CLASSIFICATION OF SPEECH UNDER STRESS BASED ON PHYSICAL CHARACTERISTICS OF VOCAL FOLDS VIBRATION

YAO Xiao

A dissertation submitted to the Graduate School of Information Science, Nagoya University.

日本語のタイトル

声帯振動の物理特徴に基づくストレス環境下での音声の検出

Department of Media Science
Nagoya University
Nagoya, Japan.

AUGUST, 2013

ABSTRACT

The performance of automatic speech recognition systems (ASR) is degraded by stress-inducing environments, such as those where noisy backgrounds, multi-tasking, fatigue, emotional situations, adverse physical conditions, and high workload stress are present. The study of speech under stress, also referred to as stressed speech, can help maintain and improve the robustness of speech recognition systems in these situations. The primary objective of this dissertation is to perform the classification of speech under stress based on a physical model. Presence of stress will cause speakers to change their physiological system to react and adapt himself to the stressed condition. Changes in physiological characteristics can result in variations in aerodynamics in the glottis and the vocal tract, and the stressed speech is produced. Therefore, a physical model is necessary to model airflow patterns in the physiological system in order to represent the process of speech production. An investigation on how physical model is used for classification of speech under stress (stress classification) is explored. This dissertation includes following objectives:

I. Explore the physical characteristics of the vocal folds during speech under stress and estimate the physical parameters of the vocal folds for stress classification (the first study);

II. Explore the physical characteristics of both the vocal folds and the vocal tract for stressed speech and propose the effective physical parameters (the second study);

III. Model the aerodynamics in the laryngeal ventricle and the false vocal folds, and estimate a parameter for classification of stressed speech (the third study);

IV. Use different classifiers to perform classification and compare their performance (the

fourth study);

In the first study, a scheme for the classification of speech under stress based on a physical model is proposed. The physical model representing the speech production models the airflow patterns in the glottis and parameters derived from the model can characterize the behavior of the vocal folds during speech production. A method that fits a two-mass model to real speech is proposed in order to estimate the physical parameters representing 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.

In the second study, the characteristics of both the vocal folds and the vocal tract are taken into consideration. Methods for fitting a physical model to real speech in order to estimate the physical parameters of the vocal folds and the vocal tract are proposed. For the physical model, a two-mass model connected to a four-tube model is used to simulate the process of speech production. The physical parameters (stiffness, vocal tract length and cross-sectional areas of vocal tract) are estimated by fitting the model to real speech. Different cost functions are proposed to evaluate the estimated physical parameters and to compare classification performance.

In the third study, we focus on variations in the aerodynamics of airflow patterns in the laryngeal ventricle and the false vocal folds 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 to estimate a parameter representing the effective area of laryngeal ventricle. Experiments results

show the estimated parameters are effective to separate stressed speech from neutral speech.

In the last study, we present a comparison for classification performance of speech under stress based on the use of different classifiers. Physical features are modeled using a linear classifier, a Gaussian Mixture Model (GMM), and Support Vector Machine (SVM), respectively.

Based on the evaluation experiments for the physical parameters, physical features represented by the proposed method are compared with traditionally used features. Results demonstrate an improvement in stress classification performance, which shows that the proposed method is more effective than conventional methods. Comparing the different parameters used to represent the vocal tract, which are effective under the vowel-dependent condition, we conclude that the vocal fold parameters are more important for stress classification. In addition, we determine that the parameter characterizing variation in the airflow patterns in the laryngeal ventricle and the false vocal folds can also be used to achieve better classification performance.

Current areas in which the proposed methods could be applied include detection of psychological suppression from workload detected in the voice of victims during remittance fraud scams. Also, they could be used for detection of stress in the voices of drivers, resulting in improved driving safety. However, uses of this technology are not limited, and could be extended to improving emotional speech recognition. Our proposed method may be combined with other methods to achieve better speech recognition performance in various different stressful environments.

ACKNOWLEDGMENTS

I would like to express my great appreciation to my academic advisor, Professors Kazuya Takeda, for his tremendous kindness whenever I needed help with my research or career. He was well qualified to guide me, and his invaluable support, motivating supervision, and constructive advice have been the most important factors in helping me sharpen my thinking, focus my research and improve my professional skills.

I would also like to express my gratitude to my supervisor, Professor Takatoshi Jitsuhiro, who has guided me throughout the years of my Ph.D. studies. From the initial idea to the final version of this dissertation, he continued to give me invaluable comments and advice to help me focus my studies.

Additionally, I would like to express my appreciation to Professor Norihide Kitaoka, who gave me important guidance for my studies and for this dissertation. He also frequently shared his personal time to help me outside of the lab. Thank you Kitaoka sensei for your kind support.

And I would like thank Professor Chiyomi Miyajima, who gave me many valuable comments on how to write my papers, and helpful suggestions for my presentations.

I extend my deep appreciation to the professors, staff, and my colleagues in the Takeda Laboratory for their continued support, help, and friendship, and for the interesting cultural exchange which helped me to better understand Japanese culture. I also thank my friends who helped me get through my four, fantastic years in Japan.

Finally, I thank my parents, who allowed me to pursue my dream in Japan. Their continuous support encouraged me to be stronger and more confident.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 OVERVIEW

1.1.1 BACKGROUND

With the rapid development and continuous improvement of information technology, a general concern has become raising the level of human-computer interaction. Emotion recognition and artificial intelligence are so closely associated that the ability of computers to recognize and adapt to the influence of the environment on humans will result in another breakthrough in this field. The development of computers which are more human-like, intelligent, and able to interact naturally with humans is one of the future goals of current development efforts. In addition to segmental information, suprasegmental information (such as pitch and juncture pattern) also plays an important role in human communication. Suprasegmental refers to properties of an utterance that apply to groups of segments, rather than to individual segments. The three main suprasegmental features are stress, intonation, and tone. If computers can be designed with such abilities to analyze the suprasegmental information from speech, and thus can, like human beings, perceive speakers' psychological and physiological states caused by environmental factors, the barriers between man and machine will be further diminished.

Traditional human-computer interaction occurs through the operation of a keyboard, mouse, or screen display. This interaction is simply the machine recognizing commands, instead of perception

of the user’s psychological state. Human-computer interaction will not be able to reach the same level of natural communication as interaction between human beings if computers lack the ability to perceive human mental states. Most studies of emotional perception, however, only focus on visual information, such as facial expressions and body gestures [1], ignoring cues contained in speech information, which is rich in information about the speaker’s emotional state. This is because speech contains not only actual semantic content, but is also a carrier of information about the psychological and physiological state of the speaker [2]. Human language, which is directly controlled by the brain, is inextricably linked with the speaker's physiological condition.

Research in the field of speech recognition has made great progress. With speech recognition systems advanced to the level of a practicality, environmental factors affecting the performance of speech recognition algorithms have become increasingly important. Some researchers have found that subtle changes in environmental factors seriously affect the accuracy of speaker identification, speech recognition, speech synthesis, and so on. It is necessary to overcome the influence of environmental factors to make speech recognition systems truly effective.

Stress is a psycho-physiological state caused by environmental factors. It is characterized by subjective strain, dysfunctional physiological activity, and deterioration of performance [3]. There are many environmental factors which affect the performance of speech recognition systems, including background noise, changes in transmission channels, psychological stress, work pressure and emotions. Stress caused by these factors results in variations in speaker’s pronunciation, making highly reliable speech recognition systems difficult to achieve. Therefore, the detection and classification of speech under stress has become a popular subject of research.

1.1.2 DEFINITION OF STRESS

Stress is an essential definition in the field of science. It is a complex term, with social, psychological and physiological elements. It is defined as a psycho-physiological state contributing to emotional reactions, subjective strain, deterioration of performance and dysfunctional physiological activity. Stress is subjective, and can be considered as the physical reaction of a speaker who is forced to adapt to an event or situation. Based on psychological theories, stress has been defined as "the balance between the perceived demands from the environment and the individual's resources to meet those demands" [4] [5].

The Cognitive Activation Theory of Stress was proposed by Ursin and Eriksen, who reported that cognitive processes and neurophysiological activation are the main factors contributing to stress arousal [6]. Physiological stress model and General Adaption Syndrome (GAS) were also proposed to associate emotional expression with physiological changes [7] [8].

Typical physiological indicators of stress include elevated levels of cortisol and/or epinephrine, elevated heart rate and elevated blood pressure. The amount of epinephrine present in the blood is regarded as the key measure of the intensity of stress, but epinephrine, blood pressure, and heart rate do not always reflect emotional stress [5]. Other studies have shown that cortisol level is a better indicator of emotional stress, but an increase in cortisol does not relieve depressive moods or help one escape from stressful environments [9]. Blood pressure and heart rate are commonly regarded as the best measures of stress in the natural environment, as they reflect the effect of stress on the sympathetic and parasympathetic nervous systems [10].

Changes in the surrounding environment or in the speaker's own physical condition can induce stress and cause variations in the speech produced. Stress can be caused by physical or psychological factors, including G-force, lack of sleep, physical workload, psychological workload, nervousness, noisy conditions, pain, confusion, doubt, emotions, and even by typical events which occur in adverse environments [11]. These stress factors are considered as impacts on the quality of speech, and can interfere with the performance of communication equipment and vocal interfaces.

When the surrounding environment or the speaker's physical or mental state change, there are variations in the speech produced [12]. In some specific environments, such as in vehicles, helicopters, warplanes, and noisy factories, the speaker is mostly focusing on a particular task, and their phonation is just a low priority auxiliary to their primary activity. Speech produced can be greatly impacted by the existence of work pressure. When a speaker tries to adjust their speaking style to achieve clearer expression due to the presence of background noise, the change in their speaking style is known as the Lombard effect [13], the influence of which depends on the strength and type of background noise. Speakers also change the style of their phonation as a result of emotions such as anger, sadness, happiness, and fear. Furthermore, stressed speech is also generated when a speaker's body is suffering from physical trauma. Therefore, stressed speech can be categorized according to variations in the physical layer, physiological layer, perceptual layer, and psychological layer [14].

In most cases, the human auditory system can correctly perceive and distinguish the stressed speech of others, capturing the speaker's psychological tension and emotional changes. However, computer is not able to resolve this problem, resulting in misunderstanding and leading to serious consequences.

Although there are many difficulties in the classification of stressed speech, many scholars have devoted themselves to studies in this field. This is because it has become increasingly important to study speech under stress in order to detect speakers' psychological and physiological states, to recognize when people are in a stressed state, and to understand the context in which a speaker is communicating.

1.1.3 Applications

Along with the deepening of the theoretical study of the stressed speech classification, this technology also gradually and successfully achieves many applications.

- Speech recognition systems: The performance of automatic speech recognition (ASR) systems degrade in stressed environments, such as with background noise, while multitasking, when suffering from fatigue, under the influence of emotions, due to physical environmental factors, and as a result of high workload stress. The study of stressed speech will improve the robustness of speech recognition systems [15].
- Intelligent in-car spoken dialogue systems: Studies have been performed with the aim of improving driver safety using the collected data via speech interfaces while driving. Monitoring the driver's mental state can improve driving safety, by reducing the occurrence of accidents. Warning information can be provided to the driver automatically when the driver is operating the car in a negative mental state [16].
- Piloting aircraft: When flying, stress is produced by the variation in gravity due to taking off, landing and diving [17] [18]. In these situations, speakers become nervous, and profound changes in

the shape of their vocal organ can occur due to the increase in gravity (G-force) [18].

- Military: Military operations are often carried out under stressed conditions, such as high workload pressure, sleeplessness, fear and battle stress. These stress factors affect the quality of phonation, and can impede the operation of communication equipment, the performance of weapons, and the interface of command and control systems [19] [20].

- Telecommunication: Stress classification can be used to improve the performance of telephone-based speech recognition systems. It can also be applied to routing emergency calls in high priority situations, and to evaluating the mental state of a caller in order to provide suitable telephone response service [21].

- Psychiatry: Stress classification can be applied in psychiatry to help evaluate the quantitative objectivity of patient assessments [22].

- Law: Stress classification is employed in the forensic analysis of speech to assess the mental state of telephone callers, and can be utilized during interviews with suspects [23].

- Prevention of remittance fraud: The judgment of human beings is negatively affected by stress, such as when they receive undesirable information (so called “overtrust”). When in a state of overtrust, people can easily be deceived, as occurs in phone phishing scams. Therefore, stress classification can be used to detect stress in the victim’s speech in order to prevent remittance fraud [24].

1.2 LITERATURE REVIEW

1.2.1 STUDIES OF SPEECH UNDER STRESS

- Previous works

In 1911, E. Lombard found that speakers strive to adjust their pronunciation to improve the articulation of speech produced in noisy environments, which is now known as the Lombard effect. The further studies were performed to analyze the physiological and psychological impacts of stress on speech. The U.S. Navy Aviation Research Center conducted an analysis of pilots' tracking capability, blood reflection, psychological reactions, and endurance. Systematic study of stressed speech began in the late 1970s. The U.S. Air Force carried out studies of speech recognition in flight at the Wright Aeronautical Laboratory, and at the Armstrong Aviation Medicine Research Center. Moreover, the Armstrong Aviation Medicine Research Center cooperated with the Lincoln Laboratory at MIT on a robust speech recognition project, sponsored by the U.S. Department of Defense, Advanced Research Projects Agency. In addition, BNN Laboratories, Texas Instruments, and Central Methodist University (CMU) also launched research to develop a robust speech recognition system.

