IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 17, 2018, accepted July 19, 2018, date of publication July 26, 2018, date of current version September 5, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2860130

Research on Speech Under Stress Based on Glottal Source Using a Physical Speech Production Model

XIAO YAO[1,2], NING XU[1,2], XIAOFENG LIU[1,2], AIMIN JIANG[1,2], AND XUEWU ZHANG[1], (Member, IEEE)

[1]College of Internet of Things Engineering, Hohai University, Changzhou 213022, China
[2]Changzhou Key Laboratory of Robotics and Intelligent Technology, Hohai University, Changzhou 213022, China

Corresponding author: Xiao Yao (yaox@hhu.edu.cn)

This work was supported in part by the National Nature Science Foundation of China under Grants 61501170, 61401148, 61471157, and 61671202, in part by the Fundamental Research Funds for the Central Universities under Grant 2015B25214, in part by the Key Development Program of Jiangsu Province of China under Grants BE2017071 and BE2017647, and in part by the Projects of International Cooperation and Exchanges of Changzhou of China under Grant CZ20170018.

ABSTRACT Speech recognition accuracy is severely reduced by the variability caused by stress. Considering the fact that speech under stress is caused by the physiological changes of the vocal folds in the physiological system whose vibration behavior is reflected by glottal flow, this paper presents a method of study on speech under stress based on a physical speech production model and characteristics of glottal flow. The physical model is used to model glottal aerodynamics in the vocal system to represent speech production. The relationship between physical parameters and glottal flow parameters is explored based on the physical model, and the glottal source and physical model are linked. Through studying on the glottal and physical parameters for the neutral and for the speech under stress, features for speech under stress characterizing the vocal folds, vortex–flow interaction, and shape of glottal flow are compared with those of neutral speech. The relations between the proposed parameters and stress-speech production mechanism are discussed. Experiments show that physical parameters representing the stiffness and viscosity of vocal folds, subglottal pressure, and laryngeal ventricle strongly influence the glottal flow. The relations for physical parameters, glottal parameters, and stress production are revealed, and theoretical and experimental bases are provided for stress detection and classification in speech recognition system.

INDEX TERMS Speech under stress, physical characteristics, glottal flow, the vocal folds, physical model.

I. INTRODUCTION

Stress, refers to sociology, psychology and physiology. Stress is our body's natural response to the inner or outer stimulations. The stimulations are caused by physical, mental, or emotional factors in nature. The occurrence of continuous stress response will result in obvious variations in vocal organ, impacting on the speech production. Stress is defined as ''the balance between the perceived demands from the environment and the individual's resources to meet those demands'' [1], [2].

Transformation of objective circumstance and the subjective psychological change of the speaker may produce stress and has impact on the speech production. The reasons of stress are varied, including emotion, Lombard effect, workload, multiple task, and physical state [3]. The speech signal will be affected substantially by these different causes. Therefore, the study of stress is believed to have important implications to many research fields like robust speech recognition, speech synthesis and emotion recognition.

Current research is mainly focused on acoustic parameters from the linear speech production model. Thomas and Thierry [4] proposed a method for the glottal analysis under the Lombard Effect. The acoustic parameters for fundamental frequency, and glottal source were proposed by Cumming and Clements [5] and Williams and Stevens [6] to detect stress with a hidden Markov model. With respect to vocal tract modeling, Hansen [7] and Hansen and Womack [8] proposed the features including formant frequency, Mel-frequency cepstral coefficients, cross-sectional areas, and spectral coefficients, representing the shape of the vocal tract.

A nonlinear model was proposed by Teager [9] and Teager and Teager [10], which considered the nonlinear characteristics of speech production. Teager and Teager [10] and Zhou *et al.* [11] believed that the existence of stress causes variability in the aerodynamics because of the changes of muscle tension of the vocal fold. Nonlinear features based on Teager energy operator (TEO) are proposed to separate the neutral and speech under stress [3]. However, the features proposed do not have an explicit physical meaning, and the methods do not take into account the speech production, which is considered as an essential issue in the research on speech under stress. So a physical model representing the whole process of speech production is helpful to study the airflow variability in the vocal system.

Our previous works have discussed that some vocal organs are affected when the speaker is under stress condition [13]. But the details about how the vocal organs are affected are not presented. Considering the fact that characteristics of vocal folds and the vortex-flow interaction between vocal folds and vocal tract play a more important role in the stress production, and that the vibration behavior of the vocal folds is reflected by glottal flow, this paper focuses on a study of speech under stress based on both physical model and glottal flow. The explanations of how the vocal system is influenced by stress are made. We discuss the relations among physical parameters, glottal flow, and stress production using the physical model. The characteristics of the vocal folds and vortex-flow interaction in laryngeal ventricle for speech under stress are analyzed. Variation of glottal flow affected by stress is studied by exploring the relationship between physical and glottal parameters. The essence and mechanism for stress production are revealed.

II. MODELING SPEECH PRODUCTION

Nonlinear theory proves that the vortex-flow interaction is essential for speech production [11]. Since the interaction depends on airflow separation around the laryngeal ventricle at the outlet from the glottis, the studies on airflow patterns in laryngeal ventricle provide a theoretical explanation for speech production.

The two-mass model proposed by Ishizaka and Flanagan [14], simulates the physical process of speech production. But the two-mass model failed to fully consider the laryngeal ventricle, which is believed extremely important in the stress production. A sketch of a modified physical model is shown in Figure. 1 In the model, the airflow patterns in the vocal system including the laryngeal ventricle and false vocal folds are modeled.

In the physical model, the vocal fold is simulated by two masses, with the stiffness and damping parameters.

