Acoust. Sci. & Tech. **34**, 5 (2013)

PAPER

Classification of speech under stress by physical modeling

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

[1]*Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464–8603 Japan*
[2]*Department of Media Informatics, Aichi University of Technology, 50–2 Umanori, Nishisako-cho, Gamagori, 443–0047 Japan*

(*Received 22 November 2012, Accepted for publication 28 January 2013*)

Abstract: In this study, we propose a method of classifying speech under stress using parameters extracted from a physical model to characterize the behavior of the vocal folds. Although many conventional methods have been proposed, feature parameters are directly extracted from waveforms or spectrums of input speech. Parameters derived from the physical model can characterize stressed speech more precisely because they represent physical characteristics of the vocal folds. Therefore, we propose a method that fits a two-mass model to real speech in order to estimate the physical parameters that represent muscle tension in the vocal folds, vocal fold viscosity loss, and subglottal pressure coming from the lungs. Furthermore, combinations of these physical parameters are proposed as features effective for the classification of speech as either neutral or stressed. Experimental results show that our proposed features achieved better classification performance than conventional methods.

Keywords: Speech under stress, Stress classification, Physical parameters, Two-mass model

PACS number: 43.72Ar [doi:10.1250/ast.34.311]

1. INTRODUCTION

The effects of stress on speech signals have been the topic of numerous studies. Many factors, such as emotional state, fatigue, physical environment, and workload, can cause people to experience stress. It has become increasingly important to study speech under stress in order to improve the performance of speech recognition systems, to recognize when people are in a stressed state, and to understand the context in which a speaker is communicating.

Researchers have attempted to probe reliable indicators of stress by analyzing acoustic variables. Some external factors (workload, background noise, etc.) and internal factors (emotional state, fatigue, etc.) may induce stress [1]. 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 [2,3]. It has been found that fundamental frequency (F_0) has different characteristics for each emotion [4], and that respiration patterns and muscle tension also change [5]. The influence of the Lombard effect on speech recognition has been examined by Bond and Moore [6] and Hansen [7], who analyzed selected acoustic features, such as amplitude and distribution of spectral energy, and 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 [8,9]. In 2011, Matsuo and Kamano *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 performance than traditional methods [10]. Furthermore, the Teager energy operator (TEO) [11] has been investigated for the purpose of stress classification. As a result, methods based on the Teager energy operator have been proposed to explore variations in the energy of airflow characteristics within the glottis [12].

We propose a new classification method, based on the working mechanisms of the vocal folds, for speech under stress using parameters estimated from a physical model. It is believed that the presence of stress can result in variations in the physical characteristics of physiological systems. The parameters of a physical model can represent the influence of speaking style more directly. Therefore a

physical model is helpful in estimating the parameters of the physiological system.

In this paper, we concentrate on a method for fitting the two-mass model to real speech in order to estimate the physical parameters that characterize the vocal folds. An algorithm based on the analysis-by-synthesis method (A-b-S) is proposed for fitting the model to real speech. The Nelder–Mead simplex method is used to estimate the optimal physical parameters, and different cost functions are proposed to compare performance in fitting and classification. As a result, the parameters of the two-mass model, representing muscle tension in the vocal folds, vocal fold viscosity loss, and subglottal pressure, are estimated as features used in the classification of neutral and stressed speech. In this work, we assume that the presence of stress has a greater impact on the vocal folds, and the parameters of the vocal folds are the most important for the classification. Furthermore, it is difficult to estimate the parameters of the vocal folds and the vocal tract at the same time. Therefore, the objective in our work is to fit the major variation in the vocal folds, and the parameters of vocal folds are chiefly considered to show their effectiveness in the classification of stressed speech.

The paper is organized as follows. In Sect. 2, physical parameters based on a two-mass model are described as features for classification. This is followed in Sect. 3 by the presentation of a fitting algorithm for real speech data to help estimate the physical parameters. In Sect. 4, experiments are performed to evaluate the obtained parameters and show their corresponding classification performance for neutral and stressed speech. Finally, we draw our conclusions in Sect. 5.

2. PHYSICAL PARAMETERS

A method for classifying speech under stress is proposed, in which a two-mass model is fitted to real speech. Some of the physical parameters that characterize the vocal folds are estimated. The two-mass vocal fold model was proposed by Ishizaka and Flanagan to simulate the process of speech production [13]. The physical parameters proposed as features for classification in the two-mass model are stiffness, damping ratio, and subglottal pressure.

2.1. Stiffness

The stiffness parameters, which represent muscle tension in the vocal folds, are the main factors related to fundamental frequency. The amplitudes of the glottal area and glottal volume velocity decrease gradually with increasing stiffness [14] because variation in the stiffness of the vocal folds influences the time span of the glottal opening and closing phases. During this time span, subglottal airflow is accelerated in the glottis, thus impacting the velocity of glottal airflow as well as the glottal source. Therefore, it is our assumption that stiffness parameters, which reflect the tension of the muscles, can be a potential factor in stress detection. In the production of speech, however, the natural frequency of the vocal folds is determined by both their mass and stiffness. So in order to simplify the estimation algorithm, the stiffness parameters are only estimated with mass fixed as a constant.

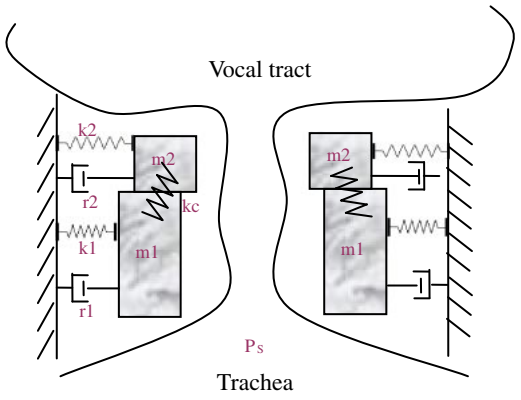


Fig. 1 The two-mass approximation of the vocal folds.