The initial studies mainly focused on the intelligibility of speech under stress, stressed speech analysis, and the impact of stressed speech on the performance of speech recognition systems. Studies found that stress reduce the intelligibility of speech [25]. It was also reported that hoods greatly influenced the performance of speech recognition systems under G-force conditions [26], and that oxygen supply systems and respiration noise were the main factors which affected system performance [27]. Through the analysis of stressed speech, the studies found that speech duration, intensity, and glottal parameters can be indicators of the presence of stress [28].

The types of stressed speech being studied have gradually expanded and more in-depth study has been done in this field. Early studies mainly concerned the Lombard effect under the condition of noisy backgrounds, but current work has begun to focus on speaking styles, workload stress and the effects of G-forces.

- Data collection

The collection of stressed speech samples is difficult, so some researchers have proposed synthesizing stressed speech using neutral speech. This method can increase the amount of available training data, improve stressed speech databases, and greatly advance stress classification studies. Research into emotional speech is currently being conducted at the ATR's Media Integration and Communication (MIC) Laboratory [29] and at Keio University [30] in Japan. Furthermore, some scholars have begun using multi-channel signals, including images and speech, to identify emotions [31].

Standard databases of stressed speech have also been established. The early corpus for stressed speech was called the "Simulated stress" speech database, collected by Texas Instruments [32], which included speakers who were asked to phonate using different simulated styles, including speaking quickly, softly, loudly, shouting, and using Lombard speech. MIT's Lincoln Laboratory established a database with 11 speaking styles and 35 words, which included neutral speech and stressed speech. The types of stressed speech included were angry, soft, fast, slow, clear, questioning, 50% workload, 70% workload and Lombard speech. Another well-known database, called SUSAS (Speech Under Simulated and Actual Stress), has been established by Duke University [11], which collected speech under simulated stress using the actual G-force stress and talking styles of amusement park roller-coaster riders. In 1994, a project called "Stressed Speech" was launched by

the Speech and Language Technology research group of NATO. A common database was established and shared internationally. The common database included the SUSAS database, the DLP database of speech under emergency conditions, and the SUSC-0/1 database collected from warplane communications. The DCIEM database developed in Canada by the University of Edinburgh [33] mainly includes stressed speech collected under adverse conditions, such as while suffering from sleep deprivation and while under the effect of drugs.

1.2.2 ANALYSIS OF STRESSED SPEECH

The presence of stress can affect both the glottal source and the modulating effect of the vocal tract. Based on the speech production model, therefore, the analysis of stressed speech can be considered from two aspects: (1) impacts on the glottal source, the typical parameters of which are glottal flow, fundamental frequency, and duration of speech; and (2) impacts on the shape of the vocal tract, the typical parameters being cross-sectional areas, spectral coefficients, formant width and position, and low-order cepstrum coefficients. Fundamental frequency (F0) and formants are the most commonly studied parameters, as they are considered to be good indicators of the presence of stress.

The first investigations of emotional speech were conducted by Van Bezooijen [34] and Scherer [35] in the mid-1980s, using the statistical properties of acoustic features to recognize emotions from speech. Cummings analyzed the impact of stress on the glottal source [36], and discovered dramatic changes in parameters representing the glottal ascending slope, descending slope, closure period, closing time, and opening time of glottis. Williams and Stevens found that fundamental frequency (F0) has different characteristics for each emotion [37]. It is believed that the value for fundamental frequency changes rapidly when the speaker is under stress, while pitch contour becomes smoother

during neutral speech. Based on the further experimental results, emotions have also been found to have different influences on fundamental frequency [38], e.g., F0 becomes lower, with a smoother contour, when the speaker is sad, while the variation range of F0 becomes larger when a speaker is angry. In 1983, Streeter et al. conducted the same experiments under flight conditions as Williams and Stevens [39]. Comparing their results with those of Williams, they did not find a consistent rise in fundamental frequency when speakers were under stress, which agreed with the results of Hecker under task-induced conditions [40]. Griffin analyzed three parameters, F0, amplitude peaks, and duration of words [41], and found that F0 and amplitude peaks increased under stress, but that word duration decreased when stress existed. Five different types of pronunciation, loud, normal, whispering, fast, and slow, were studied by Hansen [32], who found that a rise in F0 can be caused by an increase in speech loudness, but that there was less relation with variation in speaking rate. In another study, Pisoni [42] found that amplitude, duration, and F0 change significantly due to the Lombard effect, and that the spectral energy of consonants can be shifted to higher frequency bands.

Formants are the other important parameter representing the presence of stress. In 1994, Hansen studied the distribution of formants and found that formants of vowels are shifted to higher frequencies, and that the average width of formants decreased due to the Lombard effect [43]. For most phonemes, the Lombard effect results in an increase of amplitude of formants, a backward-shift in the first formant, and an increase of spectral slope. Hansen and Bou-Ghazale [44] showed how loud speech is similar to Lombard speech, and noted that variation in the duration of sounds, especially vowels, is significant. Hansen also analyzed variations in the shape of the vocal tract, cross-sectional area of the vocal tract, and Mel autocorrelation coefficients [45]. Results showed that changes in the shape of the vocal tract occur predominantly in the throat under neutral condition, while these changes occur on the edge and back of the tongue or around the lips when a speaker is

angry, which proved that the modulating effect of the vocal tract is greatly impacted by stress. The study also showed that there are significant changes in the cross-sectional areas of the vocal tract under both neutral and stressed conditions, which leads to variation in the acoustic parameters. Mel autocorrelation coefficients are also affected by stress, which can be observed by comparing neutral and stressed speech. By considering the dynamic characteristics, we see that variation in the coefficients becomes relatively slower under stress.

Analysis has also been performed using different stress factors. The influence of the Lombard effect on speech recognition was examined in [46][47][48] and [49]. Selected acoustic features were analyzed, such as amplitude and the distribution of spectral energy, and results showed that spectral energy shifted to higher frequencies for consonants in the presence of loud background noise. High workload stress would induce physical changes in the speech production. And also result in changes to the dimensions of the vocal tract [50]. Hecker showed that speaker will modify their speech under task loading although this effect varied considerably across individuals [51]. Workload stress has been proven to have a significant impact on the performance of speech recognition systems [52]. A number of studies have shown that, when a task is imposed on speakers, speech under workload becomes faster, softer, or louder than neutral speech [53][54][55][56][57][58]. Matsuo et al. examined the frequency domain and discovered that differences in the spectrum of the high frequency band under stressful workload conditions could be used to catch people committing remittance fraud, and their proposed measure achieved better stressed speech classification performance [59]. High G-force was studied in [60], showing that the first and second formants shifted to the higher frequency in the spectrum. In regards to emotional speech, studies have shown that emotional arousal results in variations in respiration patterns and muscle tension [38][47][61].

1.2.3 METHODS OF CLASSIFYING STRESSED SPEECH

1.2.3.1 Classification based on feature parameters

The parameters obtained from speech analysis can change when under stressed condition. So the basic idea for the classification of neutral and stressed speech is using the feature parameters. Classification methods can be discussed in terms of the feature parameters selected.

- Classification based on linear features

An early but still prominent physical model is the source-filter model [62] which models speech as a combination of a glottal source, such as the vocal folds, and a linear acoustic filter, representing the vocal tract (and its radiation characteristics). An essential assumption that is made with the source-filter model is the independence of the source and the filter. This model could more precisely be referred to as the “independent source-filter model”.

In 1961, Wong proposed a linear model of speech production using the lossless tube model of the vocal tract [63]. In 1979, a linear source tract model was proposed to model the glottal source, the vocal tract, and radiation impedance as linear filters, using covariance analysis [64]. However the vocal tract and vocal folds do not function independently of each other, instead there appears to be some form of interaction between them [65], which results in significant changes in fundamental frequency and formant characteristics.

The feature parameters extracted for stress classification using the traditional linear speech production model include F0, formant, energy distribution, duration of words or phonemes, and glottal excitation. The robustness of the extraction method is essential for the estimation of F0 and formants. Speaking rate of vowels is usually applied to represent the length of words or phonemes. Intensity and energy distribution are expressed by the square root of the average amplitude of the acoustic signal. Spectral slope is used for glottal characteristics and the position of first and second formants are proposed to characterize the spectrum of the vocal tract. Results suggest that F0 is the best indicator of stress, and that duration and formant can work well when they are combined with other features.

- Classification based on nonlinear features

In 1980, Teager proposed a nonlinear model, which suggested that speech production results from vortex-flow interaction [66][67]. In this study, it was assumed that airflow is unstable when it passes the wall of vocal tract. Airflow separation occurs around the region between true and false vocal folds, and concomitant vertices are generated. Teager believed that such vortex-flow interaction was a source of speech production, while also providing a modulating effect on raw speech signals; thus, it can be considered as a nonlinear component of speech production. According to this theory, speech signals include two parts; a linear part (the plane wave) and a nonlinear part (from the vortices). The Teager energy operator (TEO) was developed to reflect the instantaneous energy of nonlinear vortex-flow interaction [68]

$$\begin{aligned}\psi[x(t)] &= \left(\frac{d}{dt}x(t)\right)^2 - x(t)\left(\frac{d^2}{dt^2}x(t)\right) \\ &= [\dot{x}(t)]^2 - x(t)\ddot{x}(t),\end{aligned} \tag{1.1}$$

where $\psi[\cdot]$ is the Teager energy operator (TEO). Kaiser introduced the discrete expression,

$$\psi[x(n)] = x^2(n) - x(n+1)x(n-1), \tag{1.2}$$

Vortex-flow interaction is considered to be a source of speech production, and has a modulating effect on the speech that is finally generated. Speech signals can be divided into AM and FM components within a certain frequency band using the TEO profile [69]

$$\begin{aligned} f(n) &\approx \frac{1}{2\pi T}\arccos\left(1 - \frac{\psi[y(n)] + \psi[y(n+1)]}{4\psi[x(n)]}\right) \\ |a(n)| &\approx \sqrt{\frac{\psi[x(n)]}{\left[1 - \left(1 - \frac{\psi[y(n)] + \psi[y(n+1)]}{4\psi[x(n)]}\right)^2\right]}}, \end{aligned} \tag{1.3}$$

where $y(n) = x(n) - x(n-1)$. Kaiser proposed that speech signals can be regarded as the result of both frequency modulation and amplitude modulation within a certain carrier frequency.

On the basis of the nonlinear model, increasing emphasis should be focused on the nonlinear features of speech. Some nonlinear features have been proposed to detect stress. TEO-FM-Var is a feature representing TEO-decomposed FM variation [70][71], TEO-Auto-Env represents the instantaneous variation in glottal excitation based on the normalized TEO Autocorrelation Envelope Area [70][71][72], and TEO-CB-Auto-Env captures true stress sensitivity changes outside the first formant using critical band based frequency partitions [70][71][72].

Results have shown that the stress classification performance of TEO-FM-Var and TEO-Auto-Env is degraded due to their dependence on pitch estimation. However, TEO-CB-Auto-Env is an effective stress classification technique which is both accurate and reliable, and can maintain its performance under text independent conditions.

- Classification based on MFCC

Womack and Hansen proposed a stress classification method based on neural networks and an array of features. Parameters representing the cross-sectional area of the vocal tract, based on Mel-cepstrum, were mainly considered [73], including Mel (MFCC), Delta-Mel, Delta-Delta-Mel, and the new features Auto-Correlation Mel (AC-Mel), and Cross-Correlation Mel (XC-Mel). AC-Mel represents the difference in relative energy between the bands, and the relative variation in spectral slope between frames, while XC-Mel provides a quantity measurement for the relative variation in different scales between spectrums.

- Classification based on sub-bands

The energy of the spectrum of speech shifts when the speaker is under stress. The auditory system can detect stress because the human ear is sensitive to low frequency bands in the spectrum, which is where energy will be concentrated when speech is produced under loud and Lombard conditions. The initial studies of band partition were performed using sub-band energy from Mel filter banks for stress classification [74]. These methods were developed using wavelet packets to build the filter bank, similar to critical band based frequency partitioning, and some new features were also proposed. Erzin proposed using coefficients for sub-band energy analysis [75], and Sarikaya focused on the auto-correlation coefficient and cepstral parameters of sub-band energy, based on wavelet packets [76][77]. In 2000, Hansen explored a new parameter using the wavelet transform for the logarithm of sub-band energy [78], which was then applied to stress classification.

1.2.3.2 Classifiers

Hidden Markov Models (HMM), Artificial Neutral Networks (ANN), Gaussian Mixture Models (GMM) and Bayesian classifiers are widely applied for stress classification. Busch used a Bayesian classifier to separate neutral, loud, angry, and Lombard speech [79]. Hansen investigated features based on MFCC, and stress classification was performed using separability distance metrics and a neutral network classifier [73]. In another study, Hansen proposed a combining algorithm, and an N-dimensional Hidden Markov Model (HMM) was applied for stress classification [80]. Results showed that these stress classification techniques can improve the robustness of speech recognition systems.

Bou-Ghazale and Hansen developed perturbation models of neutral-to-stressed speech using an HMM framework [81]. Their model can simulate different speaking styles by perturbing the neutral training data. Their results showed that this method can improve classification performance. Additionally, a standard method for re-training reference models has been proposed to improve the robustness of speech recognition systems [82]. In 1987, an extended method, called a multi-style training model, was used, however evaluation has shown that multi-style training does not work well under the speaker-independent condition due to limited training data [83]. Sanjay used a reliable GMM framework to model the TEO feature TEO-CB-Auto-Env for stress classification. Using this method, the size of the training data and the number of mixtures are determined to achieve the best classification performance [84].