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

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

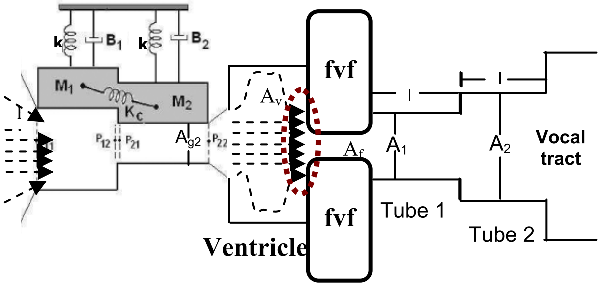

FIGURE 1. Sketch of modified two-mass model, representing the glottis, the vocal folds, laryngeal ventricle, and the vocal tract.

where m_i, r_i, s_i and F_i are the masses, the force of viscous resistance, elasticity and airflow respectively. x_i are their horizontal displacements of the two masses from the rest position. k_c here denotes the coupling stiffness between the two masses. The elasticity s_i we used are defined as:

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

where k_i are stiffness parameters.

For the damping force, related to the viscous resistance of vocal folds surface.

$$r_1 = 2\zeta_1\sqrt{m_1k_1} \quad r_2 = 2\zeta_2\sqrt{m_2k_2} \qquad (4)$$

where ζ refer to damping ratio for the viscous resistance, and m_i and k_i are the mass and stiffness of the spring we defined above.

The second part of the physical model is the laryngeal ventricle. Laryngeal ventricle is fossa of intermediate laryngeal cavity,located between the two cleft and above the vocal folds. The airflow patterns are modeled to characterize the pressure variation along the glottis, laryngeal ventricle, and the vocal tract [13].

We define the subglottal pressure as P_s, and the pressure in the inlet of glottis is expressed as:

$$P_s - P_{11} = \frac{\rho U_g^2}{2A_1^2}, \qquad (5)$$

where U_g is the volume velocity of glottal flow. A_{g1} denotes the cross-sectional glottal area, which is calculated by $A_{g1} = 2l_g(x_0 + x_1)$, where l_g is the length of the vocal folds, and x_0 is the displacement when the vocal fold is in the rest position. The pressure drop at the inlet of glottis is determined by further measurement, because of a vena contracta when the airflow enters the narrow glottis.

$$P_s - P_{11} = (1.00 + 0.37)\frac{\rho U_g^2}{2A_{g1}^2}, \qquad (6)$$

The air viscosity can cause the pressure drop, when the airflow passes through the masses.

$$P_{i1} - P_{i2} = \frac{12\mu d_i l_g^2 U_g}{A_{gi}^3}, \quad i = 1, 2 \qquad (7)$$

where μ is the air viscosity coefficient, and d_i denote the widths of masses. P_{22} is the air pressure at the outlet of glottis.

The air pressure drop is represented as follow equation in the junction between the two masses.

$$P_{21} - P_{12} = \frac{\rho U_g^2}{2} \left(\frac{1}{A_{g1}^2} - \frac{1}{A_{g2}^2} \right), \tag{8}$$

where P_{12} and P_{21} are the pressure at the higher boundary of m_1 and lower boundary of m_2, respectively.

The recovery of pressure P_{22} is achieved at outlet of glottis due to a sharp increase in cross-sectional area of laryngeal ventricle.

$$P_{22} - P_v = -\frac{\rho}{2} \cdot \frac{2}{A_{g2} A_V} \left(1 - \frac{A_{g2}}{A_V} \right) U_g^2, \tag{9}$$

where A_V denotes the cross-sectional area of laryngeal ventricle. P_v is the air pressure at this entrance. Here we neglected the pressure drop when airflow passes through the laryngeal ventricle to make a simplification of the physical model.

The pressure drops to P_{f1} when air arrives at the exit of the laryngeal ventricle, and continues to drop to P_{f2} along the false vocal folds, which given by:

$$P_v - P_{f1} = \frac{\rho}{2} \left(\frac{1}{A_f^2} - \frac{1}{A_v^2} \right) U_g^2, \tag{10}$$

$$P_{f1} - P_{f2} = 12 \frac{\mu l_f^2 d_f}{A_f^3} U_g, \tag{11}$$

where A_f, is the cross sectional area of the false vocal folds, and l_f and d_f, represent the length and thickness.

The air pressure will rise to the atmosphere value at the exit of false vocal folds, due to the steeply increase of cross-sectional area at the inlet of the vocal tract. The pressure P_1 is calculated by:

$$P_{f2} - P_1 = -\frac{\rho}{2} \cdot \frac{2}{A_f A_1} \left(1 - \frac{A_f}{A_1} \right) U_g^2, \tag{12}$$

The vocal tract is represented as four-tube acoustic resonator, which is shaped by tongue, mouth, teeth, uvula, and nasal cavity. The model is established using a transmission line analogy, terminate in a radiation load. The parameters for each section of tube calculated from cross-sectional areas $A_1 \cdots A_n$, which represent the characteristics for the vocal tract model.

III. DISCUSSION OF RELATIONSHIP AMONG PHYSIOLOGICAL CHARACTERISTICS, GLOTTAL SOURCE AND STRESS

We use the physical model to produce glottal flow, to explore how the glottal source is influenced by the physical characteristics related to stress. One physical parameter in the model is selected as a variable, and others are kept as their typical values. The glottal flow is simulated using the model with the determined parameters and three parameters from the glottal flow are calculated as measurements. The changing trend of glottal parameters with physical parameters is plotted to represent the relationship between physiological characteristics and glottal source.

A. MEASUREMENTS FOR THE GLOTTAL SOURCE

1) AREA UNDER THE AUTOCORRELATION (AUAC)

AUAC represents regularity of harmonic structure for the power spectrum. The regularity forthe harmonic periodic-structure of glottal flow spectrum, is quantified with an ‘‘area under the autocorrelation’’.

$$R_X(\Omega) = \int_{-\infty}^{\infty} S^*(\omega) S^*(\Omega - \omega) d\omega \tag{13}$$

$$AUAC = \int_{-\infty}^{\infty} R_X(\Omega) d\Omega, \tag{14}$$

where $S^*(\omega)$ is the spectrum of glottal flow within a limited high-frequency band (3000-4000Hz). AUAC is the area under the autocorrelation of $S^*(\omega)$. Therefore, the area under the ideal envelop of correlation with regular periodic-structure of spectrum is larger. In the case when periodic-structure is irregular, the autocorrelation will not present periodic characteristics, and hence the area will be smaller.