Figure 1 shows a sketch of the model. Each vocal fold is represented by a mass-spring-damper system [13], joined with a coupling stiffness, and is represented as

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

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

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. r_i denotes the equivalent viscous resistances, and s_i refers to the force related to tissue elasticity. F_i is the force of airflow, which is determined by subglottal pressure.

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

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

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

Generally, the stiffness of the vocal folds depends mainly on two muscles: the cricothyroid muscle and the thyroarytenoid muscle. CT and TA represent the weighted activities of the two muscles. In the two-mass model, coupling stiffness k_c is relative to the tension in the thyroarytenoid muscle (TA), so a high k_1 value and low k_c value represent the contraction of CT and relaxation of TA [14].

2.2. Viscosity

The viscosity of vocal fold tissue has been shown to be essential in vocal fold oscillation. During phonation, the viscosity of vocal fold tissue changes owing to hydration effects [15]. The damping ratio of viscosity has been estimated by Kaneko, *et al.* [16] and Isshiki [17]. Results show that damping ratio has a close correlation with fundamental frequency, which is a stress indicator [18]. Therefore, in this work, we assume that the damping ratio is a parameter that varies in a narrow range during phonation under different conditions. Since the viscosity of the vocal folds depends mainly on the bulk of the vocal cords (m_1 of our model), the damping ratio for m_1 is considered to be an influential parameter.

The viscous resistance of the vocal folds represents the stickiness of the soft, moist surfaces during contraction of the vocal fold. This can be represented as

$$r_1 = 2\zeta_1\sqrt{m_1 k_1} \quad r_2 = 2\zeta_2\sqrt{m_2 k_2}, \tag{4}$$

where ζ_i is a damping ratio, and k_i denotes the linear stiffness of the spring s_i.

2.3. Subglottal Pressure

Subglottal pressure is the pressure of the airflow in the trachea below the glottis. This is the main factor used by speakers to control phonation when producing speech. Subglottal pressure affects the amplitude of speech signals and fundamental frequency. Higher subglottal pressure causes higher airflow velocity, thus, it has an impact on glottal flow. It can therefore be considered as one of the feature parameters for classifying stressed speech.

Aerodynamics in the glottis is modeled with a set of equations proposed by Ishizaka and Flanagan [13]. If the subglottal pressure is represented as P_s, 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 cross-sectional area at the inlet to the glottis causes a phenomenon called vena contracta, which makes the air pressure undergo a greater drop. This 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 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. 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 input area to the vocal tract. N denotes the difference in area between the outlet of the vocal folds and inlet of the vocal tract, and is significant in the acoustic interaction between the glottal source and the vocal tract.

Finally, force F_i acting on the masses is calculated by $F_i = (P_{i1} + P_{i2})/2$. When the glottis is closed, forces are calculated by

$$F_1 = d_1 l_g P_s \quad x_1 \le -x_0 \quad or \quad x_2 \le -x_0$$
$$F_2 = \begin{cases} d_2 l_g P_s, & if \quad x_1 > -x_0, \quad x_2 \le -x_0 \\ 0, & if \quad x_1 \le -x_0 \end{cases}. \tag{9}$$

The two-mass model can be represented as a vocal fold model connected to a four-tube model. The four-tube model is constructed using a transmission line analogy involving four cylindrical, hard-walled sections terminating in the radiation load of a circular piston in an infinite baffle. The element values are determined from cross-sectional areas A and cylinder lengths L.

In this study, we consider the fitting of two-mass model to vowels because only the voiced sound can cause vibration of the vocal folds, so all of the segments for vowel /a/ are chosen as training data and testing data, and the evaluation is performed for each speaker. Since all the training and testing data are for /a/, the variation in the shape of the vocal tract is relatively minor across speakers. Our aim in this work is stress classification, therefore, an assumption is made that the effect of the vocal tract is smaller than that of the vocal folds and thus the parameters in the tube model are fixed as constants for vowel /a/.

Moreover, the objective is stress classification and our main consideration in this work is the characteristics of the vocal folds under the stressed condition. The parameters of the vocal folds are more essential and effective for stress classification because the vocal folds are mainly affected when stress occurs [12]. Therefore, in this work,

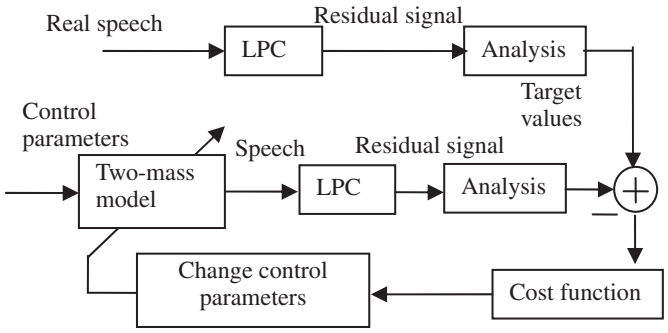


Fig. 2 Structure of algorithm.

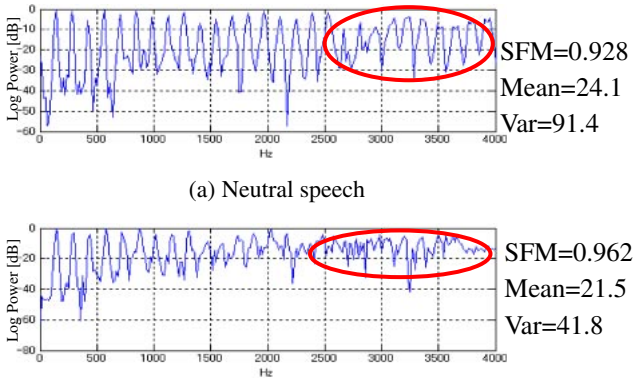


Fig. 3 Spectrum of residual signals for a male speaker.

we first concentrate on the parameters of the vocal folds, and the vocal tract parameters will be considered in the future.