1.3 RESEARCH MOTIVATION

1.3.1 SPEECH PRODUCTION

Speech production refers to the airflow from the lungs through the vocal folds in the larynx and up the vocal tract, which is then shaped by the tongue, mouth, oral cavity, palate, nasal cavity and other articulators. Thus, speech production is a process for the generation of glottal excitation in the larynx, combined with a filtering process for modulation in the vocal tract. The source can be periodic, resulting in voiced sound, or aperiodic, resulting in unvoiced sounds or silence.

The lungs generate sub-glottal pressure in the trachea below the glottis, and this pressure is aided by contraction of the laryngeal musculature, driving the air to propagate through the larynx. The vocal folds vibrate as a result of this airflow, generating periodic pulses of air, thus producing voiced sounds. The periodic pulses of air generated by the vocal folds pass through the other essential component in human speech production system, the vocal tract, which is a tube-shaped passage above the larynx made up of muscles and tissues. Variations in the length and shape of the vocal tract are mainly the result of the movement of the articulators, such as the jaw, the tongue and the mouth. The glottal source is filtered and modulated by changes in the vocal tract, and a radiation effect is produced at the lips to produce articulation in final speech. Detailed surveys of the speech production process are described in [85][86][87][88].

1.3.2 MATHEMATICAL BASIS OF MODELING SPEECH PRODUCTION

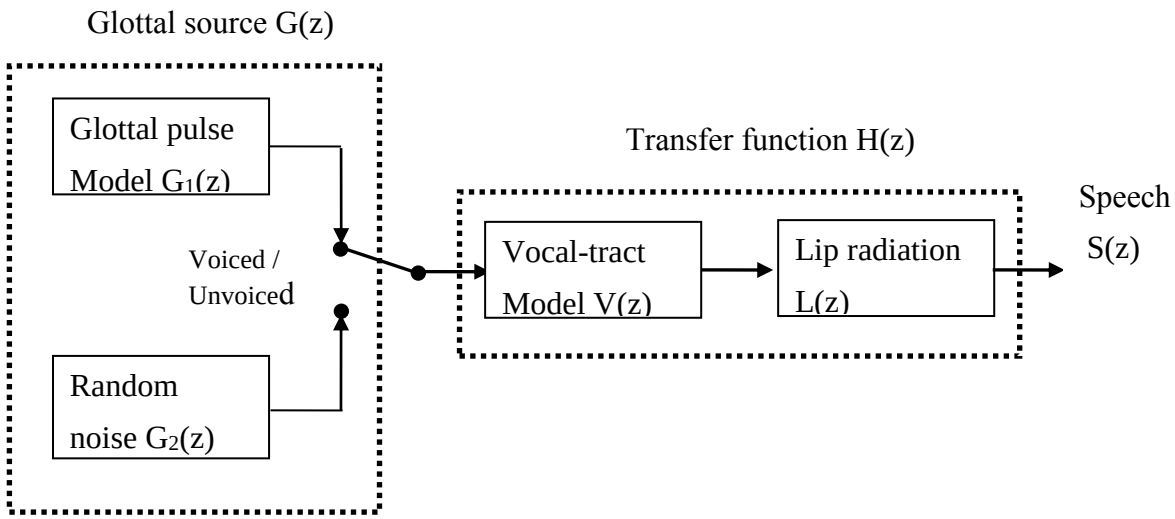


Figure 1.1 Linear source-filter model of speech production

In order to simulate the process of speech production, the traditional linear model is the source-filter model [62], which models speech as a combination of a glottal source from the vocal folds, and a linear acoustic filter representing the vocal tract (as well as radiation characteristics).

Figure 1.1 describes the general structure of the linear speech production model for the production of voiced and unvoiced speech. The glottal source is modeled by either the glottal pulse model G(z) = G_1(z), simulating the vibration of the vocal folds when generating an impulse train in voiced sound, or by a random noise generator G(z) = G_2(z), when producing turbulent airflow for unvoiced sound, yielding an excitation signal representing the volume velocity of the signal at the outlet of the glottis (glottal flow). The effect of the vocal tract is modeled by the transfer function V(z) to modulate the glottal flow. Finally, the radiation effect of lips is represented by L(z).

According to the theory of linear systems, the transfer function H(z) can be calculated,

$$H(\mathrm{z}) = \mathrm{V(z)L(z)}, \tag{1.4}$$

where V(z) is the vocal tract transfer function characterizing articulation.

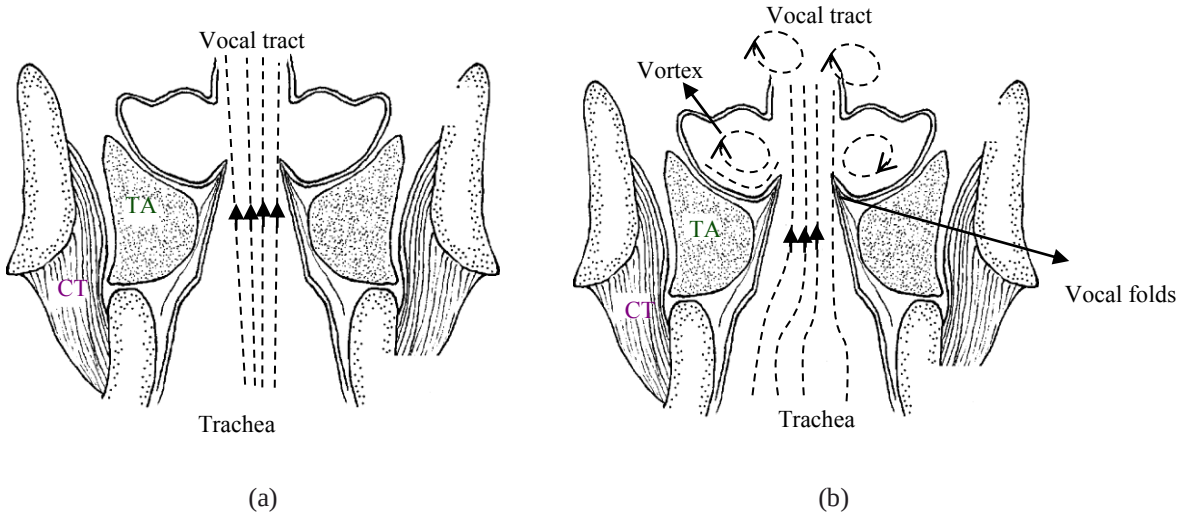


Figure 1.2 airflow propagation for speech production in the vocal system (a) Classical interpretation (b) A new aerodynamic interpretation.

The linear model is limited to simulating the vocal tract. An essential assumption of the source-filter model is that the source and filter function independently of each other. The model could more precisely be referred to as the "independent source-filter model". This independence requires us to qualify and measure the excitation source separately from the vocal tract filter. From the physiological point of view, it is assumed that airflow from the lungs always propagates as a linear plane wave in the glottis and the vocal tract, and that the pulsatile flow is the only source of speech production, as shown in Figure 1.2 (a). However, there is increasing evidence suggesting that this hypothesis may not be valid. The vocal tract and vocal folds do not function independently of each other; there is interaction between them [65], which results in significant changes in fundamental frequency and formant characteristics..

Teager provided a new theory of airflow propagation for speech production [56][57][79]. According to this theory, it is believed that the airflow does not always propagate as a plane wave in the glottis and the vocal tract. This is because the airflow coming from glottis is very unstable as it passes the

wall of vocal tract. It can attach to the wall, and then separate, and then reattach to the wall again, which can change the effective area of the vocal tract. Airflow separation occurs due to the Coanda effect [80], which causes airflow at high velocity to separate and partially attach to the wall of the ventricle between true and false vocal folds. When this occurs, viscous forces cause the airflow to "roll up" into a rotational flow structure, and airflow patterns are changed. Teager believed that variations in aerodynamics around the false vocal folds affected the process of speech production, and also provided a modulating effect on the raw speech signal.

It is believed that speakers change their physical vocal system, such as the muscle tension of the vocal folds or the shape of the vocal tract, when speaking under stressful conditions. These changes in vocal system physiology result in variations in the airflow patterns in the glottis and the vocal tract, thus the presence of stress can result in variations in airflow characteristics [61]. Therefore knowledge of the aerodynamics of speech production and the physiological characteristics of the vocal system are essential for understanding the process of stressed speech generation.

There are two methods currently used to model the process of speech production. One approach is utilization of the source-filter model. This approach assumes that the movement of the vocal folds and the shape of the vocal tract during speech production can be modeled separately. However, this approach neglects the complicated interaction between the vocal folds and the vocal tract which occurs during speech production. The other method is to model vocal airflow in order to characterize speech production. Modeling the airflow patterns mathematically may allow us to more accurately explain the process of speech production, and the aerodynamics and acoustic interaction of speech could then be modeled to characterize aerodynamics in the glottis and the vocal tract. Therefore, a physical model, like two-mass model, is necessary to model the airflow patterns in the physiological system, in order to represent the process of speech production.

Cairns suggested that variations in vocal aerodynamics resulting from airflow separation differ markedly between neutral and stressed speech [81]. Since stress is subjective, speakers under stressed conditions will have different physical reactions and will change their physical systems to adapt themselves to the stressful event or situation. In physiological systems, changes in physical characteristics such as muscle tension will significantly affect airflow patterns, which has an impact on acoustic interaction, resulting in the production of stressed speech. Measurements of the fluid flow of air within the human glottis reveal that aerodynamics are directly affected by changes in the physiological system [82]. Therefore, the two-mass model is a better alternative for representing the aerodynamics of the physiological system of speech production if we wish to comprehend the generation of stressed speech.

1.4 RESEARCH OBJECTIVE AND CONTRIBUTION

1.4.1 OBJECTIVE

The objective of this dissertation is the classification of speech under stress based on a physical model. We will mainly concentrate on the analysis of stressed speech based on a speech production model instead of on observed speech features, in order to gain a deeper comprehension of the working mechanisms of the vocal folds and the shape of the vocal tract. For this purpose, I will explore the underlying properties of the physical speech production model, and search for essential factors related to stress. Physiological characteristics are mainly examined to explain changes in speech production under stress. As a result of this study, an explanation can be made of how the physical model applies to real speech.

In this dissertation, we propose and develop a novel method for the classification of stressed speech by modeling physiological system and airflow patterns in the vocal folds and the vocal tract. It is believed that the presence of stress can result in variations in the physical characteristics of physiological systems and then cause changes in airflow patterns, thus affecting acoustic interaction between the vocal folds and the vocal tract [81]. Finally the stressed speech is produced. So, the parameters of a physical model characterizing airflow patterns and physical properties can represent the stress more directly and clearly than conventional methods.

The two-mass model is a physical model which attempts to simulate the physical process of speech production by characterizing the vocal folds and the vocal tract, and also by modeling the effect of glottis-vocal tract interaction [83]. We use the two-mass model as a physical model of speech production, and the physical parameters affected by stress are then estimated from real speech using this model. Analysis of estimated parameters is performed to explore which features are effective for the detection of speech under stress. Compared with acoustic parameters, physical parameters are more robust and precise at representing the presence of stress.

1.4.2 CONTRIBUTION OF THE DISSERTATION

In this dissertation, we propose a scheme for the classification of speech under stress from a new perspective based on a physical model. The consideration is focused on the glottal source, and then the characteristics of the vocal folds is explored. This method uses parameters estimated from a physical model to characterize the behavior of the vocal folds during speech. Therefore, we propose fitting a two-mass model to real speech in order to estimate the physical parameters which represent muscle tension in the vocal folds, vocal fold viscosity loss, and sub-glottal pressure coming from the

lungs. Furthermore, combinations of these physical parameters are proposed as features which are effective for the classification of speech as either neutral or stressed.

We develop our method by taking into consideration the characteristics of airflow patterns in the vocal folds and the vocal tract, as well as the interaction between them. Parameters derived from a physical model can characterize stressed speech more precisely because both the vocal folds and the vocal tract is considered. A two-mass model connected to a four-tube model is used to simulate the process of speech production. The physical parameters (stiffness, vocal tract length and cross-sectional areas of vocal tract) are estimated by fitting the model to real speech. Different cost functions are proposed to evaluate the estimated physical parameters in order to compare classification performance.

Furthermore, we extend our works by considering variations in the airflow patterns in the laryngeal ventricle and the false vocal folds for the classification of neutral and stressed speech. Traditional two-mass model is modified to include the laryngeal ventricle, and the physical parameters characterizing airflow variations in the laryngeal ventricle under psychological stress. The proposed model is fitted to real speech to estimate the physical parameters representing effective area of laryngeal ventricle. The estimated parameters representing airflow variation can be effective for separating stressed speech from neutral speech.

Finally, we present a comparison of the stress classification performance of different classifiers. Physical features are modeled using a linear classifier, a Gaussian Mixture Model (GMM), and a Support Vector machine (SVM) respectively. Gains should be achieved if the amount of training data is significantly increased. SVM is more advisable to propose for the improvement for the small sample size problem in stress classification.

1.5 OUTLINE OF THE DISSERTATION

In Chapter 2, we describe the stress database used in this dissertation. The psychological state of victims of phishing scams is introduced, and data collection for psychological suppression is described for different tasks and environments.

In Chapter 3, we give an introduction of the physical model. Two-mass model is applied to represent the process of speech production.

In Chapter 4, we propose a method of classifying speech under stress using parameters extracted from a physical model to characterize the behavior of the vocal folds. A method that fits a two-mass model to real speech is utilized in order to estimate the physical parameters that represent muscle tension in the vocal folds, vocal fold viscosity loss, and sub-glottal pressure coming from the lungs.

In Chapter 5, we develop the method for the consideration of both the vocal folds and the vocal tract, and also the interaction between them. Physical parameters, representing stiffness, vocal tract length and cross-sectional areas of vocal tract, are estimated with different cost function. The results for classification performance are compared.

