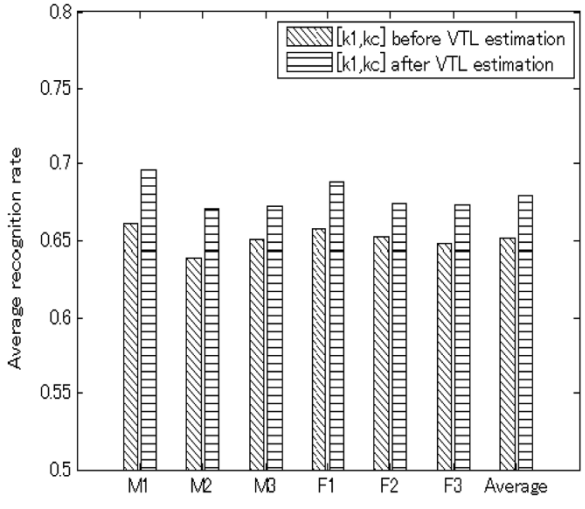

Figure 15 Comparison of performance of physical parameters k_1 and k_c before and after VTL estimation.

length is calculated from the neutral speech of that specific speaker and used as a known value for the estimation of other physical parameters, improvement in classification can be achieved by improving the accuracy of VTL estimation.

6.3.2 Evaluation of stiffness parameters of the vocal folds

In this evaluation, we focused on the stiffness parameters of the vocal folds, and the effect of each stiffness parameter on stress recognition was then examined. The physical parameters k_1, k_2, k_c, A_1, A_2, A_3, and A_4 were estimated from varied VF and varied VT values with estimated VTL, and other physical parameters were fixed at the typical values described in Database and experimental setup. We focused on the evaluation of k_1, k_2, and k_c. The classification performances of $\{[k_1]\}$, $\{[k_1, k_c]\}$, and $\{[k_1, k_2, k_c]\}$ for different speakers are shown in Figure 16. These results that stress classification performance is improved when k_c is considered. k_1 and k_c, therefore, are the parameters which are effective in stress classification. However, average classification accuracy decreases when taking k_2 into account. It suggests that k_2 is not effective in the classification of neutral and stressed speech; therefore, it is sufficient to select k_1 and k_c as feature parameters in further evaluations.

6.3.3 Evaluation of parameters of the cross-sectional areas of the vocal tract

We focused on each parameter of the cross-sectional area individually, and each area's impact on stress recognition was then examined separately. The parameters k_1, k_2, k_c, A_1, A_2, A_3, and A_4 were estimated with varied VF and varied VT values. The parameter sets $\{[k_1, k_c]\}$, $\{[k_1, k_c, A_1]\}$, $\{[k_1, k_c, A_1, A_2]\}$, and $\{[k_1, k_c, A_1, A_2, A_3]\}$ were also evaluated. Their performance is shown in Figure 17. Among the results, we first consider sets $\{[k_1, k_c]\}$ and $\{[k_1, k_c, A_1]\}$. The results show that stiffness $[k_1, k_c]$ is a better parameter for classifying stressed speech. When A_1 is taken into account, classification performance is further improved. This suggests that A_1 is an important parameter strongly related to stress. When A_1 is increasing, it indicates that the area in the supraglottis is broadening. This results in a decrease in the pressure difference inside and outside of the glottis, causing variation in the airflow pattern and further changes in the interaction around the false vocal folds. Considering the performance of sets $\{[k_1, k_c, A_1]\}$, $\{[k_1, k_c, A_1, A_2]\}$, and $\{[k_1, k_c, A_1, A_2, A_3]\}$, we found that they have roughly the same classification accuracy. This illustrates that performance cannot be greatly improved by taking A_2 and A_3 into account and that A_2 and A_3 probably have only a small effect on acoustic interaction. It appears that A_1 is sufficient to classify stressed speech from neutral speech, which agrees with the conclusion of our first evaluation.

A_2 and A_3 do affect F_0 to some extent, which was illustrated in Figure 5, so they have some influence on acoustic interaction and, further, on stress classification; however, we believe their influence is insignificant. The characteristics of the vocal tract also affect stress classification to some extent. Since A_2 and A_3 represent the shape of the vocal tract, $[k_1, k_c, A_1, A_2, A_3]$ can achieve some improvement in the recognition rate, but the increase is very small, which suggests that A_2 and A_3 are less important for stress classification than A_1.