Therefore, stiffnesses k_1, k_2, k_c, damping ratio ζ_1, and subglottal pressure P_s are selected as control parameters, which represent the parameters to be estimated, to generate the features for stress classification. After defining a target cost function, we can estimate the physical parameters by fitting the two-mass model to real speech.

3. ESTIMATION METHOD

3.1. Algorithm for Fitting

Figure 2 shows the structure of the fitting algorithm. Fitting the two-mass model to real data involves two steps. First, a pre-emphasis filter is used to flatten the speech spectrum before spectral analysis. The aim is to compensate the high-frequency part of the speech signal that was suppressed during the human sound production mechanism. The pre-emphasis filter used here is $H(z) = 1 - \alpha z^{-1}$, where $\alpha = 0.97$. Since we mainly focus on the modulation effect at the glottal source of speech, input speech is then analyzed using linear predictive coding (LPC), which removes the influence of formants and lip radiation, and emphasizes the glottal source, to obtain the residual signal. Then, some target values can be determined to measure the spectrum of the residual signal.

In the second step, each set of control parameters is considered separately. After that, simulation can be performed using the two-mass model to generate speech using the given control parameters. In order to make a comparison with the spectrum of the residual signal from the real speech, LPC analysis is also performed for the simulated speech to obtain the residual signal, and the same target values are calculated. By inverse filtering of LPC, the parameters of the vocal folds can be estimated correctly. Next, the target values are compared with the ones obtained in the first step in order to observe the difference between them. The difference between the simulated target values and the measured target values from real speech can be represented by a cost function. The control parameters are then varied and the speech is simulated until the cost function reaches a minimum.

The Nelder-Mead algorithm [19] is a simplex method of finding the minimum of a function involving several variables. It is a direct search method and it 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 k_1, k_2, k_c, ζ_1, and P_s as variables. The calculation of each time 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 for the control parameters. Optimal values of the physical parameters are estimated by the Nelder-Mead simplex method, which is implemented to search for the optimal physical parameters to minimize the cost function.

3.2. Cost Function

In this paper, we utilize four different cost functions in order to compare their performance in classification.

3.2.1. Fundamental frequency and spectral flatness measure (F_0-SFM)

When stress occurs, the fundamental frequency and spectrum of the glottal source are affected. The harmonic structure of the spectrum loses clarity in the high-frequency band, and the spectrum becomes smooth and irregular. The spectrums of residual signals are shown in Fig. 3. The part of high frequency in the spectrum is marked by red circles. This irregularity can be quantified with a "spectral flatness measure" (SFM). The spectral flatness is calculated by dividing the geometric mean by the arithmetic mean of the power spectrum:

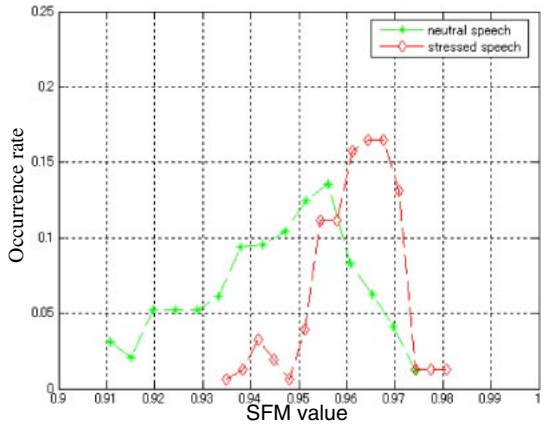


Fig. 4 Distribution of SFM for spectrum of residual signals.

$$SFM = \frac{\sqrt[M]{\prod_{n=0}^{M-1} S(n)}}{\frac{1}{M}\sum_{n=0}^{M-1} S(n)}, \tag{10}$$

where $S(n)$ is the magnitude of bin number n. The distributions of SFM for neutral and stressed speech for a male speaker are shown in Fig. 4.

The cost function can be defined as a weighted sum of the squared difference between target values from the simulated speech and from the real speech, and can be represented as:

$$\begin{aligned} C1 &= \alpha_1(F_0^* - F_0)^2 + \alpha_2(SFM^* - SFM)^2, \\ \alpha_1 &= 1/\overline{F_0},\ \alpha_2 = 1/\overline{SFM}, \end{aligned} \tag{11}$$

where the asterisk denotes the target value from real speech. The target values here denote the values of F_0 and SFM. The weights are given the values α_1, α_2 to normalize the different target values to the same range, and the overbar denotes mean values over the target region. The frequency band of the spectrum was limited to 3,000–4,000 Hz for calculating the spectral flatness measure.

3.2.2. F_0 and statistical information (F_0-Stat)

The high frequency bands of the spectrum become disordered when stress occurs. Because of the lack of clear harmonic structure, it is difficult to represent the spectrum using only fundamental frequency. Therefore, the mean and variance of the spectrum are used to describe the irregularity in the high frequency band. Figure 5 shows the distribution of mean and variance for a male speaker, sample 180. As can be seen when stress occurs, values for mean and variance fall (mean = 21.5, and variance = 41.8 in Fig. 3(b)). The cost function is defined as

$$\begin{aligned} C2 = {} & \beta_1(F_0^* - F_0)^2 + \beta_2|\text{mean}(S^*(n)) - \text{mean}(S(n))|^2 \\ & + \beta_3|\text{var}(S^*(n)) - \text{var}(S(n))|^2, \end{aligned} \tag{12}$$

