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

Xiao Yao¹, Takatoshi Jitsuhiro¹², Chiyomi Miyajima¹, Norihide Kitaoka¹, Kazuya Takeda¹

¹Department of Media Science, Graduate School of Nagoya University, Nagoya-shi, Aichi, Japan
²Department of Media Informatics, Aichi University of Technology, Gamagori-shi, Aichi, Japan

xiao.yao@g.sp.m.is.nagoya-u.ac.jp, jitsuhiro@aut.ac.jp,
miyajima@nagoya-u.jp, kitaoka@nagoya-u.jp, kazuya.takeda@nagoya-u.jp

Abstract

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

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

1. Introduction

The effect of stress on speech signals has been the topic of numerous studies. Many factors, such as emotional state, fatigue, physical environment, and workload can cause people to experience stress. By studying speech under stress, we can improve the performance of speech recognition systems, recognize when people are in a stressed state, and better understand the context in which speakers are communicating.

Researchers have attempted to probe reliable indicators of stress by analyzing acoustic variables. The first investigations of emotional speech were conducted by Van Bezooijen [1] and Scherer [2], who used the statistical properties of acoustic features to recognize emotions from speech in the mid-1980s. Williams and Stevens found that the fundamental frequency (F0) has different characteristics for each emotion [3], and that respiration patterns and muscle tension also change due to a speaker’s emotional state [4]. The influence of the Lombard effect on speech recognition has been examined in [5], which analyzed selected acoustic features, such as amplitude and distribution of spectral energy, and found that spectral energy shifted to higher frequencies for consonants when speakers increase the volume of their voices. High workload stress has been proven to have a significant impact on the performance of speech recognition systems, with speech under workload sounding faster, softer, or louder than neutral speech [6]. In 2011, Matsuo, et al. examined the frequency domain, and showed how differences in the spectrum of the high frequency band of speech of speakers under stressful workload conditions aimed to catch people committing remittance fraud, and the proposed measure achieved better performance [7].

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

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

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

2. Physical modeling

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

\[
m_1 \frac{d^2 x_1}{dt^2} + r_1 \frac{dx_1}{dt} + s(x_1) + k_1(x_1 - x_2) = F_1, \quad (1)
\]

\[
m_2 \frac{d^2 x_2}{dt^2} + r_2 \frac{dx_2}{dt} + s(x_2) + k_1(x_1 - x_2) = F_2, \quad (2)
\]

where \(m\) are the masses, \(x_i\) are their horizontal displacements measured from the rest (neutral) position \(x_0 > 0\), and \(k_1\) is the coupling stiffness. In this equation, \(si\) are the equivalent tensions with non-linear relations given by

\[s_i(x_i) = k_i \left( x_i + \eta x_i^3 \right), \quad (3)\]
where $k_i$ are stiffness coefficients and $\eta$ is a coefficient of the nonlinear relations.

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

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

2.1. Modeling airflow aerodynamics

2.1.1. Pressure drop at the glottis

The aerodynamics of the glottis are modeled using the two-mass model. If subglottal pressure is represented as $P_1$, then the pressure drops to $P_{11}$ when entering the glottis (at the edge of $m_1$) according to Bernoulli’s equation:

$$P_2 - P_{11} = \frac{\rho U_g^2}{2} A_1,$$

(4)

where $\rho$ is the air density, and $U_g$ is the volume velocity of glottal airflow. $A_1$ is the cross-sectional lower glottal area, which can be represented by $A_1 = 2 l_1 (x_0 + x_i)$, where $l_1$ is the length of the vocal folds, and $x_i$ is the displacement when the vocal fold is in the rest position. Abrupt contractions in the cross-sectional area at the inlet to the glottis cause a vena contracta to occur, which causes an even greater drop in pressure. The drop is determined using the flow measurements from van den Berg:

$$P_2 - P_{11} = (1.00 + 0.37) \frac{\rho U_g^2}{2} A_1,$$

(5)

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

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

(6)

where $\mu$ is the air viscosity coefficient, and $d_i$ are the widths of $m_1$ and $m_2$, respectively. $P_{22}$ is air pressure at the glottal exit.

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

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

(7)

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

2.1.2. Pressure drop around laryngeal ventricle and false vocal folds

Next, we model airflow patterns around the laryngeal ventricle and false vocal folds. At the glottal outlet, expansion causes air pressure to recover because of the relatively larger area of the laryngeal ventricle. This pressure rise is represented by:

$$P_{22} - P_r = \frac{\rho}{2} \frac{1}{A_{g2}} \left( \frac{1}{A_{g1}} - \frac{1}{A_{g2}} \right) U_g^2,$$

(8)

where $A_{g1}$ is the area at the entrance to the laryngeal ventricle, and $P_r$ is the pressure at this inlet. In order to simplify our model, we disregard the pressure changes when air enters the laryngeal ventricle. Therefore, we assume airflow is uniform without any expansion $A_{g2} = A_{g1}$.

When air passes the laryngeal ventricle between the true vocal folds and false vocal folds, it is very unstable because of the negative pressure difference. Airflow separation occurs along the wall of laryngeal ventricle. After passing this region, the airflow propagates as a plane wave entering the false vocal folds. Separation causes variations in the effective area of the laryngeal ventricle into the false vocal folds. Therefore, it is hypothesized that the effective area of the ventricle changes in relation to airflow separation in this area. Here, we use $A_{vf}$ to represent the effective area of the ventricle into the false vocal folds. The pressure drop at the inlet of the false vocal folds is calculated according to Bernoulli’s equation:

$$P_{22} - P_{f1} = \frac{\rho}{2} \frac{1}{A_{vf}} \left( \frac{1}{A_{vf}} - \frac{1}{A_{g2}} \right) U_g^2,$$

(9)

where $A_{vf}$ is the area of the false vocal folds. Since the false vocal folds do not vibrate during the process of phonation, $A_{vf}$ can be fixed to a constant.