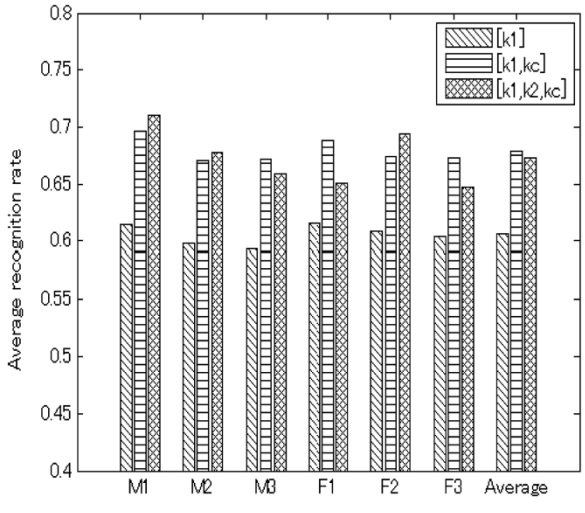

Figure 16 Illustration of classification results for physical parameters of the vocal folds. The performance of stiffness parameters k_1 and k_c shows their effectiveness for stress classification.

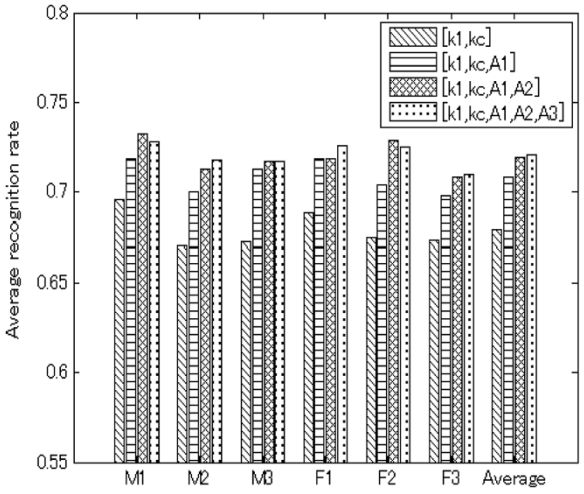

Figure 17 Classification results for physical parameters of the vocal tract. The performance of cross-sectional area parameter A_1 shows its effectiveness for stress classification.

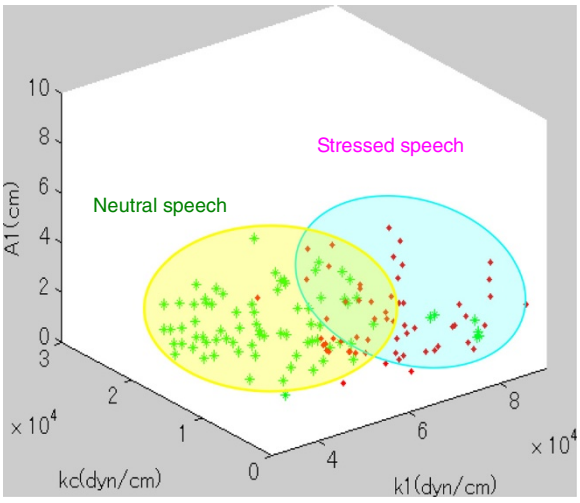

Figure 18 Distributions of estimated parameters k_1, k_c, and A_1 for neutral and stressed speech.

6.3.4 Evaluation for proposed physical parameters

As a result of our evaluation process, parameter set [k_1, k_c, A_1] was proposed. Figure 18 shows the distribution results for k_1, k_c, and A_1 with an estimated VTL. These results show that the proposed parameters are effective for stress classification. The estimated values of the parameters are limited in range, and these ranges correspond to the actual range of human beings. As this distribution shows, stiffness and area of the entrance to the vocal tract are good indicators of stressed speech. Under stressed conditions, the value of k_1 becomes relatively large, k_c smaller, and A_1 increases compared with the same parameters under neutral conditions. This indicates that stress causes variation in the muscle tension of the vocal folds and that the area at the entrance to the vocal tract in the supraglottis becomes wider when the speaker is under stress.

Table 4 Classification rates with different numbers of mixtures

	Number of mixtures					
	1	2	3	4	5	6
Classification rate (%)	61.57	66.63	71.47	71.88	71.22	71.24

We then compared the performance of proposed parameters [k_1, k_c, A_1] with traditionally proposed features, namely [SFM, F_0], [TEO], and [MFCC]. The results are shown in Figure 19. As our experimental results show, [SFM, F_0], which characterizes the vocal folds, works well in classifying stressed speech. This shows that the characteristics of the vocal folds play a very important role in stress classification. MFCC, which represents vocal tract information, is also effective for stress classification, illustrating that the characteristics of the vocal tract also affect stress classification to some extent, which agrees with our previous results in Figure 17. The results shown in Figure 19 demonstrate that our proposed physical parameters outperform the features traditionally used for stress detection, which suggests that parameters estimated from a physical model are more effective at representing stress during phonation than traditional methods. Results show that [k_1, k_c, A_1] has the best stress recognition performance of the physical parameter sets. This illustrates that stiffness of the vocal folds and the cross-sectional area at the entrance to the vocal tract in the supraglottis are the factors which are most impacted when a speaker is under stress.