2) NORMALIZED AMPLITUDE QUOTIENT (NAQ)

The time domain parameter NAQ is representing the closing phase of the vocal folds [15], which is defined as:

$$NAQ = \frac{AQ}{T} = \frac{f_{ac}}{d_{peak} \cdot T} \tag{15}$$

where T is defined as pitch period, AQ (amplitude quotient) is defined by the maximum amplitude of the glottal flow and the largest negative peak of the first derivative.f_{ac}: The maximum amplitude of the glottal waveform. d_{peak}: The largest negative peak of the first derivative of the glottal waveform

3) SPEED QUOTIENT (SQ)

Speed quotient will be used as a measure of the glottal flow skewness and symmetry. It is defined as

$$SQ = \frac{t_{max} - t_o}{t_c - t_{max}} \tag{16}$$

where t_{max} denote the instant when glottal flow reach maximum, and t_{o} and t_{c} is the opening and closing instant respectively.

B. THE VOCAL FOLDS

The vocal folds are located at the center of glottis in an anterior-posterior orientation. The right and left folds show a ‘‘V’’-shaped structure. The ‘‘V’’ shaped opening forms the inlet to the trachea. On both sides of the larynx, each vocal fold is connected to an arytenoid cartilage with muscles. Through contracting or relaxing the muscles, the glottal flow is produced, which is believed as glottal source.

The structure of the vocal folds is described by body-cover theory [16]. Since the existence of the two tissue structures (cover, body) of the vocal folds, the body-cover theory suggests that the stiffness property of the vocal folds mainly depends on relative activations of the thyroarytenoid (TA) and

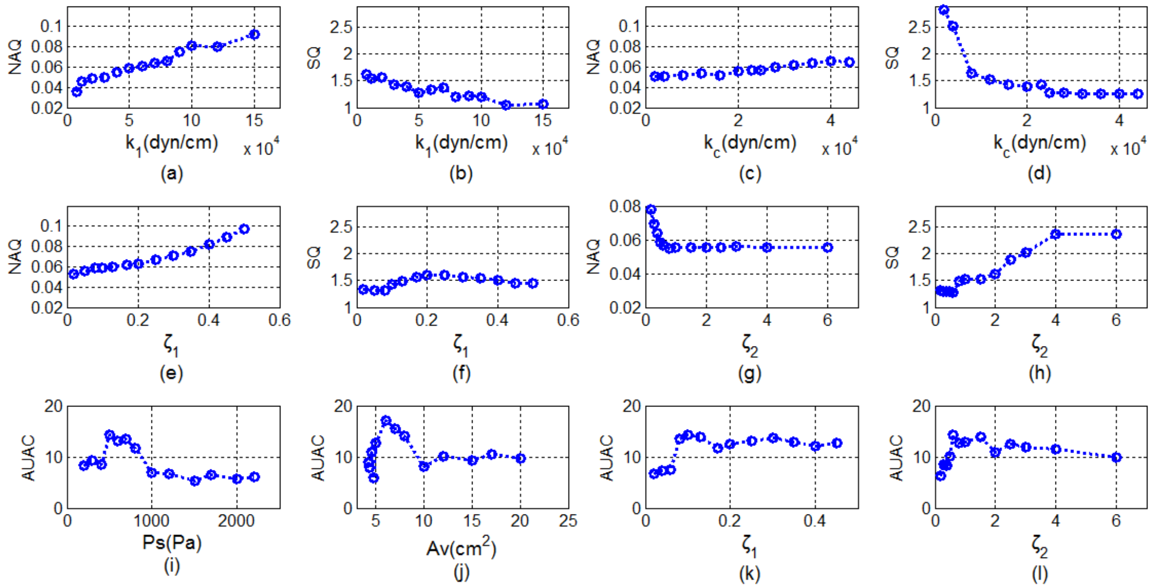

FIGURE 2. The relationship between physiological characteristics and glottal source. Physical parameters include the stiffness of muscle tension and damping ratio of viscosity for the vocal folds, area of laryngeal ventricle, and subglottal pressure. The glottal source is represented by parameters for area under autocorrelation (AUAC), normalized amplitude quotient (NAQ), and speed quotient (SQ).

cricothyroid (CT) muscles, corresponding body and layer. Therefore, the vibration characteristics of the vocal folds is represented by the couple oscillators of body and cover layers. Body-cover model functionally divides the vocal folds into body and cover layers. In order to simplify the model and reduce the number of degrees of freedom of the system, we consider only symmetric two-mass vocal fold model.

Vocal fold vibration is the source of phonation, and the glottal flow reflects the vibration styles of vocal folds. Since the physiological stress results in variations of stiffness of the vocal folds ($k_1 k_c$) [13], the changes of airflow patterns in the glottis will impact on the production of glottal flow. Therefore, discussing the parameters k_1, k_c of the vocal folds can lead to revelation of the effect of physiological stress on glottal source.

The relationships between vocal fold parameters and glottal source are shown in Figure 2. The normalized k_1, k_c versus NAQ and SQ in Figure 2(a)-(d) is a convenient medium to demonstrate the effect of muscle tension of the vocal folds on the glottal flow. An increase in k_1 raises the NAQ and reduces the SQ value. The change of NAQ with k_1 is significant, while SQ does not show a remarked decrease in Figure 2 (a) and (b). In contrast, an increase in k_c, also similarly increase NAQ and decrease SQ, but the difference is slight increase for NAQ and significant decline of SQ are achieved, as shown in Figure.2(c) and (d).

Generally, k_1 represents the tension in the cricothyroid muscle (CT), while coupling stiffness k_c is relative to the thyroarytenoid muscle (TA) [17], so a larger k_1 and a lower k_c indicate contracting CT and relaxing TA if the speaker is under stress. The relationship of k_1 and NAQ for stress production illustrates contraction of CT will cause the slow vocal folds closure during vibration. A decrease in k_c increases the phase difference between the opening and closing phase, raising the steep falling slope of glottal flow, and tends to make the flow more asymmetrical. Therefore, stress production indicates that the abduction and adduction behavior of vocal folds are decelerated and glottal flow become more asymmetrical and flat, shown in Figure 3.