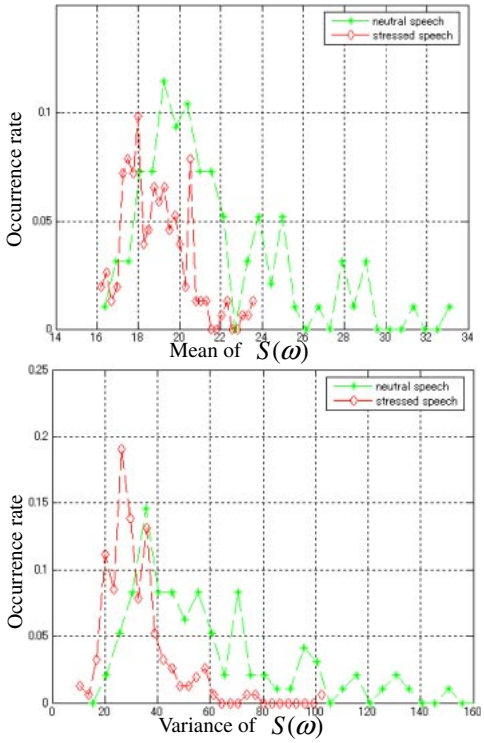


Fig. 5 Distributions of mean and variance of spectrum of residual signal for neutral (green) and stressed speech (red).

where $\beta_1 = 1/\overline{F_0}$, $\beta_2 = 1/\overline{\text{mean}(S(n))}$ and $\beta_3 = 1/\overline{\text{var}(S(n))}$ are used to normalize target values to the same range. The overbar denotes mean values over the target region. The frequency band of the spectrum was limited to 3,000–4,000 Hz.

3.2.3. Spectrum and histogram (Spect-Histo)

A histogram can be used to calculate statistical characteristics, including mean, variance, entropy, and third-order moments. It more accurately represents the spectrum of the glottal source. A frequency histogram refers to the probability mass function of the magnitude of the spectrum. More formally, the frequency histogram is defined by

$$H(k) = M \cdot B(X = k), \tag{13}$$

where X represents the magnitude of the spectrum, M is the number of frequency bins in the spectrum, and B denotes the probability of $X = k$. Thus a concatenated cost function can be defined as the spectral distance in the low-frequency band and the histogram distance in the high-frequency band, which can be represented as

$$\begin{aligned} C3 &= W_1 \sum_{n=1}^{M}(S^*(n) - S(n))^2 + W_2 \sum_{j=1}^{L}(H^*(k_j) - H(k_j))^2 \\ W_1 &= 1 \Big/ \sum_{n=1}^{M}(S(n))^2,\ W_2 = 1 \Big/ \sum_{j=1}^{L}(H(k_j))^2, \end{aligned} \tag{14}$$

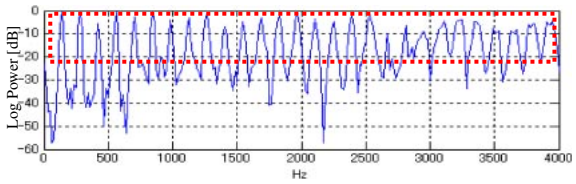


Fig. 6 Cut-off spectrum with a threshold. Spectrum within the dotted line is emphasized for calculation of cost function.

where $S(n)$ and $S^*(n)$ represent the spectrums of simulated speech and real speech, respectively. Note that M and L are the number of bins for the spectrum and the histogram. A partition of the speech frequency band for F_0-Stat was performed to determine the high-frequency band between 3,000–4,000 Hz; however, this partition is coarse. Automatic separation of the low- and high-frequency bands might help us derive a more effective cost function for fitting. This separation is performed by detecting the periodic feature of the harmonic, described as follows

Step 1: Spectrum is split into (overlapping) frames. Frame length is fixed as the frequency band including three harmonic structures.

Step 2: Autocorrelation is calculated for each frame.

Step 3: Zero-crossing for the autocorrelation is computed to classify whether it has a clear harmonic structure in this frame.

Step 4: Separation point is determined by an abrupt increase in zero-crossing.

3.2.4. Modified spectrum (Spectrum)

The spectrum of the residual signals has a flat upper envelope, and information on harmonic structure mainly exists in spectral peaks rather than in spectral valleys. Therefore, the spectrum is cut with a threshold to remove the lower valley section, and only the upper section representing harmonic structure is used to calculate spectral distance, as shown in Fig. 6, which is a power spectrum of speech from a male speaker, with the threshold chosen as -20 dB. Spectral distance can then be calculated to evaluate the similarity between the spectrums of real and simulated speech.

Let $P(n)$ and $P^*(n)$ represent the cut-off spectrum of simulated speech and real speech, respectively. The normalized cost function can be defined as

$$C4 = \frac{\sum_{n=1}^{M} |P^*(n) - P(n)|^2}{\sum_{i=1}^{M} |P(n)|^2}, \tag{15}$$

where M is the number of bins for the power spectrum.

Figure 7 shows the simulated results with these four cost functions. In this experiment, the neutral and stressed speech in Fig. 3 from a male speaker are used to estimate

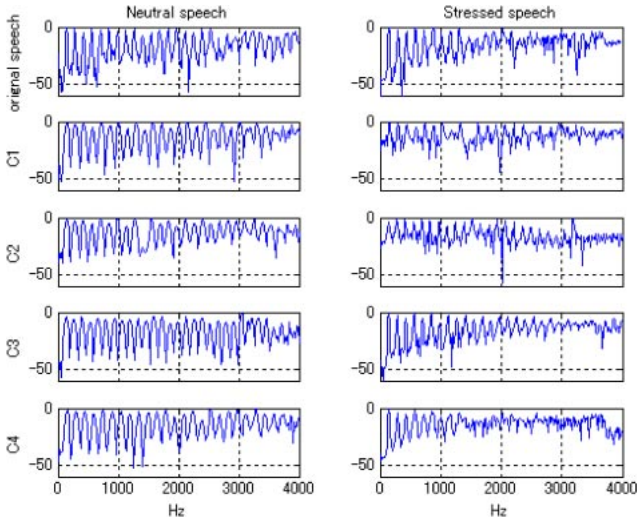


Fig. 7 Spectrums of residual signals for original speech (top) and for simulated speech with different cost functions under neutral (left column) and stressed (right column) conditions.

the corresponding physical parameters by fitting the two-mass model. The simulated spectrums of residual signals obtained using the estimated parameters are shown. The estimated values are shown in Table 1.