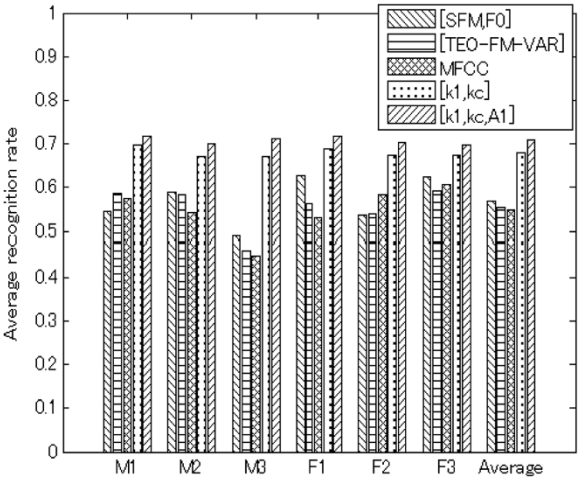

Figure 19 Performance of proposed features compared with traditional methods.

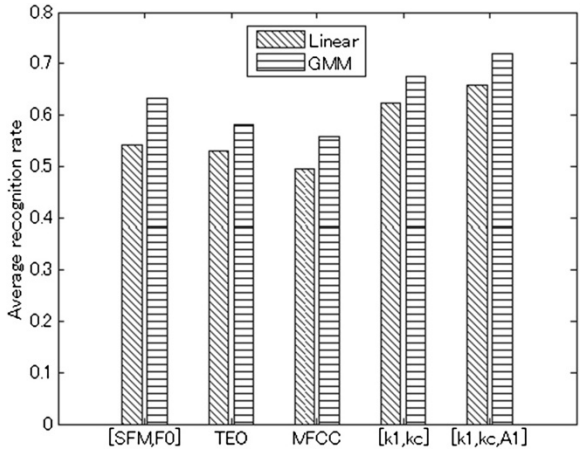

Figure 20 Performance of proposed features using Gaussian mixture models. The number of mixtures is set to 4.

6.4 Results of Gaussian mixture modeling

In this section, we modeled the features using Gaussian mixture model (GMM), which are widely used statistical classifier. Two GMM models were trained, one for neutral speech the other for stressed speech.

The data set for each speaker was divided evenly into four subsets, and for each classification, one of the subsets was used as a test set and the other three subsets were combined to form a training set. The final result was obtained by calculating the average classification rate across four trials by a *K*-fold cross-validation method. In order to increase the amount of training data, the GMMs were trained using training set from three male speakers. The testing set of three male speakers and all of the data from female speakers were combined to generate the testing data used in this experiment.

We performed an experiment to find the best number of mixtures which corresponds to the best performance for proposed features [k_1, k_c, A_1]. Table 4 shows that the best performance is obtained when the number of mixtures equals to four. When we increased the number of mixtures, the classification rate decreased, and it also makes the GMM more complicated. Therefore, the number of mixture components of the GMM was set to four, which obtained the best performance. The features for [SFM, F_0] [TEO-FM-VAR], [MFCC], [k_1, k_c], and [k_1, k_c, A_1] were modeled using GMMs with four mixture components. Classification performance is shown in Figure 20, which shows that improvement is achieved for each feature. However, the increase in classification rates is small because of the lack of training data. If we increase the size of training data significantly, major gains in classification rate should be achieved. Here, it is recommended that a GMM with four mixture components is acceptable for improving stress classification.

7. Conclusion

In this paper, we explored more effective features for the classification of neutral and stressed speech based on a physical model. To achieve this target, a two-mass model characterizing the properties of the vocal folds and the vocal tract was used to simulate speech production. Physical parameters including stiffness of the vocal folds, vocal tract length, and cross-sectional area of the vocal tract were investigated and estimated using a method that fits the two-mass model to real data. Cost functions were used as targets to reach more reliable results. The obtained parameters were used as physical features to classify stressed speech. We concluded that the two parameters: (1) stiffness of the vocal folds and (2) the area at the entrance to the vocal tract in the supraglottis, which is related to the velocity of glottal airflow and acoustic interaction between the vocal folds and the vocal tract, are key indicators of stress during phonation. The average performance in the classification of speech under stress was improved by 10% to 15% using the proposed features, compared to traditional methods of stressed speech classification. In the future, our work should be focused on the exploration of parameters for a speaker-independent stressed speech classification system.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

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 allowing us to use their database and for his valuable suggestions.