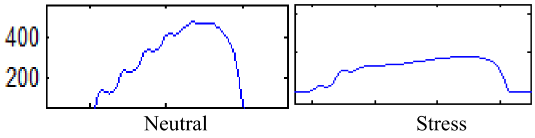

FIGURE 3. Glottal flow with characteristics of the vocal folds under neutral and stress condition.

The viscosity of vocal fold tissue is important in the vocal fold vibration, which represents the glutinosity of the wetted surfaces during adduction of the vocal fold, changes owing to hydration effects [18]. Damping ratio ζ_1 and ζ_2 is main factor to determine the viscosity of vocal folds tissues. Stress production will result from a larger damping ratio and lead to make the surface of vocal folds stickier.

Figure 2(e)-(h) show the relationship between damping ratio and glottal source. An increase in ζ_2 will produce a decreased NAQ and an increased SQ, shown in Figure 2(g) (h). It indicates that a more asymmetrical glottal flow is produced and time span for abduction is lengthened and adduction of the vocal folds is shortened. In contrast, an increase of ζ_1 for stress will result in a raised NAQ and

SQ (It is a slight decline after 0.3, but overall upward trend). In this case, it demonstrates both the adduction and abduction behavior of the vocal folds are prolonged, which indicate the vibration speed is slowed down and open quotient (OQ) of the glottal flow in a vibration period is increased, shown in Figure 3. It may induce the partial abduction and incomplete closure of the vocal folds, which is correspondent with the results of stiffness characteristics under stress we discussed above.

C. THE LARYNGEAL VENTRICLE

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

Negative pressure at the outlet of the glottis leads to instability of airflow, leading to the occurrence of airflow separation in the laryngeal ventricle. The separated airflow propagates along the wall of laryngeal ventricle, and the reattachment causes the turbulence, resulting in production of vortices. The airflow separation and vortices production will have impact on the effective area of laryngeal ventricle, and contribute significantly to the vortex-flow interaction between the vocal folds and the vocal tract, which is considered as essential distinction between neutral and speech under stress [3]. So, the parameter A_v representing the effective cross-sectional area of laryngeal ventricle is considered as an indicator of stress during speech production.

Figure 2(j) shows the changing trend of AUAC is raising first and then drop with the increase of A_V. The peak is reached at 6cm^2, which is the typical value during the speech production. It indicates that the periodic-structure of the spectrum of glottal flow shows its irregularity if A_V is too small or too large. The smaller A_V implies the abrupt contraction in laryngeal ventricle at the outlet from glottis, which causes that the amount of airflow cannot propagate fast through the narrow space, wandering at the inlet of the laryngeal ventricle, and changes their original direction. The vortices are produced and interaction between vocal folds and vocal tract is affected. The vortex-flow interaction pattern is represented by the ripples in the glottal flow, which is shown in Figure 4. When A_V gets larger, it indicates that the effective area of laryngeal ventricle is broadened, causing an increment of the amount of airflow separation, and the vortex-flow interaction between the vocal folds and the vocal tract is also affected. Therefore, stress production of speaker presents a larger or smaller A_V values, represented by the ripple in the glottal flow. It depends on the area ratio of vocal folds, laryngeal ventricle, and vocal tract, which determines the vortex-flow interaction patterns.

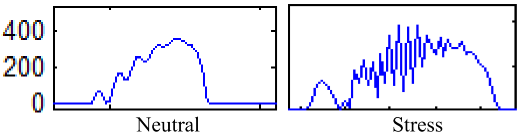

FIGURE 4. Glottal flow with characteristics of the laryngeal ventricle under neutral and stress condition.

D. SUBGLOTTAL PRESSURE

Subglottal pressure is the air pressure below the glottis, which is the source from lungs to make the vocal folds vibration. Subglottal pressure determines airflow velocity in glottis, thus, it has a great effect on glottal flow. In the model, P_s is used to represent the subglottal pressure under glottis.

In Figure 2 (i), the relationship between subglottal pressure and glottal source is illustrated. The changing trend of AUAC rises first and then declines with an increase of P_s. The peak explains the periodic-structure of the spectrum of glottal flow is regular, signifying the speaker is under the neutral condition.

When a speaker is under the stress condition, lower P_s is preference because the attention of speaker is distractive without the concentrate on the task at hand. The lower subglottal pressure induces the partial abduction of vocal folds because the lower pressure is not able to support the complete vibration. It makes the glottal flow excessively smooth, results from the incomplete open and closure of the vocal folds, and turns to the increased open quotient, which is in consistence with the results in damping ratio discussed above. The similar phenomenon is shown in damping ratio ζ_1 and ζ_2 in Figure 2 (k) (l), but the impact on AUAC is not evident.

The higher subglottal pressure for stress produces the airflow with high velocity. Turbulence flow at outlet of glottis and airflow separation in the laryngeal ventricle will occur caused by high speed airflow, which bring about the variability in the vortex interaction between vocal folds and vocal tract. The massive ripple in glottal flow, shown in Figure 5 can prove the explanation.

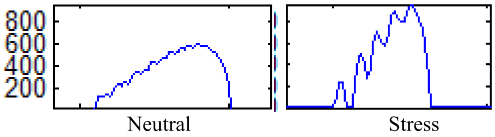

FIGURE 5. Glottal flow with characteristics of the subglottal pressure under neutral and stress condition.

IV. EVALUATIONS AND ANALYSIS

In this section, we use the real speech in database to verify the relationship between the physical parameters and glottal flow parameters. An evaluation and analysis are performed for the proposed method for speech under stress based on both physical model and glottal flow.

A. CORPUS DESCRIPTION

A database we used is the corpus that was gathered by Fujitsu [19], containing speech data from telephone conversations with topics performing dissimilar tasks. In order to simulate environment to produce psychological pressure, three tasks were introduced, which were completed by the speaker chatting with the telephone operator. The first dialog is relaxed conversation without any task, followed by dialog 2 and 3, in which the speaker was asked to complete the task to collect the speech data under workload. In dialog 4, there is an easy-flowing conversation without any tasks.