4. EVALUATION

4.1. Database and Experimental Setup

In the experiments, we used a database collected by the Fujitsu Corporation containing speech samples from eleven subjects (four male and seven female) [10]. To simulate mental pressure resulting in psychological stress, we introduced three different tasks, which were performed by the speakers while conversing on the telephone 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 were 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 was relaxed chat without any task.

All of the data is acquired from telephone calls, so the sampling frequency was 8 kHz. The segments with the vowel /a/ were cut from the speech and selected as training samples and testing samples. The experiments were performed for each speaker. The number of samples was different for each speaker, and the total number of samples ranged about 100–250 for each person. We randomly chose six speakers (three male, three female) from eleven subjects to show the classification performance. Linear classifiers based on the minimum Euclidean distance to reach the classification performance were used.

Table 1 Estimated values of physical parameters for four cost functions.

	Neutral speech					Stressed speech				
	P_S [Pa]	k_1 [dyn/cm]	k_2 [dyn/cm]	k_c [dyn/cm]	ζ_1	P_S [Pa]	k_1 [dyn/cm]	k_2 [dyn/cm]	k_c [dyn/cm]	ζ_1
$C1$	438	75,460	7,840	22,640	0.16	299	90,780	8,040	8,260	0.32
$C2$	455	74,030	8,250	21,980	0.14	276	87,440	8,277	7,260	0.32
$C3$	416	74,270	7,730	20,810	0.17	306	84,290	7,800	7,740	0.31
$C4$	446	77,360	8,000	22,600	0.14	279	89,170	8,480	7,650	0.34

A K-fold cross-validation method was used in classification experiments, in which K was set to 4. By this method, the data set was divided into 4 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 4 trials. The samples were analyzed with 12-order LPC and the frame size chosen to perform the experiment was 64 ms, with 16 ms frame shift.

For the configuration of the two-mass model, the following values were adopted, using typical values for males: $m_{1M} = 1.25 \times 10^{-4}$ kg, $m_{2M} = 2.5 \times 10^{-5}$ kg, $l_{gM} = 0.014$ m, $d_{1M} = 0.0025$ m, $d_{2M} = 5 \times 10^{-4}$ m, $\zeta_{2M} = 0.6$, and $x_0 = 2 \times 10^{-4}$ m. The vocal tract model was represented by a standard tube configuration for the vowel /a/ [20], and the number of elements was limited to four cylindrical sections of equal length. In order to reduce the number of parameters to be estimated, and simplify the proposed method, the typical values are adopted for the configuration of the tube model. For males, the length of the vocal tract was assumed to be $L_M = 0.18$ m, with each element set to $l_i = 0.045$ m, and the cross-sectional areas were $A_1 = 8 \times 10^{-5}$ m^2, $A_2 = 4 \times 10^{-5}$ m^2, $A_3 = 3 \times 10^{-4}$ m^2, and $A_4 = 8 \times 10^{-4}$ m^2. For the configuration for females, the typical values were as follows: $m_{1F} = 4.56 \times 10^{-5}$ kg, $m_{2F} = 9.1 \times 10^{-6}$ kg, $l_{gF} = 0.01$ m, $d_{1F} = 1.79 \times 10^{-3}$ m, $d_{2F} = 3.6 \times 10^{-4}$ m, $\zeta_{2F} = 0.6$, $x_0 = 2 \times 10^{-4}$ m. For the vocal tract model, the length of the vocal tract was set to $L_F = 0.14$ m, with each element $l_i = 0.035$ m, and the cross-sectional areas were $A_1 = 4.85 \times 10^{-5}$ m^2, $A_2 = 2.4 \times 10^{-5}$ m^2, $A_3 = 1.8 \times 10^{-4}$ m^2, and $A_4 = 4.85 \times 10^{-4}$ m^2.

In the experiments, the following ranges for the control parameters were used for all speakers: P_S: 200–1,900 Pa, k_1: 10,000–140,000 dyn/cm, k_2: 2,000–14,000 dyn/cm, k_c: 4,000–45,000 dyn/cm, ζ_1: 0.05–0.6. The ranges here selected for the control parameters are sufficiently large to ensure that our search method is able to simulate different types of speech. Moreover, they can make our work applicable to emotional speech recognition. Emotional speech has larger ranges for physical parameters (e.g., the standard value of subglottal pressure for phonation is 2–8 cmH_2O, but 10–12 cmH_2O for loud speech), so the greater search range is advisable for the search method.

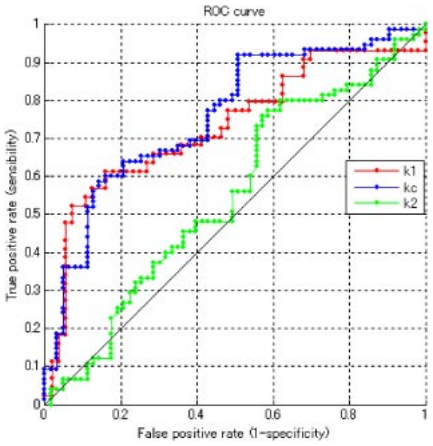


Fig. 8 ROC curve for stiffness parameters (k_1, k_2, k_c).

4.2. Comparison of Feature Parameters

By fitting the model to real data, the physical parameters of speech can be estimated. The obtained parameters were used as features for classifying speech into neutral or stressed speech. The purpose of our first evaluation was to verify which parameters are related to stress, and whether these parameters are dependent on speakers. The proposed parameter sets were then compared to show their classification performance using $C4$ as the cost function.