Author details

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

Received: 30 October 2012 Accepted: 21 June 2013
Published: 5 July 2013

References

1. HJM Steeneken, JHL Hansen, Speech under stress conditions: overview of the effect on speech production and on system performance, in *Proc. ICASSP* (Atlanta, Georgia, 1996)
2. D Cairns, JHL Hansen, Nonlinear analysis and detection of speech under stressed conditions. J. Acoust. Soc. Am. **96**(6), 3392–3400 (1994)
3. RV Bezooijen, *The characteristics and recognizability of vocal expression of emotions* (Foris, Drodrecht, 1984)
4. FJ Tolkmitt, KR Scherer, Effect of experimentally induced stress on vocal parameters. J. Exp. Psychol. **12**(3), 302–313 (1986)
5. CE Williams, KN Stevens, Emotions and speech: some acoustical correlates. J. Acoust. Soc. Am. **52**(4), 1238–1250 (1972)
6. SE Bou-Ghazale, JHL Hansen, Generating stressed speech from neutral speech using a modified CELP vocoder. Speech Commun. **20**, 93–110 (1996)
7. ZS Bond, TJ Moore, *A note on loud and Lombard speech*. International Conference on Spoken Language Processing (Kobe, 1990), pp. 969–972
8. JHL Hansen, *Analysis and compensation of stressed and noisy speech with application to robust automatic recognition*. Ph.D. dissertation (Georgia Institute of Technology, Atlanta, 1988)
9. IR Murray, C Baber, A South, Toward a definition and working model of stress and its effects on speech. Speech Commun. **20**, 3–12 (1996)
10. J Whitmore, S Fisher, Speech during sustained operations. Speech Commun. **20**, 55–70 (1996)
11. A Kamano, N Washio, S Harada, N Matsuo, *A study of psychological suppression detection based on non-verbal information*. IEICE Technical Report IEICE-SP2010-64 (IEICE, Tokyo, 2010), pp. 107–110. in Japanese
12. JF Kaiser, On Teager's energy algorithm and its generalization to continuous signals, in *Proceedings of the 4th IEEE Digital Signal Processing Workshop* (New Paltz, 1990)
13. G Zhou, JHL Hansen, JF Kaiser, Nonlinear feature based classification of speech under stress. IEEE Trans. Speech Audio Process. **3**, 201–206 (2001)
14. G Fant, *Acoustic Theory of Speech Production* (Mouton, The Hague, 1960)
15. HK Dunn, Methods of measuring vowel formant bandwidths. J. Acoust. Soc. Am. **33**(12), 1737–1746 (1961)
16. DY Wong, JD Markel, AH Gray, Glottal inverse filtering from the acoustic speech waveform. IEEE Trans. Acoust. Speech Signal Process **27**(4), 350–355 (1979)
17. JF Kaiser, Some observations on vocal tract operation from a fluid flow point of view, in *Vocal Fold Physiology: Biomechanics, Acoustics, and Phonatory Control*, ed. by IR Titze, RC Scherer (Denver Center for the Performing Arts, Denver, 1983), pp. 358–386
18. K Ishizaka, JL Flanagan, Synthesis of voiced sounds from a two-mass model of the vocal cords. Bell. Syst. Tech. J. **51**, 1233–1268 (1972)

19. D Kincaid, W Cheney, *Numerical Analysis: Mathematics of Scientific Computing*, 3rd edn. (Brook/Cole, Pacific Grove, 2002), pp. 722–723
20. C Lucero, Chest- and falsetto-like oscillations in a two-mass model of vocal folds. J. Acoust. Soc. Am. **100**, 3355–3399 (1996)
21. IR Titze, Acoustic interpretation of resonant voice. J. Voice **15**, 519–528 (2001)
22. JL Flanagan, *Speech Analysis, Synthesis, and Perception* (Springer-Verlag, New York, 1972)
23. A de Cheveigne, H Kawahara, YIN, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Am. **111**(4), 1917–1930 (2002)
24. IR Titze, BH Story, Acoustic interactions of the voice source with the lower vocal tract. J. Acoust. Soc. Am. **101**, 2234–2243 (1997)

doi:10.1186/1687-4722-2013-17
Cite this article as: Yao *et al.*: Classification of speech under stress based on modeling of the vocal folds and vocal tract. *EURASIP Journal on Audio, Speech, and Music Processing* 2013 **2013**:17.