In Chapter 6, we focus on variations in the aerodynamics in the laryngeal ventricle and the false vocal folds. Two-mass model is modified to include the laryngeal ventricle, and the physical parameters characterizing airflow variations in the laryngeal ventricle under psychological stress are explored.

In Chapter 7, we make a comparison of the stress classification performance of different classifiers. Physical features are modeled using a linear classifier, a Gaussian Mixture Model (GMM), and a Support Vector machine (SVM) respectively.

Finally, Chapter 8 contains our conclusion and suggestions for future work.

CHAPTER 2: DATA COLLECTION FOR SPEECH UNDER STRESS

2.1 PSYCHOLOGICAL STATE OF PHISHING SCAM VICTIMS

According to statistical data from the national police agency in Japan, there were 3,770 acts of remittance fraud by the end of July 2010, with economic losses of up to 256,500,000 Japanese yen (approx. 2.5 million U.S. dollars). In 2012 alone, 6,401 acts of remittance fraud were committed, with 16,200,000,000 yen (approx. 162 million U.S. dollars) in total economic losses. These crimes are increasing, and becoming a serious social problem, making it increasingly important to focus on stress detection and to develop methods to prevent it. One possible solution would be to detect phone phishing scams by using the victim's voice information. Phone calls could be monitored, and the features representing stress could be extracted in order to discriminate between neutral and stressed speech. This would allow automated detection of attempted wire fraud in real time.

It is essential to have access to speech data in order to conduct stress classification studies, so experiments were designed which evoked suppression stress similar to that of phishing scam victims, and the resulting data was then used in this study.

2.1.1 TECHNIQUES USED BY CRIMINALS FOR REMITTANCE FRAUD

Remittance fraud is a criminal act which occurs when a victim is requested to transfer money at the request of someone misrepresenting themselves as part of a phone phishing scam. There are four techniques used by criminals to commit remittance fraud, according to reports from Japan's national police agency:

- The "It is me, it is me!" scam.

The perpetrator pretends to be the victim's relative or friend, claims to be in trouble, and asks the victim for a money transfer.

- False billing

The perpetrator demands payment for false claims or non-existent bills.

- Advance-fee load scam

The perpetrator offers to loan the victim money, and asks for a loan processing fee, but the victim never receives the loan.

- Refund scam

The perpetrator requests the payment of a service charge in order to process a refund which the victim is told they are entitled to.

Figure 2.1 shows how detection technology could be used to prevent remittance –solicitation phone scams, using the example of the "It is me, it is me!" scam. The victim is tricked into making a remittance using a credit card or ATM as the result of being given false information by phone. The stressed state of the victim is detected during the phone phishing scam, and a warning is sent to

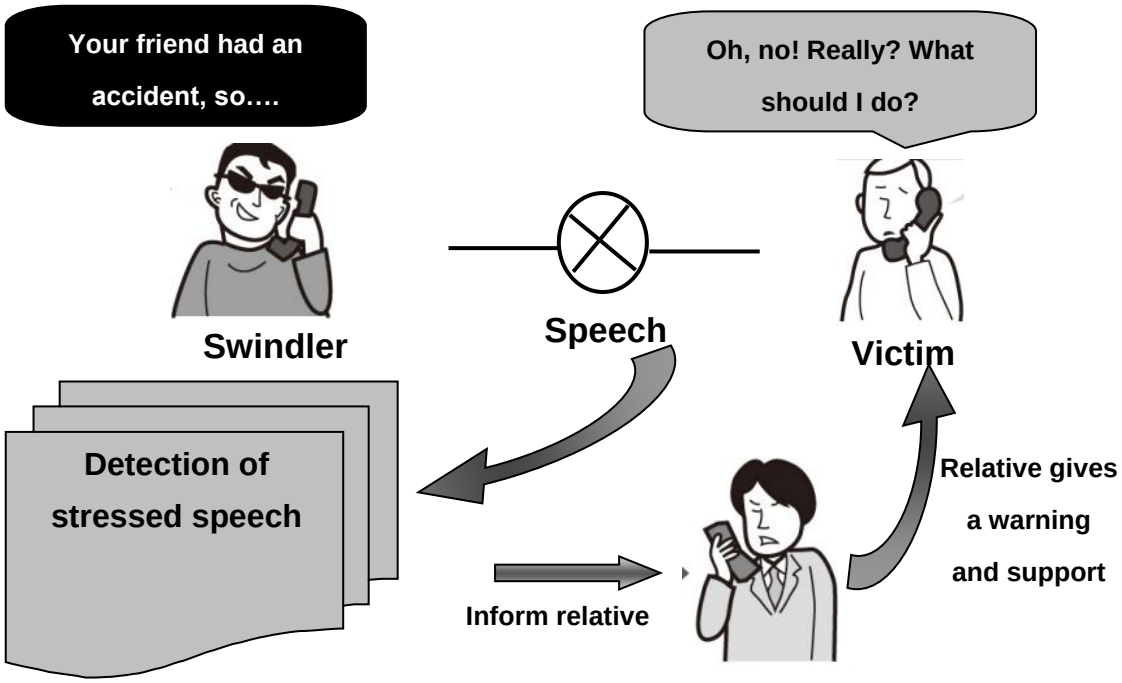


Figure 2.1 Detection of remittance –solicitation phone scams

inform a relative or friend of the victim, who then contacts the victim in order to prevent them from being victimized.

2.1.2 PSYCHOLOGICAL SUPPRESSION

An abnormal mental state is evoked when victims suffer from psychological suppression during phone phishing scams. There are three mental states which occur due to the suppression experienced as a result of the techniques used by criminals to commit remittance fraud:

(1) The state in which victims are distracted by the false situation described in the phone phishing scam (some emergency or mistake they are accused of having made), which causes the victim to lose their good judgment. When in this state, the victims are unable to focus on what is happening.

(2) The state in which victims are being hurried to perform tasks under time pressure. When in this state, the victims cannot make appropriate decisions due to the anxiety caused by the limited amount of time they have.

(3) The state in which the victim's desire for monetary gain is evoked, giving them the hope of obtaining money easily. When in this state, the victims are not able to make appropriate decisions due to their strong desire for money, a technique which is usually used in advance-fee load scams and refund scams.

2.2 COLLECTION OF PSYCHOLOGICAL SUPPRESSION SPEECH

2.2.1 WORKLOAD TASKS

In this section, three tasks corresponding to different mental states were introduced in order to simulate mental pressure resulting in psychological stress. These tasks were performed by a speaker having a telephone conversation with an operator, to simulate a situation involving pressure during telephone communication.

- Concentration

The speaker is asked to finish tasks including solving logic puzzles and spotting the differences between two pictures.

A logic puzzle task is shown in Figure 2.2. During the telephone calls, the speaker is required to give logical answers to the puzzles using the given hints, and to explain his or her reasoning process.

An example of the task involving spotting the differences between two pictures is shown in Figure 2.3. Detailed information about the differences is requested.

In the experiment, the operator first explains the whole procedure in detail. The operator then provides some hints if the speaker has trouble solving the problems. The speaker is given problems of increasing difficulty, and cannot abandon a problem until the whole process is completed.

- Time pressure

The speaker is asked to answer questions under time pressure, as shown in Figure 2.4. In the experiment, the operator asks the speaker to find the differences between two pictures, and some questions are posed at the same time. Elapsed time is shown on a display to generate time pressure on the speaker, who must answer the questions in a limited amount of time. The questions disappear when time is running out. If there are problems which the speaker failed to solve within the given time limit, the speaker is asked to answer them again.

- Speculative spirit (Desire for monetary gain)

Gambling games are played to evaluate the speaker's desire for monetary gain. A deck of playing cards is used for gambling, and the speaker's objective is to win a targeted amount of money. The speaker must borrow money from the operator by phone if the speaker loses all his or her money, in order to keep playing the game, or if the speaker would like to wager more money. According to the rules of the game, it is easy to win if the speaker borrows more money from the operator to play the game. Speech data was collected from the conversations between the speaker and the operator.

Q. Fill in the table according to the following hints:

Hint 1: Which sport does Kaaspar play?

Hint 2: Football is located to the right of baseball.

Hint 3: Mark does not like football

Location	Left	Right
Person		
Sport		

Figure 2.2 Logic puzzle

2.2.2 DATA COLLECTION PROCEDURE AND ENVIRONMENT

A database collected by the Fujitsu Corporation, containing speech samples from telephone conversations with subjects performing different tasks, was also used. To simulate mental pressure resulting in psychological stress, three different tasks were introduced, which were performed by the speakers while having conversations with an operator. For each speaker, there are four dialogues with different tasks. The first dialogue involves relaxed chat without any task, followed by dialogues 2 and 3, in which the speaker is asked to finish tasks, generating speech under workload conditions. Finally, in dialogue 4 there is also relaxed chat without any tasks involved. Evaluations were then performed using the data collected with workload.

During the relaxed chat, the speech data collected was small talk, without the speaker experiencing any mental suppression. During the chat, the speaker discussed easy topics, and other topics were provided if the speaker felt uncomfortable answering questions. Furthermore, quiet and elegant

Figure 2.3 Spot the differences

surroundings were provided, and air-conditioning was used to make the speaker feel relaxed by the comfortable temperature and low humidity.

This database contains speech samples from 100 people, including 50 males and 50 females. Data from 11 speakers was chosen by subjective assessment for use in the experiments. Speakers were selected whose tone of voice had obviously changed under the pressure of the different tasks, as compared to normal speech. During the subjective evaluations, it was assumed that stressed speech would be generated under workload conditions, and that neutral speech would be collected during the relaxed chat, without any suppression. The speech of these 11 speakers was selected for further analysis and evaluation.

Answer the following questions in 7 minutes.

(1) What is meaning of the proverb “Failing to plan is planning to fail”?

(2) How to use the Internet?

(3) Please tell me about your fantasy lover

(4) Please tell me about your friend. What are some things he likes to do and doesn’t like to do?

(5) What is the most important thing to remember when making travel plans?

(6) Please tell me the most impressive news you’ve heard recently.

(7) Do you know where the telephone was invented?

Figure 2.4 Answer the questions with time pressure

CHAPTER 3: PHYSICAL MODEL FOR SPEECH PRODUCTION

The process of speech production is essential for stress classification, so a physical model is helpful to model airflow patterns in order to characterize speech production. Here, a two-mass model is used to represent the characteristics of the physiological system. In this chapter, we give an introduction about the two-mass model.

3.1 MECHANICAL RELATIONS

Two-mass vocal fold model was proposed by Ishizaka and Flanagan to simulate the process of speech production [93]. Figure 3.1 shows a sketch of the model. Each vocal fold is represented by a mass-spring-damper system, joined with a coupling stiffness, and is represented as

$$m_1 \frac{d^2 x_1}{dt^2} + r_1 \frac{dx_1}{dt} + s_1(x_1) + k_c(x_1 - x_2) = F_1 \tag{3.1}$$

$$m_2 \frac{d^2 x_2}{dt^2} + r_2 \frac{dx_2}{dt} + s_2(x_2) + k_c(x_2 - x_1) = F_2, \tag{3.2}$$

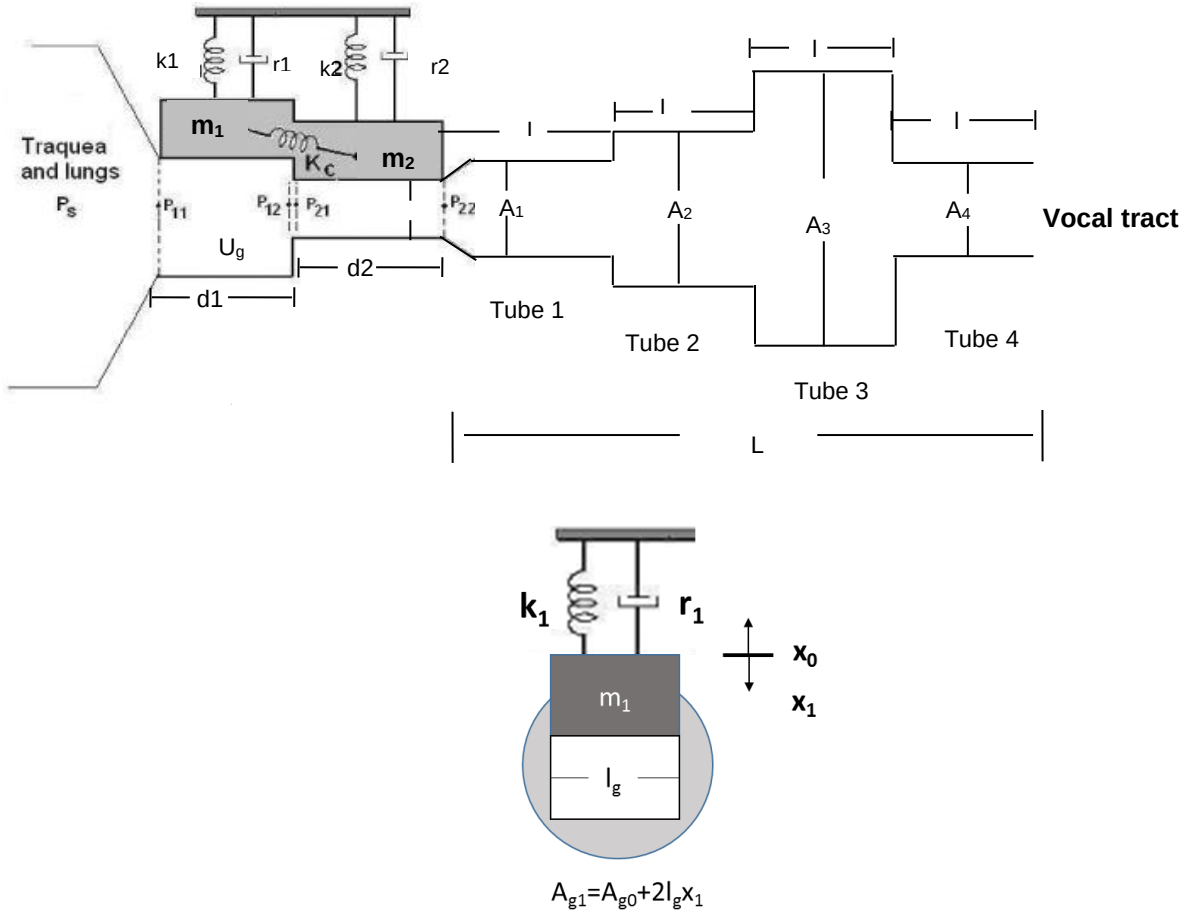


Figure. 3.1 The structure of two-mass model.