In the relaxed conversation, collected speech was the data when making a small talk with a speaker without any psychological pressure. In the small talk, a light topic was discussed, and the subject would be changed if the speaker felt embarrassed or unpleasant. Moreover, we provided a quiet and comfortable environment, where the air conditioner is used to make the speaker feel easy through maintaining appropriate temperature and humidity.

Three tasks corresponding to different mental states were introduced in order to simulate mental pressure resulting from psychological stress. The speaker was asked to complete the tasks in the conversation with the operator, and the attention of speaker was distractive without the concentration on the conversation. The stress under the workload condition is simulated.

1) CONCENTRATION

The task included the solution of the logical problem and spotting the differences of two pictures. Figure 6 gives an example for the logical problem task, in which the speaker needed to answer the logical question according to the hints provided and explained the whole process of reasoning.

Q. Fill in the table with the following hints:
Hint 1: Which musical Instruments does Kelly play?
Hint 2: Violin is located to the left of piano.
Hint 3: Alice does not like violin

Location	Left	Right
Person		
Instruments		

FIGURE 6. Logic puzzles for the task of concentration.

Figure 7 shows an example of a task finding differences between two pictures. In the experiment, the whole procedure was first explained in detail by the operator. The speaker was then given some clues when encountering troubles in spotting the differences. The difficulty would be upgraded with the game process, but the speaker cannot give up halfway before the whole process was completed.

2) TIME PRESSURE

The speaker was required to answer a few questions for a certain time. In this experiment, the speaker was asked to spot the difference in the provided two pictures, and meanwhile answer some question raised by the operator. Progress bar of elapsed time is displayed on the screen to produce the time pressure on speaker, who had to finish the questions in a certain time limit. The pictures would disappear on the screen if the time is exhausted. This experiment simulates the multitasking environment to make the speaker produce the stress for workload.

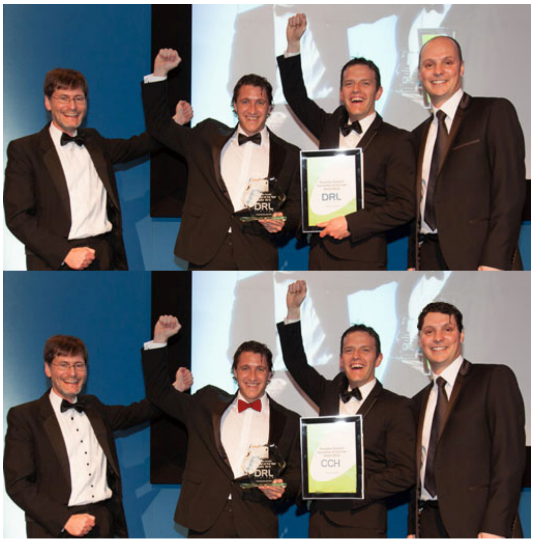

FIGURE 7. Spot the differences between two pictures for the task of concentration.

3) SPECULATIVE SPIRIT

Gambling games are used to measure the speaker's desire for money. The speaker used the poker to perform the gambling game with the target to win a certain amount of money. The operator kept phone connected with the speaker through the whole process of game. The speaker could like to borrow money from the operator in order to continue the game if the speaker gambled away all the money, or in order to win more money based on a principle that it is easier to win with more bets borrowed from the operator. The conversation between speaker and operator was recorded during the whole game.

The database contains speech data from 100 subjects, including 50 male and 50 female speakers. Based on the strict subjective evaluation, the speech from 11 speakers are selected as experimental data, whose voice show the marked changes when under the stress pressure of multitasking, comparing with the speech under relax condition. In the subjective evaluation, we define that the stress data were collected under workload condition, while neutral data were recorded from the relax chat

The segments for the vowels /a/, /i/, /u/, /e/, /o/ were cut manually from the speech data and selected as samples. The samples from 11 speakers constitute the database. The total number of sample is around 700-1500 for each speaker, which depends on different speech data from every speaker.

B. METHOD FOR FITTING TO THE MODEL

The physical parameters are estimated by fitting the model using the method of Analysis by Synthesis (A-b-S). Fitting the model to real speech poses a difficulty: the stability and performance of the fitting method may be influenced with too many parameters estimated simultaneously. An alternative way is to separate the fitting process into two parts with different cost functions: vocal folds fitting and vortex-flow interaction fitting. However, the vortex-flow interaction between the vocal folds and vocal tract makes it difficult to fit the two parts separately. Variations in the stiffness and damping ratio can influence both the vocal folds and interaction, and changes in subglottal pressure also have an impact on glottal source. The solution is to perform iteration for fitting the vocal folds and the interaction. Therefore, an iteration method is applied to estimate the physical parameters.

In the fitting method, cross-sectional areas of the four-tube model: A_1, A_2, A_3, and A_4 are estimated in the first step. Cost function 1 (C_1) is defined as the root mean square (RMS) distance between the spectral envelope of simulated and original speech ($P(\omega)$ and $P^*(\omega)$).

$$C_1 = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|\log P(\omega_i) - \log P^*(\omega_i)|^2} \tag{17}$$

In the second step, A_1, A_2, A_3, and A_4 are fixed at obtained values. k_1, k_c, ζ_1 and ζ_2 are considered as the control parameters in vocal folds fitting, and cost function 2 (C_2) is defined as:

$$C_2 = \frac{1}{N}\sum_{i=1}^{N}\left|Ug(t) - Ug^*(t)\right|^2 \tag{18}$$

where $Ug(t)$ and $Ug^*(t)$ are the glottal flow of the signals for simulated and real speech, respectively. A_V and P_s are selected as control parameters for the interaction fitting, and the cost function 3 (C_3) is defined as:

$$C_3 = \frac{2}{N}\sum_{i=N/2+1}^{N}\left|\log S(\omega_i) - \log S^*(\omega_i)\right|^2 \tag{19}$$

In the cost function, the power spectrum in the high frequency is used.