In the first evaluation, the stiffness parameters were first focused on and the effect of each stiffness on stress recognition was then examined. The parameters k_1, k_2, and k_c were estimated with $P_S = 500$ Pa, and $\zeta_1 = 0.1$, and other physical parameters were fixed at the typical values described in Sect. 4.1. In Fig. 8, receiver operating characteristics (ROC curves) are shown to compare the classification performances of k_1, k_2 and k_c separately for a male speaker. In this result, k_1 and k_c perform better than k_2 in classifying stressed speech from neutral speech. The classification performances of $\{[k_1]\}$, $\{[k_1, k_2]\}$ and $\{[k_1, k_2, k_c]\}$ for different speakers are shown in Fig. 9. It is illustrated that the average classification accuracy decreas-

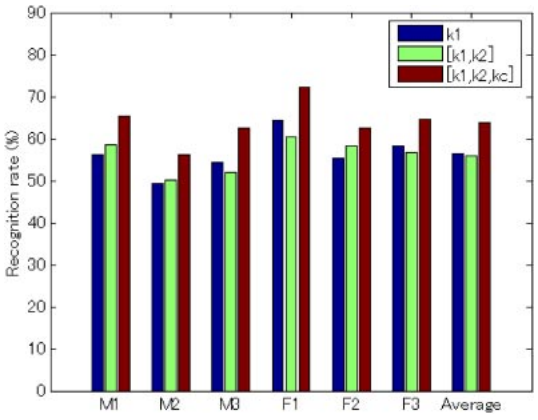


Fig. 9 Classification performance for stiffness parameters.

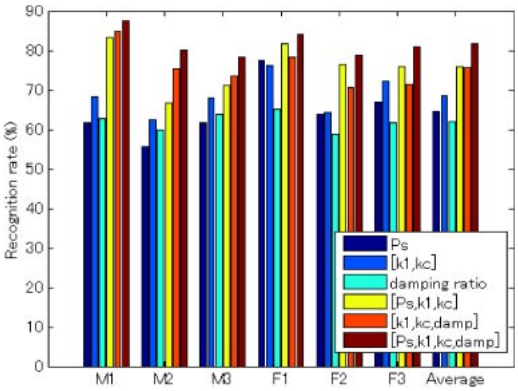
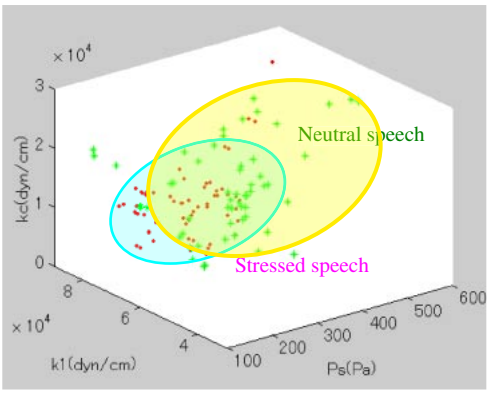
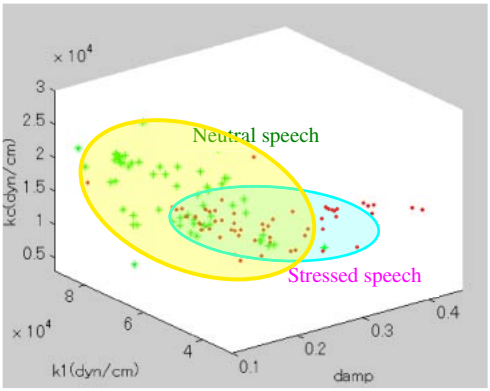


Fig. 10 Performance of each parameter and parameter set.



(a) Distribution for P_S, k_1 and k_c.



(b) Distribution for k_1, k_c and damping ratio

Fig. 11 Distributions of estimated parameters.

es when taking k_2 into account, and the performance for stress classification is improved when k_c is considered. It is proved 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 evaluation.

Next, we focused on subglottal pressure, stiffness, and damping ratio individually, and fixed the other parameters at typical values. Then the effect of each parameter on stress recognition was examined. The results are shown in Fig. 10 (in Figs. 10–16, "damp" is the abbreviation for "damping ratio"). For these physical parameters, the results show that stiffness (k_1, k_c) achieves the best classification performance, which means it is strongly linked to stress. The other two parameters vary in performance depending on the sex of the speaker. For males, the results show that the damping ratio can classify stressed speech well, so it plays a more important role when male speakers are under stress, whereas for females, the stress classification rate of P_S is higher, which indicates that subglottal pressure is a better indicator of stress. Furthermore, classification performance among speakers differs significantly, which proves that these physical parameters are dependent on the speakers.

F_0 is dependent on stiffness and subglottal pressure, while the viscosity of vocal folds is determined by stiffness and damping ratio. Therefore, the following parameter sets are proposed: $\{[P_S, k_1, k_c]\}$, $\{[k_1, k_c, \zeta_1]\}$, and $\{[P_S, k_1, k_c, \zeta_1]\}$. Figure 11(a) shows the distribution results for $\{[P_S, k_1, k_c]\}$, in which we estimated P_S, k_1, and k_c with a fixed damping ratio, while Fig. 11(b) shows the distribution for $\{[k_1, k_c, \zeta_1]\}$, with subglottal pressure fixed at a typical value. It is illustrated that the proposed parameters are effective for the stress classification and the estimated values of parameters are limited in some range, and these ranges agree with the actual situation for human beings.

As this distribution in Fig. 11 shows, stiffness, subglottal pressure, and damping ratio are all good indicators of stressed speech. Under stressed conditions, the value of P_S becomes smaller, k_1 becomes relatively large, k_c smaller, and the damping ratio increases compared with the same parameters under neutral conditions. This indicates that high stress causes variation in the muscle tension of the vocal folds. There is also lower subglottal pressure from the lungs and the vocal folds become more viscous than under neutral conditions.