m_1, m_2 : the equivalent mass

d_1, d_2 : the thickness of m_1, m_2, respectively

U_g: the average volume velocity across the glottal area

L : the vocal tract length

k_1, k_2, k_c : the stiffness of the vocal folds

A_1, A_2, A_3, A_4 : the cross-sectional area for the vocal tract

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

$$s_i(x_i) = k_i(x_i + \eta x_i^{\,3}), \qquad i = 1, 2 \tag{3.3}$$

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

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

$$r_1 = 2\zeta_1\sqrt{m_1 k_1} \qquad r_2 = 2\zeta_2\sqrt{m_2 k_2}\,, \tag{3.4}$$

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

3.2 GLOTTAL AND VOCAL TRACT AERODYNAMINCS

Aerodynamics in the glottis is modeled with a set of equations proposed by Ishizaka and Flanagan [93]. 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:

$$P_s - P_{11} = (1.00 + 0.37)\frac{\rho U_g^2}{2A_{g1}^2}, \tag{3.5}$$

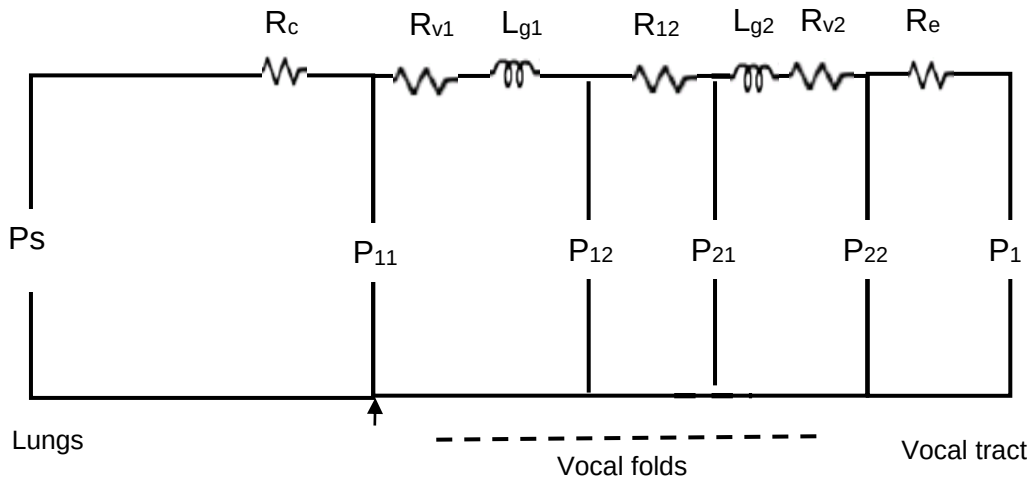


Figure 3.2 Equivalent circuit for the glottis

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:

$$P_{i1} - P_{i2} = \frac{12\mu d_i l_g^2 U_g}{A_{gi}^3}, \qquad i = 1,2 \tag{3.6}$$

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

$$P_{21} - P_{12} = \frac{\rho U_g^2}{2}\left(\frac{1}{A_{g1}^2} - \frac{1}{A_{g2}^2}\right), \tag{3.7}$$

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

$$P_1 - P_{22} = \frac{1}{2}\rho\frac{U_g^2}{A_{g2}^2}\left[2N(1-N)\right], \tag{3.8}$$

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.

Based on these pressure difference relations, an equivalent circuit is shown in Figure 3.2, with different acoustic impedance elements.

The elements of circuit are described as

$$R_c = 1.37\frac{\rho}{2}\frac{|U_g|}{A_{g1}^2}, \tag{3.9}$$

$$R_{v1} = 12\frac{\mu l_g^2 d_1}{A_{g1}^3}, \tag{3.10}$$

$$L_{g1} = \frac{\rho d_1}{A_{g1}}, \tag{3.11}$$

$$R_{12} = \frac{\rho}{2}\left(\frac{1}{A_{g2}^2} - \frac{1}{A_{g1}^2}\right)|U_g|, \tag{3.12}$$

$$R_{v2} = 12\frac{\mu l_g^2 d_2}{A_{g2}^3}, \tag{3.13}$$

$$L_{g2} = \frac{\rho d_2}{A_{g2}}, \tag{3.14}$$

$$R_e = -\frac{\rho}{2}\cdot\frac{2}{A_{g2}A_1}\left(1-\frac{A_{g2}}{A_1}\right)\left|U_g\right|. \tag{3.15}$$

The total acoustic impedance of the glottis is expressed as

$$Z_g = \frac{\rho}{2}\left|U_g\right|\left\{\frac{0.37}{A_{g1}^2}+\frac{1-2\frac{A_{g2}}{A_1}\left(1-\frac{A_{g2}}{A_1}\right)}{A_{g2}^2}\right\}+(R_{v1}+R_{v2})+j\omega(L_{g1}+L_{g2}+L_c), \tag{3.16}$$

or

$$Z_g = (R_{k1}+R_{k2})\left|U_g\right|+(R_{v1}+R_{v2})+j\omega(L_{g1}+L_{g2}+L_c), \tag{3.17}$$

where

$$R_{k1} = \frac{0.19\rho}{A_{g1}^2} \quad R_{k2} = \frac{\rho\left[0.5-\frac{A_{g2}}{A_1}\left(1-\frac{A_{g2}}{A_1}\right)\right]}{A_{g2}^2}, \tag{3.18}$$

In general, L_c is neglected compared to $(L_{g1}+L_{g2})$.

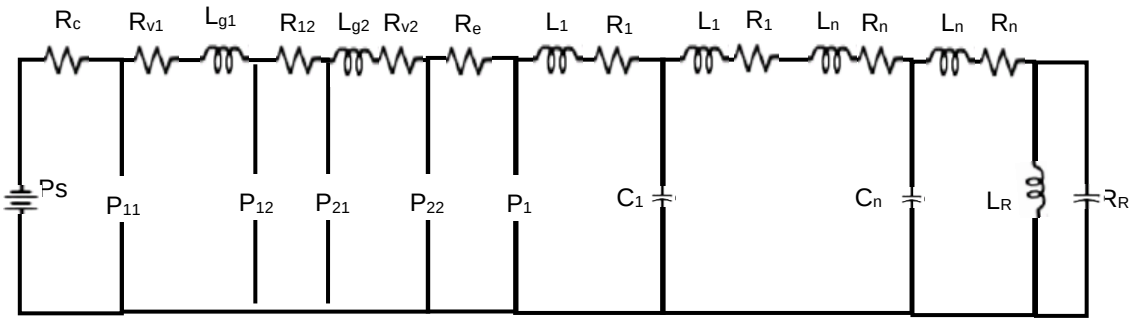


Figure 3.3 Network model for simulation of voiced sound

3.3 MODEL SYSTEM

The network model for simulating the voiced sounds is represented in Figure 3.3. In this network, subglottal system including trachea, bronchi, and lungs is neglected, and pressure from lung is approximated by subglottal pressure.

The two-mass model is connected to a four tube model representing the vocal tract [19]. The tube model is constructed using a transmission line analogy involving n cylindrical, hard-walled sections. The elemental values of the model are determined by cross-sectional areas $A_1 \cdots A_n$, and cylinder lengths $l_1 \cdots l_n$. The total length of the vocal tract is defined as L_{VT}. The tube model can be represented by an equivalent circuit, as shown in Figure 3.3. The inductances are $L_i = \rho l_i / 2A_i \quad (i = 1,2,\ldots,n-1)$, the capacitances are $C_i = l_i \cdot A_i / \rho c^2$, and the resistances $R_i = (S_i / A_i^2)\sqrt{\rho\mu\omega/2}$ where c is the velocity of sound. Here, the tube model has been limited to four cylindrical sections of equal length, $n = 4$. In this study, the model is limited to only vowel articulation (as vowels were the subject of the experiments) and modal voice production. These assumptions greatly simplify the modeling of the vocal tract and the glottal source. The model is

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

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

$$
\begin{aligned}
&(R_{k1} + R_{k2})|U_g|U_g + (R_{v1} + R_{v2})U_g + (L_{g1} + L_{g2})\frac{dU_g}{dt} + \\
&L_1 \frac{dU_g}{dt} + R_1 U_g + \frac{1}{C_1}\int_0^t (U_g - U_1)dt - P_s = 0, \\
&(L_1 + L_2)\frac{dU_1}{dt} + (R_1 + R_2)U_1 + \\
&\frac{1}{C_2}\int_0^t (U_1 - U_2)dt + \frac{1}{C_1}\int_0^t (U_1 - U_g)dt = 0, \\
&(L_2 + L_3)\frac{dU_2}{dt} + (R_2 + R_3)U_2 + \\
&\frac{1}{C_3}\int_0^t (U_2 - U_3)dt + \frac{1}{C_2}\int_0^t (U_2 - U_1)dt = 0, \\
&(L_3 + L_4)\frac{dU_3}{dt} + (R_3 + R_4)U_3 + \\
&\frac{1}{C_4}\int_0^t (U_3 - U_L)dt + \frac{1}{C_3}\int_0^t (U_3 - U_2)dt = 0, \\
&(L_4 + L_R)\frac{dU_L}{dt} + R_4 U_L - L_R \frac{d(U_R)}{dt} \\
&\frac{1}{C_4}\int_0^t (U_L - U_3)dt = 0, \\
&L_R \frac{d}{dt}(U_R + U_L) + R_R \cdot U_R = 0,
\end{aligned}
\tag{3.19}
$$

The mean pressures acting on the masses of the vocal folds are expressed as

$$P_{m1} = \frac{1}{2}(P_{11} + P_{12}) = P_s - 1.37\frac{\rho}{2}\left(\frac{U_g}{A_{g1}}\right)^2 - \frac{1}{2}\left(R_{v1}U_g + L_{g1}\frac{dU_g}{dt}\right) \tag{3.20}$$

$$\begin{aligned} P_{m2} &= \frac{1}{2}(P_{21} + P_{22}) = P_{m1} - \frac{1}{2}\left\{(R_{v1} + R_{v2})U_g + (L_{g1} + L_{g2})\frac{dU_g}{dt}\right\} \\ &- \frac{\rho}{2}U_g^2\left(\frac{1}{A_{g2}^2} - \frac{1}{A_{g1}^2}\right), \end{aligned} \tag{3.21}$$

Finally, force F_i acting on the masses is calculated by the displacement and the length of the vocal folds. When the glottis is closed, forces are calculated by

$$\begin{array}{lllll} F_1 = d_1 l_g P_{m1} & F_2 = d_2 l_g P_{m2} & if & x_1 > x_{1\min} & or \quad x_2 > x_{2\min} \\ F_1 = d_1 l_g P_s & F_2 = 0 & if & x_1 \le x_{1\min} & or \quad x_2 > x_{2\min} \\ F_1 = d_1 l_g P_s & F_2 = d_2 l_g P_s & if & x_1 > x_{1\min} & or \quad x_2 \le x_{2\min} \\ F_1 = d_1 l_g P_s & F_2 = 0 & if & x_1 \le x_{1\min} & or \quad x_2 \le x_{2\min}\,. \end{array} \tag{3.22}$$

where $x_{1\min} = -(A_{g01}/2l_g)$, $x_{2\min} = -(A_{g02}/2l_g)$, and A_{g01} and A_{g02} are the neutral values of the glottal area. The cross-section areas of the glottis A_{g1}, A_{g2} are calculated from the displacement of the masses.

$$A_{g1} = (A_{g01} + 2l_g x_1),\ A_{g2} = (A_{g02} + 2l_g x_2), \tag{3.23}$$

3.4 MALE AND FEMALE CONFIGURATION

For the configuration of the two-mass model, the following values were adopted, using typical values for males: $m_{1M} = 1.25\times10^{-4}$ kg , $m_{2M} = 2.5\times10^{-5}$ kg , $l_{gM} = 0.014$ m , $d_{1M} = 0.0025$ m , $d_{2M} = 5\times10^{-4}$ m , $\zeta_{2M} = 0.6$, and $x_0 = 2\times10^{-4}$ m . The vocal tract model was represented by a standard tube configuration for the vowel /a/ [94], 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 $L_M = 0.18\,\mathrm{m}$, with each element set to $l_i = 0.045\,\mathrm{m}$, and the cross-sectional areas were $A_1 = 8\times10^{-5}\ \mathrm{m}^2$, $A_2 = 4\times10^{-5}\ \mathrm{m}^2$, $A_3 = 3\times10^{-4}\ \mathrm{m}^2$, and $A_4 = 8\times10^{-4}\ \mathrm{m}^2$. For the configuration for females, the typical values were as follows: $m_{1F} = 4.56\times10^{-5}\ \mathrm{kg}$, $m_{2F} = 9.1\times10^{-6}\ \mathrm{kg}$, $l_{gF} = 0.01\,\mathrm{m}$, $d_{1F} = 1.79\times10^{-3}\ \mathrm{m}$, $d_{2F} = 3.6\times10^{-4}\ \mathrm{m}$, $\zeta_{2F} = 0.6$, $x_0 = 2\times10^{-4}\ \mathrm{m}$. For the vocal tract model, the length of the vocal tract was set to $L_F = 0.14\,\mathrm{m}$, with each element $l_i = 0.035\,\mathrm{m}$, and the cross-sectional areas were $A_1 = 4.85\times10^{-5}\ \mathrm{m}^2$, $A_2 = 2.4\times10^{-5}\ \mathrm{m}^2$, $A_3 = 1.8\times10^{-4}\ \mathrm{m}^2$, and $A_4 = 4.85\times10^{-4}\ \mathrm{m}^2$.