The accuracy of the fitting method is verified by comparing with conventional methods: formant synthesis and LPC synthesis. Log-spectral distance (LSD) as objective evaluation was proposed to represent the spectral distortion for real and simulated speech

$$\text{LSD} = \sqrt{\frac{1}{f(b)}\sum_{\omega_i \in B(b)}\left(10\log_{10}|S^*(\omega_i)| - 10\log_{10}|S(\omega_i)|\right)^2} \tag{20}$$

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

The results for the average values of log-spectral distance are illustrated in Figure. 8. It shows the fitting method is effective through the whole frequency band. Besides an improvement in the accuracy of spectrum simulation in high frequency due to study on harmonic structure of power spectrum in cost function, a better fitting effect in the low frequency is also achieved benefiting from the full consideration of characteristics of the glottal flow. This indicates that the proposed method provides reliable accuracy for the fitting to real speech.

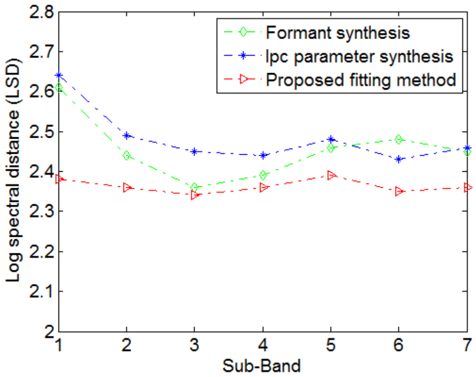

FIGURE 8. Evaluation of accuracy of the fitting method, comparing with the traditional methods.

C. EVALUATION RESULTS AND DISCUSSION

The evaluation for relationship establishment based on glottal flow source and physical parameters is performed, which is shown in Figure 9.

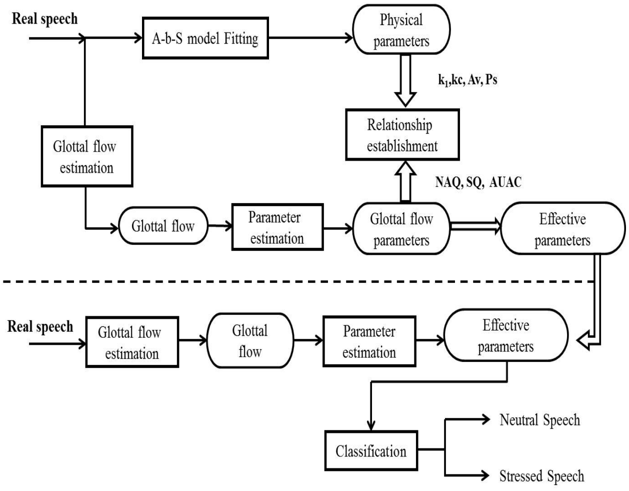

FIGURE 9. Diagram of evaluation for relationship establishment based on glottal flow source and physical parameters.

The physical parameters are calculated by fitting to a physical model using the method of Analysis-by-Synthesis. And meanwhile the glottal flow is estimated from the real speech using the method of iterative adaptive inverse filtering (IAIF), and the parameters representing the characteristics of glottal source are calculated. For the real speech data, LPC residual signal for the real speech is used to compute AUAC,

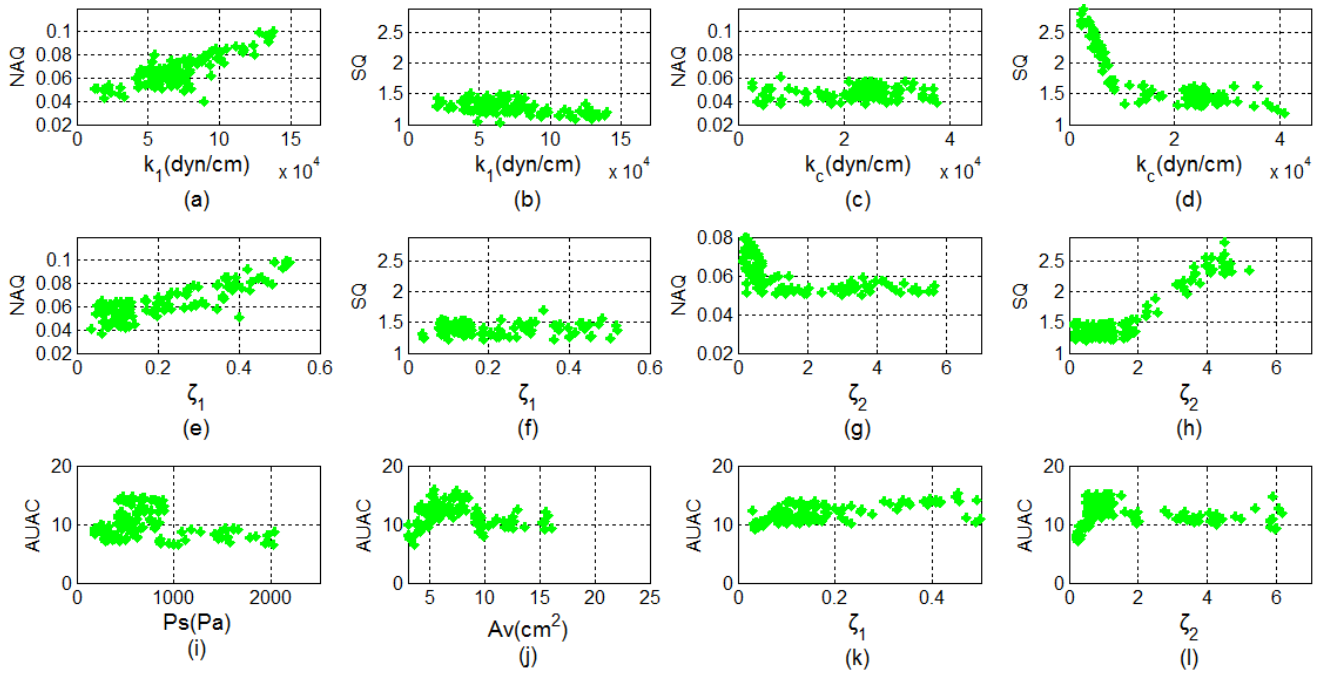

FIGURE 10. The evaluation of relationship between physiological characteristics and glottal source using the real speech. 120 samples are randomly selected from a database including mixed neutral and stress data for the vowel-independent condition.

representing the interaction between the vocal folds and the vocal tract.