Along the false vocal folds, pressure drops from $P_{f1}$ to $P_{f2}$ due to the loss from air viscosity:

$$P_{f1} - P_{f2} = \frac{12 \mu U_g^2}{A_{vf}} d_i,$$

(10)

where $l_i$ and $d_i$ are the length and thickness of the false vocal folds, respectively.

Since the area of the vocal tract is relatively large compared with the glottal area, an abrupt expansion cause the pressure to recover toward the atmospheric value at the inlet to the vocal tract.

$$P_{f2} - P_1 = \frac{\rho}{2} \frac{2}{A_{g1}} \left( \frac{1}{A_{g1}} - \frac{1}{A_{vf1}} \right) U_g^2,$$

(11)

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

The effective area of the ventricle into the false vocal tract $A_{vf}$ can represent the variation in airflow pattern, which has a modulating effect on produced speech. Therefore, it is our assumption that this area parameter can be used as an indicator for stress classification.
In this section, we describe experiments which were performed to represent the effect of proposed parameter $A_V$ on the formants and SFM. We selected the formants (F1, F2, F3), the fundamental frequency ($F_0$) and the spectral flatness measure (SFM) as stress measurements. Formants depend on the shape of the vocal tract, while $F_0$ and SFM represent characteristics of the glottal source generated from the vocal folds.

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

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

where $S(\omega)$ and $S^*(\omega)$ are the power spectrums of the signals for simulated and real speech, respectively, after Fourier transform. Optimal values of the physical parameters are estimated using a Nelder-Mead simplex method [19], which is implemented to search for the optimal stiffness parameters which minimize the cost function.

4. Experiments

4.1. Database and experimental setup

In our experiments, we used a database collected by the Fujitsu Corporation containing speech samples from seven subjects (three male, and four female). To simulate mental pressure resulting in psychological stress, three different tasks were introduced, which were performed by the speakers while having telephone conversations with an operator, in order to simulate a situation involving pressure during a telephone call. The three tasks involved (A) Concentration; (B) Time pressure; and (C) Risk taking. For each speaker, there are four dialogues with different tasks. In two dialogues, the speaker is asked to finish the tasks within a limited amount of time, and in the other dialogues there is relaxed chat without any task.

The segments with the Japanese vowels /a/, /i/, /u/, /e/, /o/ were cut from the speech and selected as samples. All of the vowels were mixed for the vowel-independent condition. The experiments were conducted for each speaker, and all of the results were speaker dependent. Here, we used samples from seven subjects (three male, four female) to show the classification performance for each speaker, respectively, in this speaker-dependent system. The number of samples depended on the speakers, and the total amount is about 450-700 for each person. In order to increase the significance level of the experimental results, a K-fold cross-validation method was used in the classification experiments, with 60% of samples used for training, and the rest used for testing. K was set to 4. Linear classifiers based on minimum Euclidean distance, which fit a multivariate normal density to each group, with a pooled estimate of covariance, were used to determine classification performance. The samples were analyzed with 12th-order LPC and the frame size chosen to perform the experiment was 64ms, with 16ms for frame shift.

4.2. Results and analysis

4.2.1. Effect of parameter $A_V$

In this section, we describe experiments which were performed to represent the effect of proposed parameter $A_V$. We selected the formants (F1, F2, F3), the fundamental frequency ($F_0$) and the spectral flatness measure (SFM) as stress measurements. Formants depend on the shape of the vocal tract, while $F_0$ and SFM represent characteristics of the glottal source generated from the vocal folds.

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

In order to further evaluate the influence of $A_V$ on the spectrum, comparison of spectral distortion for real speech
and simulated speech with and without estimation of AV, was made. Log-spectral distance (LSD) was used to describe the difference in spectral distortion between real and simulated.

\[
LSD = \sqrt{\frac{1}{f(b)_{\text{sub-band}}} \sum_{i} \left( 10 \log_{10} |S'(\omega_i) - 10 \log_{10} |S(\omega_i)| \right)^2}
\]

where \(f(b)\) denotes the bandwidth of sub-band \(b\) and \(B(b)\) consists of a set containing all the discrete frequencies in sub-band \(b\). \(S(\omega_i)\) and \(S'(\omega_i)\) are the power spectrums of the residual signals for simulated and real speech, respectively.

The sub-bands are described in Table 1.

<table>
<thead>
<tr>
<th>Sub-Band</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>0-1000 (Hz)</td>
<td>500-1500 (Hz)</td>
<td>1000-2000 (Hz)</td>
<td>1500-2500 (Hz)</td>
<td>2000-3000 (Hz)</td>
<td>2500-3500 (Hz)</td>
<td>3000-4000 (Hz)</td>
</tr>
</tbody>
</table>

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

4.2.2. Evaluation of physical parameters

In this section, we compared the performance of two physical parameter sets, \([k_1, k_2]\) and \([k_1, k_2, AV]\), to evaluate the effectiveness of proposed parameter AV. The results are shown in Figure 5. When AV is taken into account, classification performance is improved. Since the samples selected in the experiment are mixed data from all the vowels, the results show that AV can maintain its performance under vowel-independent conditions, because the area of the ventricle has less impact on the vocal tract, and thus does not rely on vowel information. From these results, it is believed that AV is an essential parameter strongly related to stress. Larger AV values indicate that the amount of airflow separation is increasing, causing the effective area at the inlet of the false vocal folds to broaden. This results in variations in the airflow pattern around the false vocal folds, causing a stronger modulation effect on the speech produced.

5. Conclusions

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

6. Acknowledgements

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