We checked the performances of the above parameters and compare them. In the proposed sets, the stress

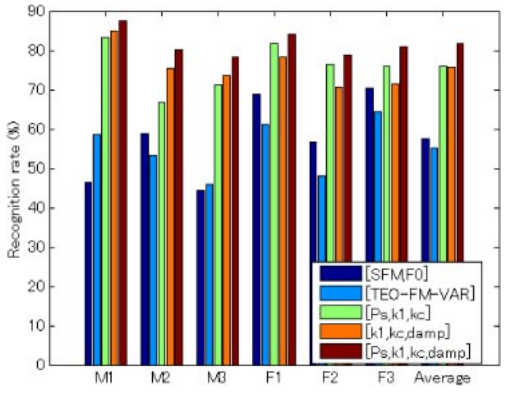


Fig. 12 Classification performance for F_0-SFM.

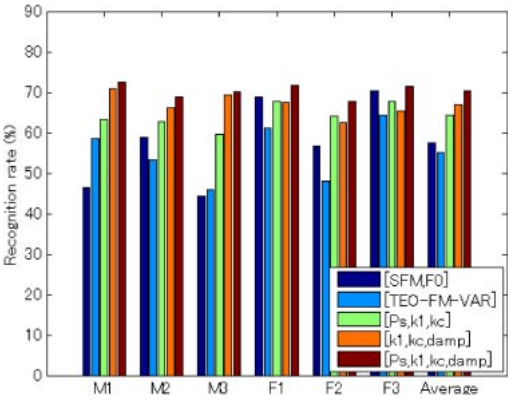


Fig. 14 Classification performance for Spect-Histo.

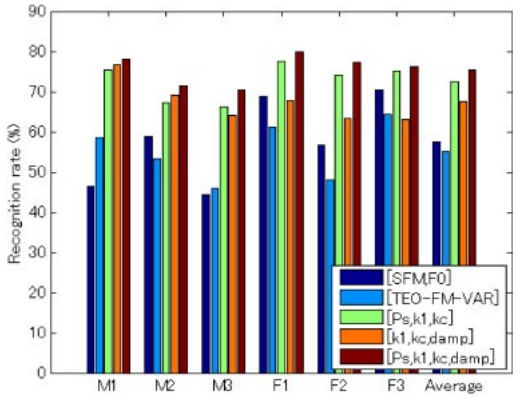


Fig. 13 Classification performance for F_0-Stat.

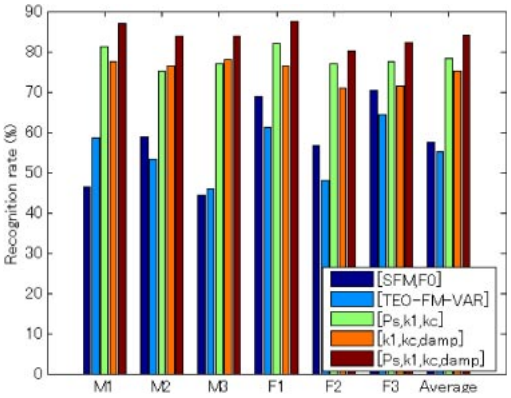


Fig. 15 Classification performance for Spectrum.

classification rate of $\{[P_S, k_1, k_c]\}$ was higher than that of $\{[k_1, k_c, \zeta_1]\}$ for female data. This suggests that females are more likely to exhibit stress vocally through variation in F_0 than male speakers, which agrees with the results above. Furthermore, results show that $\{[P_S, k_1, k_c, \zeta_1]\}$ had the best stress recognition performance among the physical parameter sets. This illustrates that stiffness, damping ratio of the vocal folds, and subglottal pressure are the factors that are affected when a speaker is under stress.

4.3. Comparison of Different Cost Functions

In the second evaluation, we also compare $\{[P_S, k_1, k_c]\}, \{[k_1, k_c, \zeta_1]\}, \{[P_S, k_1, k_c, \zeta_1]\}$ with different cost functions. For cost functions $C2$ and $C3$, the low- and high-frequency bands were separated on the basis of periodic characteristics of the harmonic in the spectrum. A linear classifier was used to examine their performance, and we took the average classification rate for all of the speakers to compare different cost functions. The results for different cost functions are shown in Figs. 12–15, and the average classification performance is shown in Fig. 16. Results show that cost function $C4$ yields the best improvement in classification performance.

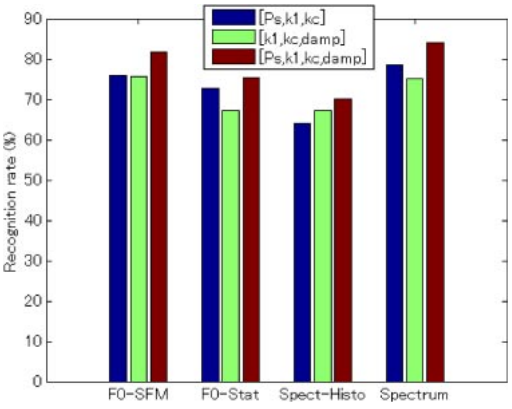


Fig. 16 Average results for proposed parameter sets with different cost functions.

Since the proposed features are based on physical characteristics, it would be helpful to compare their performances with those of traditional features. The traditional methods include the parameter sets $[F_0, SFM]$, and $[TEO - FM - VAR]$. $[TEO - FM - VAR]$ is the feature based on the Teager energy operator (TEO) to detect stress. It represents the frame-based variation of the frequency modulation (FM) component of the filtered signal [12]. The

results of this comparison are shown in Figs. 12–15. The proposed physical parameters perform better than the traditional features used for stress detection. This shows that parameters estimated from a physical model are more effective in representing speech under stress.