The vowels (/a/, /i/, /u/, /e/, /o/) from real data (700 for each speaker) constitute a database. All of the vowels including neutral and stress samples were mixed for the vowel-independent condition. In this experiment, we selected randomly 120 samples from the database to evaluate each relationship. For a same speech data, the corresponding physical parameters and glottal parameters are extracted respectively, and are mapped into the two dimensional planes to verify the accuracy for the explored relationship we discussed above. The results are shown in Figure 10.

The evaluation results show the presenting variation trends from the real speech data are all fit the relation curves in Figure 2 with a high accuracy. Since the lack of enough stress samples in different levels, the changing trend presents the discontinuity.

The neutral and stress vowel samples are separated into the two databases, one is neutral database and the other is stress database. The classification performance for speech under stress is evaluated using the physical and glottal parameters. We estimated respectively the physical parameters (stiffness, subglottal pressure, area of laryngeal ventricle) and parameters from glottal flow (NAQ, SQ) and from LPC residual signal (AUAC) corresponding to neutral and stress data in real speech. The estimated parameters are vowel-independent and speaker independent. Table 1 shows the average values of the parameters for stress and neutral speech.

TABLE 1. Parameters for neutral and stress data from real speech.

Speech / Parameters	Neutral speech	Speech under stress
k_1 (dyn/cm)	62930	84635
k_c (dyn/cm)	23062	10538
ζ_1	0.1275	0.3529
ζ_2	0.6764	0.4581
Ps (Pa)	655	483
Av (cm^2)	6.2844	8.5206
NAQ	0.6473	0.7753
SQ	1.2005	1.8230
AUAC	13.5527	7.0684

In Table 1, the larger k_1 and smaller k_c for speech under stress raised from the distortion of muscle tension of the vocal folds indicate that the contraction of CT and relaxation of TA when the speaker under stress. It leads to deceleration for the abduction and adduction behavior of vocal folds and asymmetrical glottal flow, represented by a bigger NAQ and SQ. Damping ratio ζ_1 and ζ_2 for stress is significantly larger than neutral speech, which proves that stress leads to make the surface of vocal folds stickier. The produced larger NAQ and SQ in glottal flow indicates the asymmetrical glottal flow and the partial abduction and incomplete closure of the vocal folds. However, the bigger damping ratio does not cause distortion in periodicity of harmonic structure for spectrum, so the influence of stress on damping ratio is not represented in AUAC.

Table 1 shows that smaller P_s is produced when speaker under the stress. In fact, stress production of speaker presents a larger or smaller P_s values. The lower P_s is not able to support the complete vibration, inducing the partial abduction of the vocal folds, and higher P_s will accelerate airflow to cause turbulence flow and airflow separation in the laryngeal ventricle, impacting on the variability in the vortex interaction between vocal folds and vocal tract. Both the two cases bring about the smaller AUAC, represented by the incomplete open and closure of the vocal folds and the ripples in the glottal flow for lower and higher subglotal pressure. The similar case is found in A_v. A smaller A_v for stress implies the amount of airflow wanders at the inlet of the laryngeal ventricle, and produces the vortices, affecting interaction between vocal folds and vocal tract. The ripples in the glottal flow are produced. While larger A_v indicates that the broadened effective area at the inlet of the false vocal folds results in increase in airflow separation, impacting on the vortex-flow interaction, which leads to occurrence of frequency doubling in glottal flow. So stress production presents a larger or smaller A_v, but both are represented by a smaller AUAC. In this evaluation, the estimated P_s and A_v parameters from the stress database are clustered by the method of k-means in feature space. The smaller P_s and larger A_v are selected to calculate the average value for stress. Correspondingly, the larger P_s and smaller A_v are neglected and not described in the Table 1.

The samples of the stress class present great randomness because of a variety of stress levels and stress styles. According to the discussion above, different P_s and A_v present the various stress styles, and some stress samples in high-dimensional could not be clustered together and show great clustering distinctiveness. The different stress levels and styles determine the discontinuity of samples distribution in the feature space. Therefore, the GMM is used for stress classification since the various probability density functions following Gaussian distribution fit the different stress levels and styles.

In the evaluation of classification, two GMM models, corresponding to neutral data and stress data respectively, were used. Speech segments with all the vowels were mixed for the vowel-independent and speaker independent condition. The neutral model and stress model were trained from 3000 data, respectively. K-fold cross-validation method (K is 4) was applied to compute the average recognition rate.

The physical feature for $[k_1, k_c, \zeta_1, \zeta_2, P_s, A_V]$ and glottal feature for $[NAQ, SQ, AUAC]$ were modeled using GMMs with four mixture components. Classification performance is compared with traditional features. The traditional methods include the parameter sets $[SFM, F_0]$, $[TEO-FM-VAR]$, and [MFCC].

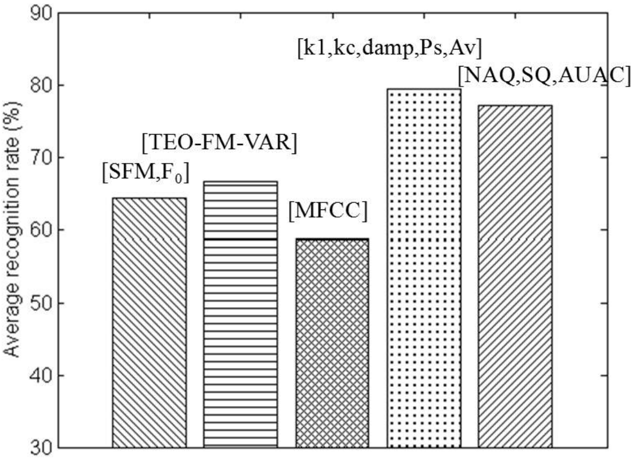

FIGURE 11. Evaluation of performance of stress classification under the speaker-independent and vowel-independent condition.