In the experiments, the following ranges for the control parameters were used for all speakers: P_S : $200-1900\ \mathrm{Pa}$, k_1 : $10000-140000\ \mathrm{dyn/cm}$, k_2 : $2000-14000\ \mathrm{dyn/cm}$, k_c : $4000-45000\ \mathrm{dyn/cm}$, 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 $\mathrm{cmH_2O}$, but 10-12 $\mathrm{cmH_2O}$ for loud speech), so the greater search range is advisable for the search method.

CHAPTER 4: CLASSIFICATION OF SPEECH UNDER STRESS BASED ON MODELING OF THE VOCAL FOLDS

4.1 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.

Among the various approaches previously proposed for stress classification, features derived from a linear speech production models are currently under extensive research. However, the features examined in these previous studies lack a physical basis, and the methods do not consider the whole process of speech production, which is believed to be essential for effective classification of speech under stress.

We propose a new classification method, based on the working mechanisms of the vocal folds, for speech under stress using parameters estimated from the two-mass 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 chapter, 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. When a speaker is under stressed condition, muscle tension of the vocal folds will become higher, subglottal pressure from the lungs will lower and the vocal folds become more viscous than under neutral conditions. So 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.

This chapter is organized as follows. In Section 3.2, the theoretical basis for two-mass model is introduced. This is followed in Section 3.3 by the presentation of a fitting algorithm for real speech data to help estimate the physical parameters. In Section 3.4, 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 Section 3.5.

4.2 PHYSIOLOGICAL BASIS FOR THE VOCAL FOLDS

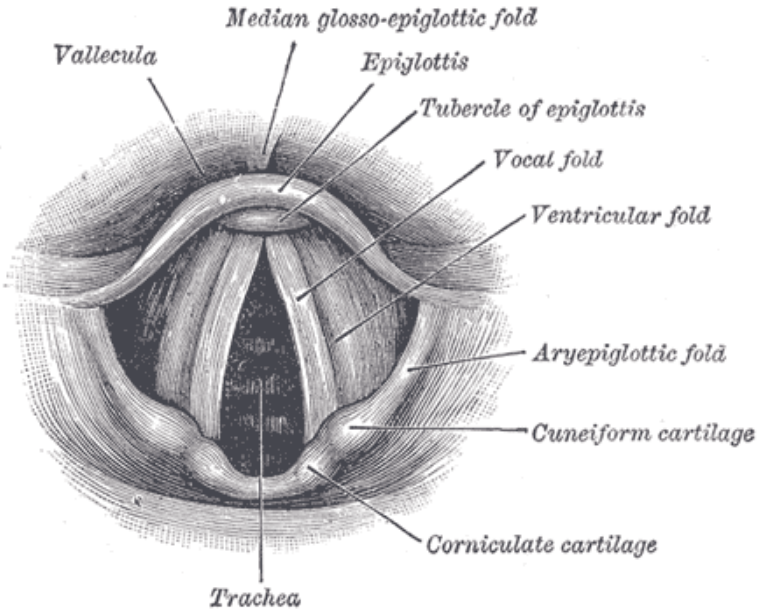


Figure. 4.1 View of the vocal folds [95]

The larynx is located in the anterior of the neck, which is formed by cartilages including epiglottis, thyroid, cricoid, arytenoids, comiculate, and cuneiform. It can extend from the epiglottis to the cricoid cartilage and separated into supraglottis, subglottis, and glottis.

The vocal folds consist of two infoldings of mucous membrane, located in the larynx above the trachea. Figure 4.1 shows the view of the vocal folds. They connect anterior to the thyroid cartilage and anterior to the arytenoids cartilages. The main tissue of the vocal folds is formed from epithelium with some muscle fibres inside (vocalis muscle), which tightens the ligament closed to the thyroid cartilage to get the infoldings together. The false vocal folds and epiglottis are located above the both sides of the vocal folds.

The infoldings open when inhalation, closed during breathing, and vibrate as the production of speech. The opening and closing of the vocal folds cause the oscillation to produce the speech. During the oscillation, the vocal folds are brought to get together by the muscles, and then are pushed

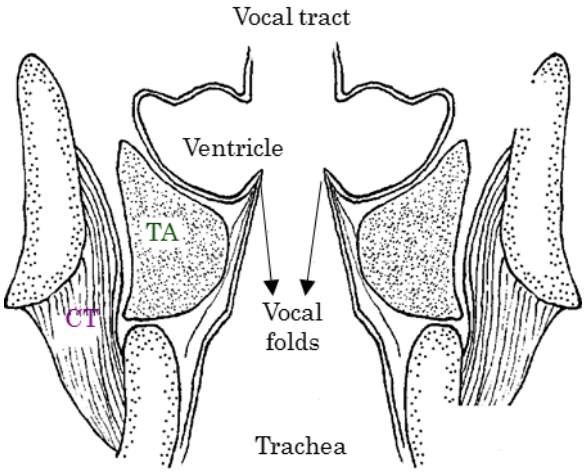


Figure. 4.2 Structure of the vocal folds

apart by the continuous subglottal pressure generated from lungs. The air expelled from the lungs cause the vibration of the two inflodings, generating the periodic impulse during the phonation.

The size of the vocal folds is different depending on speakers, especially for male and female. Generally speaking, the adult male has longer and thicker folds producing the lower pitched sound, and shorter and thinner folds for female to generate the high frequency sound. The length of the vocal folds for male is ranged from 1.75cm to 2.5cm, and 1.25cm to 1.75cm for female. The variation in the size of the vocal folds results in a difference in pitch of produced pitch. And cause the different tone for male and female speakers.

Generally, the stiffness of vocal folds mainly depends on two muscles: the cricothyroid muscle and the thyroarytenoid muscle. The weighted activities of cricothyroid and thyroarytenoid muscles can be denoted as CT and TA. TA is the main factor causing the variation in the fundamental frequency (F0) in neutral phonation. Thus, in the normal speech phonation modus, an increase in F0 results from higher TA activity while CT activity is relatively low. The contraction of the thyroarytenoid muscle (TA) tends to make the vocal fold stiff decrease and with increasing viscous, while the contraction of

cricothyroid muscle (CT) tense the vocal fold, resulting in a rise of stiffness. When CT is contracting, the tension of ligament is becoming large, and the ligament will vibrate instead of TA muscle [96].

Because the contract of the vocal folds, the viscous loss of the vibrating cords is caused to represent the “stickiness” of the soft, moist contracting surfaces as they form together. An osmotic gradient resulting form mucus is established as the vocal folds is contracting which induce different hydration effect. When the vocal fold mucosa was dried, causing dehydration, result in the increase of both stiffness and viscosity of vocal folds. When the vocal folds are exposed in the humid air, it is re-moistened, and decreases the stiffness and viscosity [97].

4.3 ESTIMATION METHOD FOR PHYSICAL PARAMETERS

4.3.1 PHYSICAL PARAMETERS FOR THE VOCAL FOLDS

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 physical parameters proposed as features for classification in the two-mass model are stiffness, damping ratio, and subglottal pressure.

- **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 [98] 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.

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

$$s_i(x_i) = k_i(x_i + \eta x_i^{\ 3}), \qquad i = 1,2 \tag{4.1}$$

where k_i are stiffness coefficients.

In the two-mass model, the coupling stiffness k_c is relative to the tension in the thyroarytenoid muscle (TA), and stiffness coefficients are represented as k_1=CT+TA and k_2=CT-TA [96]. Therefore, high TA and low CT values during normal phonation lead to high values for k_1, while k_2 is relatively low.

- **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 [99]. The damping ratio of viscosity has been estimated by Kaneko *et al.* [100] and Isshiki [101]. Results show that damping ratio has a close correlation with fundamental frequency, which is a stress indicator [102]. 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

$$r_1 = 2\zeta_1\sqrt{m_1 k_1} \qquad r_2 = 2\zeta_2\sqrt{m_2 k_2}\,, \tag{4.2}$$

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

- **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. In two-mass model, the subglottal pressure is denoted as P_s.

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

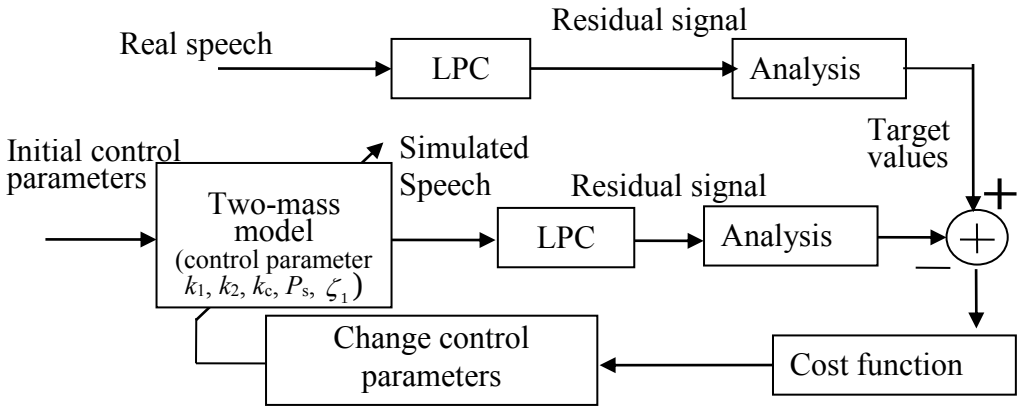


Figure. 4.3 Structure of algorithm.

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 [61]. Therefore, in this work, 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.

4.3.2 ALGORITHM FOR FITTING THE MODEL TO REAL SPEECH

Figure 4.3 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, parameters sets $[k_1, k_2, k_c]$, $[\zeta_1]$ and $[P_S]$ are 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. 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 (see 4.3.3). The control parameters are then varied and the speech is simulated until the cost function reaches a minimum.

The Nelder-Mead algorithm [103] 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.

4.3.3 COST FUNCTIONS

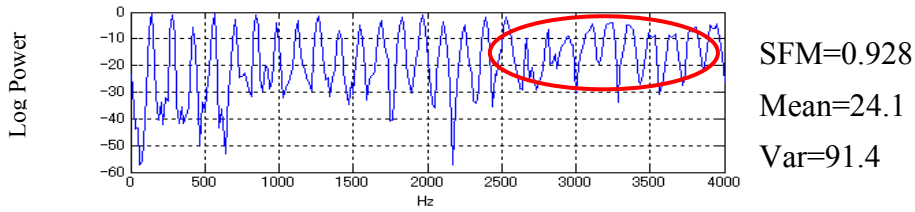
We utilize four different cost functions in order to compare their performance in classification.

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

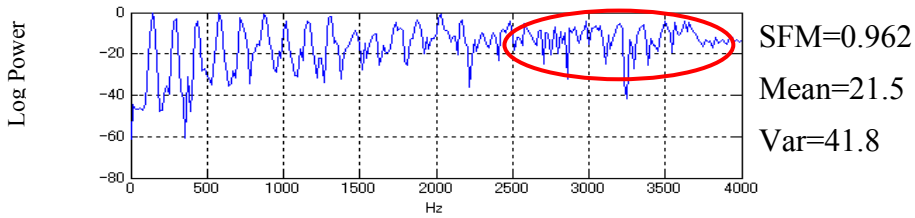
When stress occurs, the fundamental frequency and spectrum of the glottal source are affected. The harmonic structure of the spectrum loses clarity in the high-frequency band, and the spectrum becomes smooth and irregular. The spectrums of residual signals are shown in Figure 4.4. The part of high frequency 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:

$$SFM = \frac{\sqrt[M]{\prod_{n=0}^{M-1} S(n)}}{\frac{1}{M}\sum_{n=0}^{M-1} S(n)}, \tag{4.3}$$

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 Figure 4.5. The distribution is calculated using 180 samples from a male speaker.



(a) Neutral speech



(b) Stressed speech

Figure. 4.4 Spectrum of residual signals for a male speaker.