5. CONCLUSION

In this paper, we proposed more effective features for the classification of neutral and stressed speech. A physical model that characterizes the behavior of the vocal folds was used to simulate speech production. Physical parameters (stiffness, damping ratio, and subglottal pressure) were estimated using a method that fits the two-mass model to real speech, and different cost functions were used as targets to make a comparison. The obtained parameters were used as physical features for the classification of neutral and stressed speech. The conclusion drawn is that subglottal pressure from the lungs, muscle tension, and viscosity of the vocal folds, all of which affect the glottal source, are key indicators of stressed speech. Future work should be focused on the estimation of the parameters of both vocal folds and vocal tract for the classification of speech under stress.

REFERENCES

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

[2] R. Van Bezooijen, *The Characteristics and Recognizability of Vocal Expression of Emotions* (Foris, Drodrecht, 1984).

[3] F. J. Tolkmitt and K. R. Scherer, "Effect of experimentally induced stress on vocal parameters," *J. Exp. Psychol.*, **12**, 302–313 (1986).

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

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

[6] Z. S. Bond and T. J. Moore, "A note on loud and Lombard speech," *Int. Conf. Speech Language Processing '90*, pp. 969–972 (1990).

[7] J. H. L. Hansen, "Analysis and compensation of stressed and noisy speech with application to robust automatic recognition," *Ph.D. dissertation, Georgia Inst. Technol.*, Atlanta (1988).

[8] I. R. Murray, C. Baber and A. South, "Toward a definition and working model of stress and its effects on speech," *Speech Commun.*, **20**, 3–12 (1996).

[9] J. Whitmore and S. Fisher, "Speech during sustained operations," *Speech Commun.*, **20**, 55–70 (1996).

[10] A. Kamano, N. Washio, S. Harada and N. Matsuo, "A study of psychological suppression detection based on non-verbal information," *IEICE Tech. Rep.*, IEICE-SP2010-64, pp. 107–110 (2010) (in Japanese).

[11] J. F. Kaiser, "On Teager's energy algorithm and its generalization to continuous signals," *Proc. 4th IEEE Digital Signal Processing Workshop*, New Paltz, NY (1990).

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

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

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

[15] B. K. Finkelhor, I. R. Titze and P. L. Durham, "The effect of viscosity change in the vocal folds on the range of oscillation," *J. Voice*, **1**, 320–325 (1988).

[16] T. Kaneko, H. Asano, J. Naito, N. Kobayashi, K. Hayashi and T. Kitamura, "Biomechanics of the vocal cords-on damping ratio," *J. Jpn. Bronchoesophagol. Soc.*, **25**, 133–138 (1972).

[17] N. Isshiki, *Functional Surgery of the Larynx* (Kyoto University, Kyoto, 1977), pp. 62–67.

[18] P. H. Dejonckere and J. Lebacq, "Damping coefficient of oscillating vocal folds in relation with pitch perturbations," *Speech Commun.*, **3**, 89–92 (1984).

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

[20] J. L. Flanagan, *Speech Analysis, Synthesis, and Perception* (Springer-Verlag, New York, 1972).

Xiao Yao received B.E. degree from University of Shanghai for Science and Technology, Shanghai, China, 2004, and the M.E. degree from Tongji University, Shanghai, China in 2008. He is currently a Ph.D. in the Department of Media Science, Graduate School of information Science, Nagoya University, Nagoya, Japan. His research interests are focused on in the field of speech recognition and speech classification under stress.

Takatoshi Jitsuhiro received the B.E. and M.E. degrees in electrical engineering from Nagoya University, Japan, in 1991 and 1993. In 1993, he joined the Human Interface Laboratories, Nippon Telegraph and Telephone Corporation (NTT), Kanagawa, Japan. From 2000, he has been a researcher at ATR Spoken Language Translation Research Laboratories, Kyoto, Japan. He received the Ph.D degree in engineering from Nara Institute of Science and Technology in 2005. He is currently an Associate Professor at Aichi University of Technology since 2008. He also works on a part-time basis for Nagoya University from 2009. His research interests include speech recognition and speech signal processing. He is a member of ASJ, IEICE, and IPSJ.

Chiyomi Miyajima received the B.E., M.E., and Dr.Eng. degrees in computer science from Nagoya Institute of Technology, Nagoya, Japan, in 1996, 1998, and 2001, respectively. From 2001 to 2003, she was a Research Associate with the Department of Computer Science, Nagoya Institute of Technology. She is currently an Assistant Professor with the Graduate School of Information Science, Nagoya University. Her research interests include human behavior signal processing and speech processing.

Norihide Kitaoka received the B.E. and M.E. degrees from Kyoto University, Kyoto, Japan, in 1992 and 1994, respectively, and the Dr. Eng. degree from Toyohashi University of Technology, Toyohashi, Japan, in 2000. He joined DENSO CORPORATION, Kariya, Japan, in 1994. He joined the Department of Information and Computer Sciences, Toyohashi University of Technology, as a Research Associate in 2001 and was a Lecturer from 2003 to 2006. Since 2006, he was been an Associate Professor with the Department of Media Science, Graduate School of Information Science, Nagoya University, Nagoya, Japan. His research interests include speech processing, speech recognition, and spoken dialog.

Kazuya Takeda received his B.E., M.E. in Electrical Engineering and Doctor of Engineering degrees from Nagoya University, Nagoya Japan, in 1983, 1985, and 1994, respectively. From 1986 to 1989 he was with the Advanced Telecommunication Research laboratories (ATR), Osaka, Japan. His main research interest at ATR was corpus based speech synthesis. He was a Visiting Scientist at MIT from November 1987 to April 1988. From 1989 to 1995, he was a researcher and research supervisor at KDD Research and Development Laboratories, Kamifukuoka, Japan. From 1995 to 2003, he was an associate professor of the faculty of engineering at Nagoya University. Since 2003 he has been a professor at the Department of Media Science, Graduate School of Information Science, Nagoya University. His current research interests are media signal processing and its applications include; spatial audio, robust speech recognition and driving behavior modeling.