The classification results are shown in Figure 11. The proposed physical and glottal features show the advantage over the traditional features used for stress classification. Furthermore, the physical feature performs slightly better than the glottal features. It is achieved by the fact that the physical parameters are more direct and more essential to represent the physical characteristics in the physiological system. So the physical features are more accurate but the estimation method is a complex process for the fitting with iteration and depends on the selection of cost function. The glottal parameters are also work well in stress classification. The estimation method is simple and the theoretic basis of glottal parameters is provided based on the discussion of explored relationship with physical parameters, so the parameters of glottal flow can carry explicit physical meanings and show their effectiveness in the detection system of speech under stress.

V. CONCLUSION

This paper proposes a method for a research of speech under stress based on a physical model and glottal flow. The characteristics of the vocal folds and vortex-flow interaction in laryngeal ventricle for speech under stress and neutral speech are analyzed. And variation of glottal flow affected by stress is studied by linking the physical parameters with glottal flow based on the speech production model. The muscle tension and viscosity of the vocal folds, subglttal pressure, and area of laryngeal ventricle can be affected by stress, and have impact on the characteristics of glottal flow. The change properties of the glottal flow can represent the vibration characteristics of the vocal folds. When under stress, time span of abduction and adduction of vocal folds are prolonged and asymmetrical glottal flow are achieved, and the partial abduction and incomplete closure of the vocal folds are induced.

Furthermore, stress influences vortex-flow interaction between the vocal folds and the vocal tract, resulting from area-difference for vocal folds, laryngeal ventricle, and vocal tract. The results achieved in this paper reveal the essence and mechanism for stress production, and establish theoretical and experimental foundation for speech classification under stress.

REFERENCES

[1] M. Frankenhaeuser, "A psychobiological framework for research on human stress and coping," in *Dynamics of Stress: Physiological, Psychological and Social Perspective*. New York, NY, USA: Plenum, 1986, pp. 101–116.

[2] U. Lundberg, "Methods and applications of stress research," *Technol. Health Care*, vol. 3, no. 1, pp. 3–9, 1995.

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

[4] T. Drugman and T. Dutoit, "Glottal-based analysis of the Lombard effect," in *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. Chiba, Japan: Makuhari, Sep. 2010, pp. 2610–2613.

[5] K. E. Cumming and M. A. Clements, "Application of the analysis of glottal excitation of stressed speech to speaking style modification," in *Proc. ICASSP*, Apr. 1993, pp. 207–210.

[6] C. E. Williams and K. N. Stevens, "On determining the emotional states of pilots during flight: An exploratory study," *Aerosp. Med.*, vol. 40, pp. 1369–1372, Jan. 1969.

[7] J. H. L. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 598–614, Oct. 1994.

[8] J. H. L. Hansen and B. D. Womack, "Feature analysis and neural network-based classification of speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 4, pp. 307–313, Jul. 1996.

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

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

[11] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 201–206, Mar. 2001.

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

[13] X. Yao, T. Jitsuhiro, C. Miyajima, N. Kitaoka, and K. Takeda, "Modeling of physical characteristics of speech under stress," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1801–1805, Oct. 2015.

[14] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell Syst. Tech. J.*, vol. 51, no. 6, pp. 1233–1268, Jul. 1972.

[15] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *J. Acoust. Soc. Amer.*, vol. 112, no. 2, pp. 701–710, 2002.

[16] M. Hirano, "Morphological structure of the vocal cord as a vibrator and its variations," *Folia Phoniatrica Logopaedica*, vol. 26, no. 2, pp. 89–94, 1974.

[17] J. C. Lucero, "Chest- and falsetto-like oscillations in a two-mass model of the vocal folds," *J. Acoust. Soc. Amer.*, vol. 100, no. 5, pp. 3355–3399, 1996.

[18] B. K. Finkelhor, I. R. Titze, and P. L. Durham, "The effect of viscosity changes in the vocal folds on the range of oscillation," *J. Voice*, vol. 1, no. 4, pp. 320–325, 1988.

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

XIAO YAO received the M.E. degree from Tongji University, Shanghai, China, in 2008, and the Ph.D. degree from Nagoya University, Nagoya, Japan, in 2013. He is currently with the College of Internet of Things Engineering, Hohai University, Changzhou, China. His research interests include speech modeling, speech recognition, neutral network, artificial intelligence, computer vision, and natural language understanding.

NING XU received the B.E., M.E., and Ph.D. degrees from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2005, 2007, and 2010, respectively. He is currently with the College of Internet of Things Engineering, Hohai University, Changzhou, China. His research interests include speech signal processing, voice conversion, and natural language understanding.

XIAOFENG LIU received the B.S. degree in electronics engineering and the M.S. degree in computer science from the Taiyuan University of Technology in 1997 and 1999, respectively, and the Ph.D. degree in biomedical engineering from Xi'an Jiaotong University in 2006. He is currently a Professor with the the College of Internet of Things Engineering, Hohai University, China, where he is also the Leader of the Cognition and Robotics Laboratory and the Director of the joint Laboratory of Aldebaran Robotics and Hohai University. He has authored 20 accredited journal papers. He holds over 11 grants as a PI and over 12 grants as a Researcher, including the National High-Tech Research and Development Program (863) and the National Basic Research Program (973). He holds 15 granted patents. His current research interests focus on human–robot interactions, social robotics, and neural engineering.

AIMIN JIANG received the B.E. and M.E. degrees in electrical engineering from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2001 and 2004, respectively, and the Ph.D. degree in electrical engineering from the University of Windsor, Windsor, Canada, in 2010. He is currently with the College of Internet of Things Engineering, Hohai University, Changzhou, China. His research interests include mathematical optimization and its applications to digital signal processing and communications. He has served as a member for the Digital Signal Processing Technical Committee of the IEEE Circuits and Systems Society.

XUEWU ZHANG received the B.E., M.E., and Ph.D. degrees from the College of Computer and Information, Hohai University, China, in 1998, 2006, and 2011, respectively. He is currently a Professor with the College of Internet of Things Engineering, Hohai University. His research interests include computer vision and artificial intelligence, signal processing, and human behavior perception modeling.