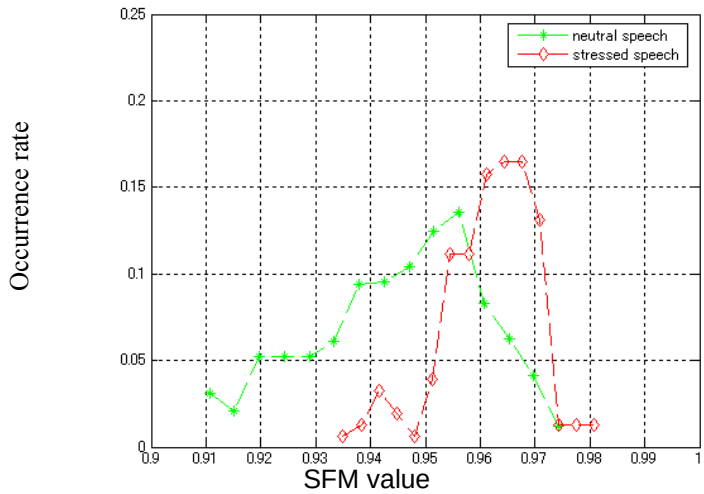


Figure. 4.5 Distribution of SFM for spectrum of residual signals

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

$$\begin{aligned} C_1 &= \alpha_1 \left(F_0^* - F_0\right)^2 + \alpha_2 \left(SFM^* - SFM\right)^2, \\ \alpha_1 &= 1/\overline{F_0}, \alpha_2 = 1/\overline{SFM}, \end{aligned} \tag{4.4}$$

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 the mean over the range of target values. The frequency band of the spectrum was limited to 3000-4000 Hz for calculating the spectral flatness measure.

- 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 4.6 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 Figure 4.4(b)). The cost function is defined as

$$\begin{aligned} \mathrm{C}_2 = {} & \beta_1 (F_0^* - F_0)^2 + \beta_2 \left| \mathrm{mean}(S^*(n)) - \mathrm{mean}(S(n)) \right|^2 \\ & + \beta_3 \left| \mathrm{var}(S^*(n)) - \mathrm{var}(S(n)) \right|^2, \end{aligned} \tag{4.5}$$

where $\beta_1 = 1/\overline{F_0}$, $\beta_2 = 1/\overline{\mathrm{mean}(S(n))}$ and $\beta_3 = 1/\overline{\mathrm{var}(S(n))}$ are used to normalize target values to the same range. The overbar denotes the mean over the range of target values. The frequency band of the spectrum was limited to 3000-4000 Hz.

- 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

$$H(k) = M \cdot B(X = k), \tag{4.6}$$

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

$$\begin{aligned} \mathrm{C}_3 &= W_1 \sum_{n=1}^{M} \left(S^*(n) - S(n)\right)^2 + W_2 \sum_{j=1}^{L} \left(H^*(k_j) - H(k_j)\right)^2 \\ W_1 &= 1 \Big/ \sum_{n=1}^{M} \left(S(n)\right)^2, W_2 = 1 \Big/ \sum_{j=1}^{L} \left(H(k_j)\right)^2, \end{aligned} \tag{4.7}$$

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 3000-4000 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

Step1: Spectrum is split into (overlapping) frames. Frame length is fixed as the frequency band including three harmonic structures.

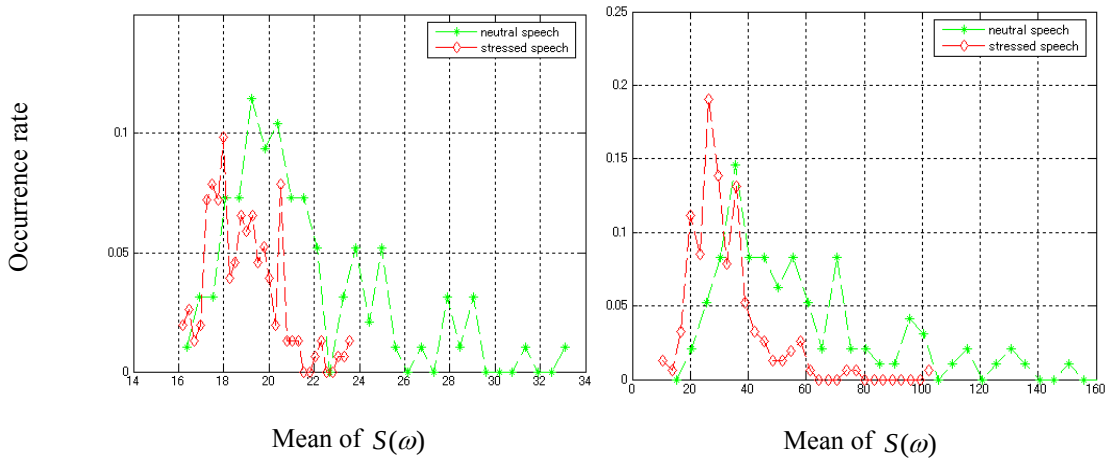


Figure. 4.6 Distributions of mean and variance of spectrum of residual signal for neutral (green) and stressed speech (red).

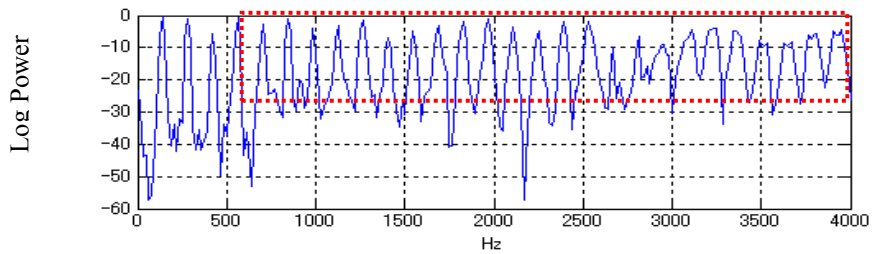


Figure. 4.7 Cut-off spectrum with a threshold. Spectrum within the dotted line is emphasized for calculation of cost function.

Step2: Autocorrelation is calculated for each frame.

Step3: Zero-crossing for the autocorrelation is computed to classify whether it has a clear harmonic structure in this frame.

Step4: Separation point is determined by an abrupt increase in zero-crossing.

- Modified spectrum (Spectrum)

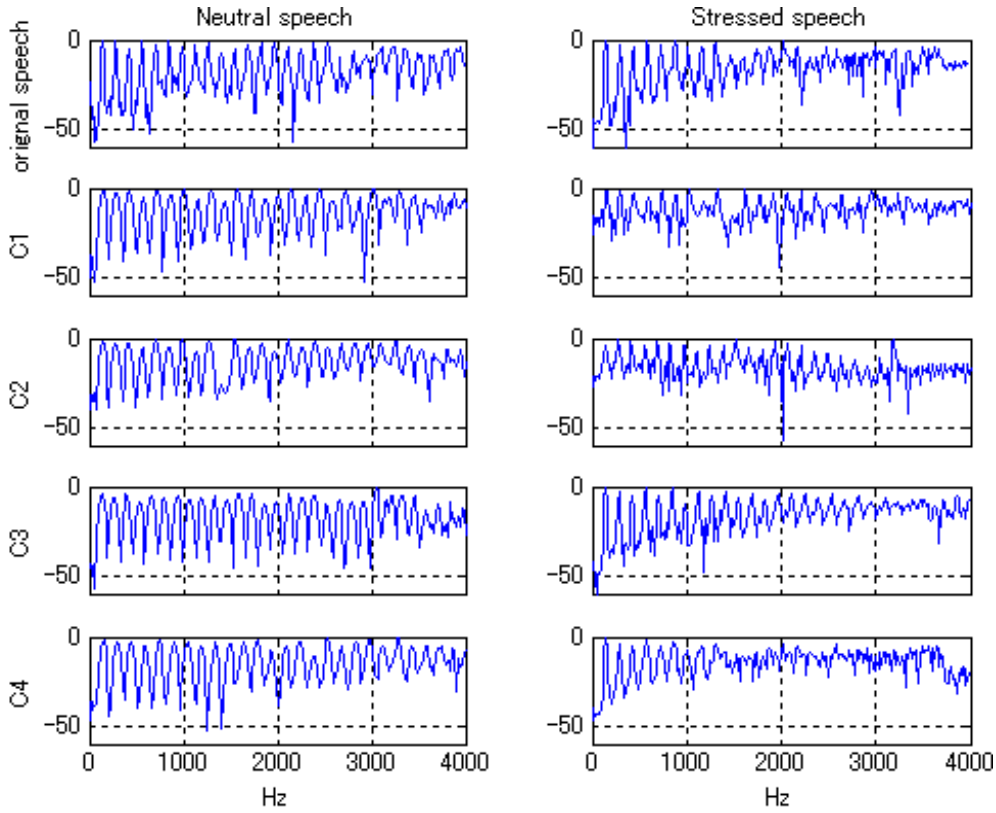


Figure. 4.8 Spectrums of residual signals for original speech (top) and for simulated speech with different cost functions under neutral (left column) and stressed (right column) conditions.

The 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 Figure 4.7, which is a power spectrum of speech from a male speaker. The threshold is 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

$$C_4 = \frac{\sum_{n=1}^{M} \left| P^*(n) - P(n) \right|^2}{\sum_{i=1}^{M} \left| P(n) \right|^2}, \tag{4.8}$$

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

Figure 4.8 shows the simulated results with these four cost functions. In this experiment, the neutral and stressed speech in Figure 4.4 from a male speaker are used to estimate 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 4.1.

Table 4.1 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
C_1	438	75460	7840	22640	0.16	299	90780	8040	8260	0.32
C_2	455	74030	8250	21980	0.14	276	87440	8277	7260	0.32
C_3	416	74270	7730	20810	0.17	306	84290	7800	7740	0.31
C_4	446	77360	8000	22600	0.14	279	89170	8480	7650	0.34

4.4 EXPERIMENTAL EVALUATIONS

4.4.1 DATA SELECTION AND EXPERIMENTAL SETUP

In the experiments, we used a database collected by the Fujitsu Corporation containing speech samples from eleven subjects (four males and seven females). 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 speech 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 males, three females) from eleven subjects to show the classification performance. Linear classifiers based on the minimum Euclidean distance to reach the classification performance were used. 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.

4.4.2 EVALUATION FOR THE 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

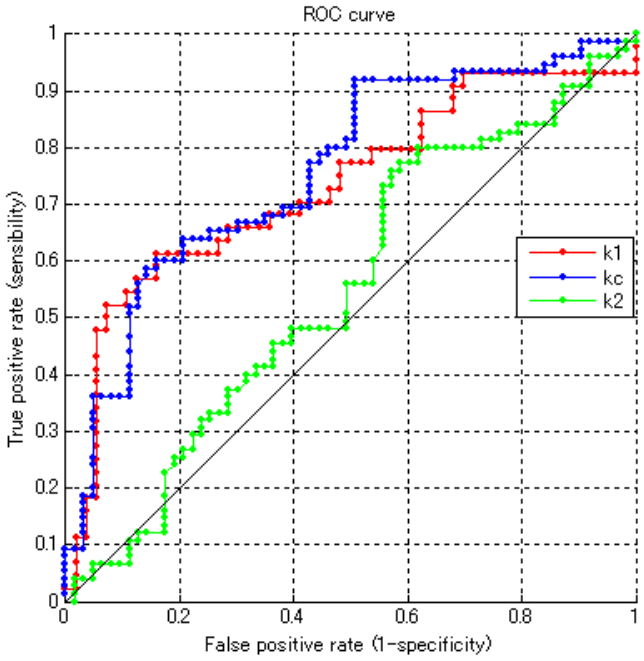


Figure. 4.9 ROC curve for stiffness parameters (k_1, k_2, k_c).

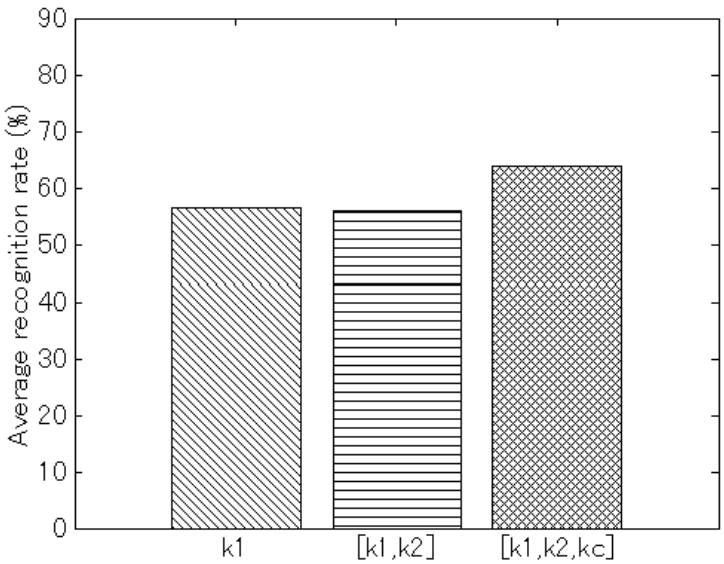


Figure. 4.10 Classification performance for stiffness parameters.

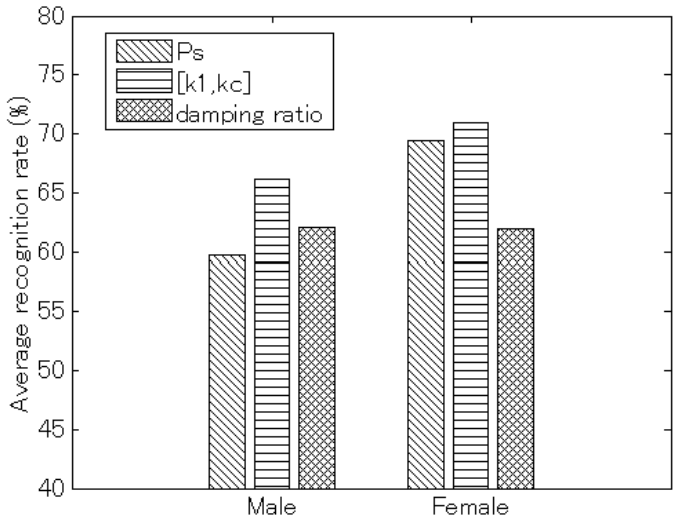


Figure. 4.11 Performance of each parameter for each gender.

parameters are dependent on speakers. The proposed parameter sets were then compared to show their classification performance using C4 as the cost function.

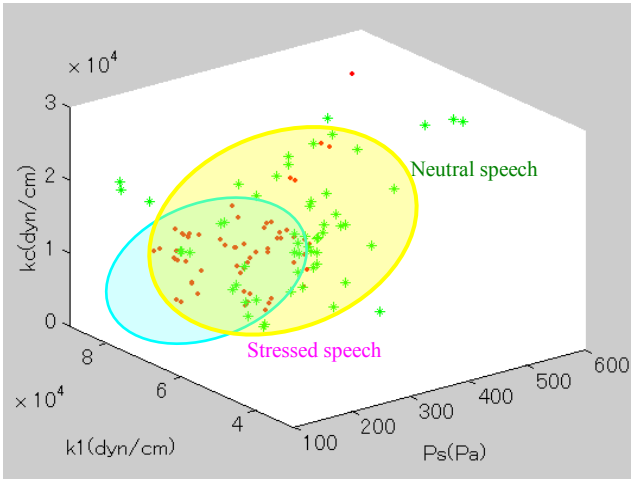
In the first evaluation, the stiffness parameters were first focused on and the effect of each stiffness on stress recognition was then examined. The parameters k_1, k_2, and k_c were estimated with P_S = 500Pa, and $\zeta_1 = 0.1$, and other physical parameters were fixed at the typical values described in Sec. 4.1. In Figure 4.9, 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 Figure 4.10. It is illustrated that the average classification accuracy decreases 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

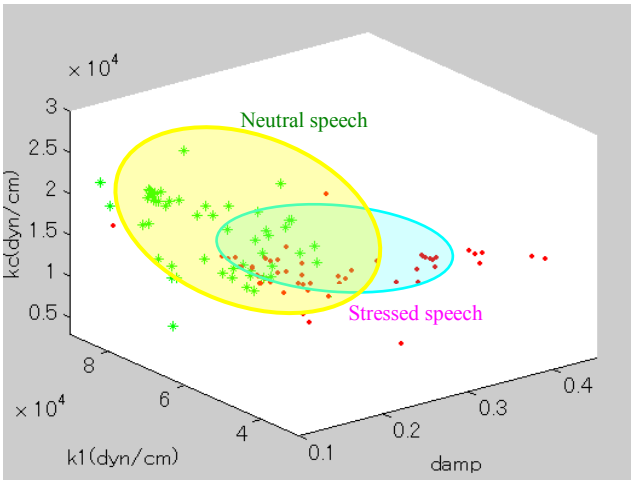
other parameters at typical values. Then the effect of each parameter on stress recognition was examined. The results are shown in Figure 4.11 (in Figure 4.12-4.14, “damp” is the abbreviation for “damping ratio”). For these physical parameters, the results show that stiffness ($k1$, kc) achieves the 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.

F0 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 4.12 (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 (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 Figure 4.12 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.



(a) Distribution for P_S, k_1 and k_c.



(b) Distribution for k_1, k_c and damping ratio

Figure. 4.12 Distributions of estimated parameters.

We checked the performances of the above parameters and compare them. In the proposed sets, the stress 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 F0 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.

4.4.3 COMPARISON WITH 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 C_2 and C_3, 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 average classification performance for different cost functions is shown in Figure 4.13. Results show that cost function C_4 yields the best improvement in classification performance.

Since the proposed features are based on physical characteristics, it would be helpful to compare their performances with those of traditional features. Here the cost function C_4 is selected, which achieves the best performance in the last evaluation. 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 [61]. The results of this comparison are shown in Figures 4.14. 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.

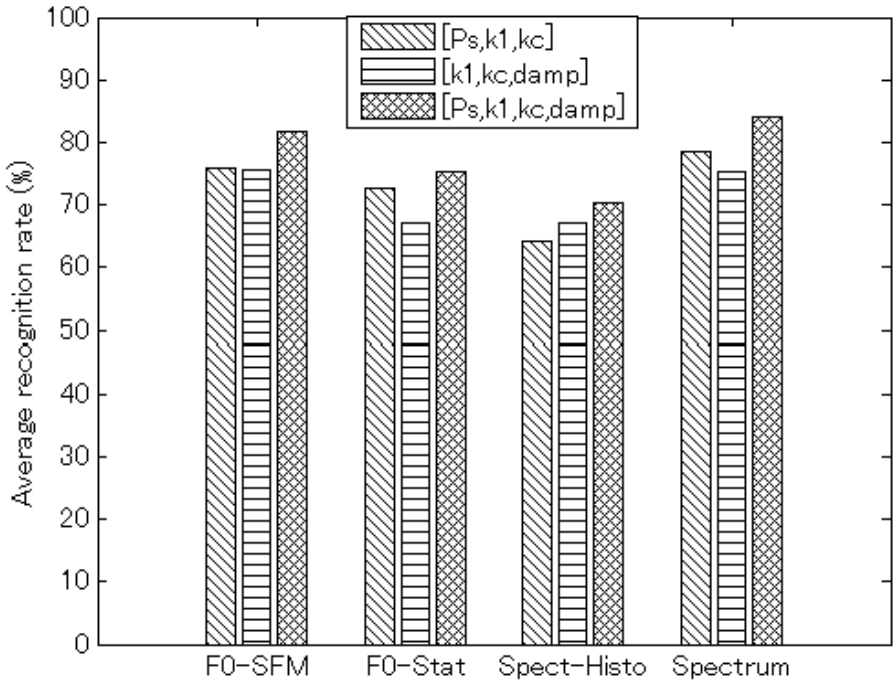


Figure. 4.13 Average results for proposed parameter sets with different cost functions.

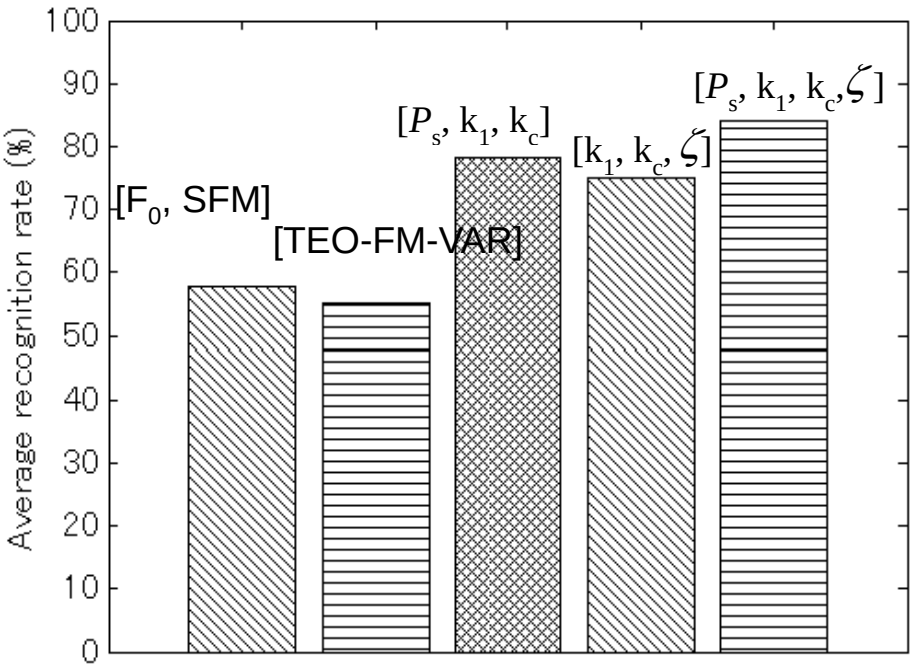


Figure. 4.14 Average classification performance comparing with traditional features.

4.5 SUMMARY

In this chapter, 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, which affect the glottal source, are key indicators of stressed speech. Stress causes higher tension in the muscle tension of the vocal folds, lower subglottal pressure from the lungs and the vocal folds become more viscous than under neutral conditions. 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.

CHAPTER 5: STRESS CLASSIFICATION BASED ON MODELING OF THE VOCAL FOLDS AND THE VOCAL TRACT

5.1 INTRODUCTION

In the previous chapter, we propose a new classification method, based on the working mechanisms of the vocal folds, for speech under stress. The vocal fold parameters are estimated from a two-mass model for the classification of stressed speech. However, we assumed 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. This is something of an oversimplification. Therefore, in this paper, we concentrate on estimation of vocal tract parameters representing cross-sectional areas and vocal tract length.

We propose a stressed speech classification method based on a physical model characterizing the vocal folds (VF) and the vocal tract (VT). This method can represent the process of speech production and model airflow patterns in the vocal folds and the vocal tract, which are essential for stress classification. In this physical model, changes in the physical characteristics of the vocal folds, such as muscle tension, have a modulating effect on the formants, while the shape of the vocal tract can also influence the glottal source because of the interaction between the vocal folds and the vocal tract. It is believed that the presence of stress can result in variations in the physical characteristics of physiological systems. When a speaker is under stressed condition, muscle tension of the vocal folds will become higher and the area at the entrance to the vocal tract become wider than under the neutral condition. The variations will result in an impact on the acoustic interaction between the vocal folds and the vocal tract [91]. The parameters of the physical model are also helpful to

represent the influence of speaking style more directly and clearly. Therefore, a physical model is helpful to estimate the parameters of the physiological system.

The two-mass model is a physical model, which attempts to simulate the physical process of vocal fold vibration, which characterizing the vocal folds and the vocal tract, and to also model the effect of glottis-vocal tract interaction [93]. Parameters affected under stressed conditions are extracted from the physical model and are used as features to identify speech under stress more precisely. We use the two-mass model as a physical model, and our proposed method estimates the values of parameters included in the model from input speech. To identify speech under stress, we evaluate parameters affected by stress.

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

The chapter is organized as follows. In Section 5.2, an overview of our method is presented Physical parameters, related to the vocal folds and the vocal tract, based on the two-mass model, are described as features for classification in Section 5.3. This is followed by the presentation of a fitting algorithm for real speech data in Section 5.4 to help estimate the physical parameters. Section 5.5 describes the

classification method used for evaluation. In Section 5.6, experiments are performed to evaluate the obtained parameters and show their corresponding classification performances when separating neutral and stressed speech. Finally, we draw our conclusions in Section 5.7.

5.2 PHYSIOLOGICAL BASIS FOR THE VOCAL TRACT

The vocal tract is defined as the structure bound by soft and hard tissues, which can be shaped by tongue, mouth, teeth, oral cavity, palate, nasal cavity and other articulators, which is shown in Figure 5.1. Some structures such as teeth and hard palate are immoveable. The articulators including tongue, jaw, lips, oral cavity and nasal are moveable, and associate with speech production to generate different sounds. The movement of these articulators causes the variation in the shape of the vocal tract leading to articulations. However, the shape of the vocal tract is caused by the movement of other physical structures. For example, the vibration of the vocal folds can result in a shortening or lengthening of the vocal tract, and a downward motion of glottis with block of the oral and nasal cavities, can generate negative pressure in the vocal tract. Vortices are produced around the false vocal folds resulting from the negative pressure can also generate sound, which can be considered as the other source to produce the speech.

Differences in human voices are caused by the different sizes of vocal tract (VT), thus their generated speech signals contain frequencies that are not always constantly the same. The physiology and linguistic differences between speakers are known to be the inter-speaker variability, affecting the overall performance for continuous Automatic Speech Recognition (ASR) and stress

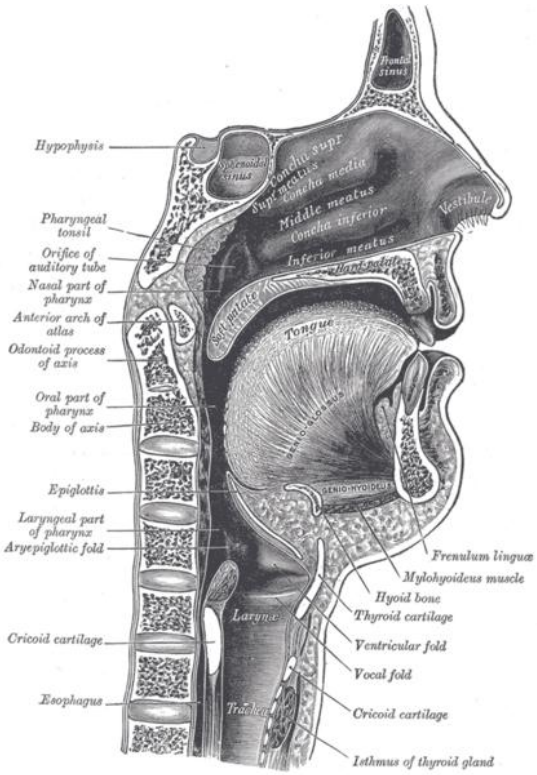


Figure 5.1 Structure of the vocal tract [104]

classification system.

One physical source of inter-speaker variability is the vocal tract length (VTL). Physical difference in VTL is more noticeable between male and female speakers. Male speakers have longer VT that generates lower frequency speech spectrum. On the other hand, female speakers have shorter VT which generates higher frequency speech spectrum. According to Lee & Rose [105] [106], VTL can vary from approximately 13 cm for adult females to over 18 cm for adult males. Due to these VTL differences, it is necessary to estimate speaker's vocal tract length in speaker dependent system.

The supraglottic area includes the structure that lies above the true vocal folds and below the tongue base. The anatomical structures present in this area that are important to speech production lie posterior to the epiglottis. They include: the ventricle; the false vocal folds; the epilarynx; the arytenoids; the laryngeal aspects of the aryepiglottic folds; and the vestibule [107]. Figure 5.2 shows

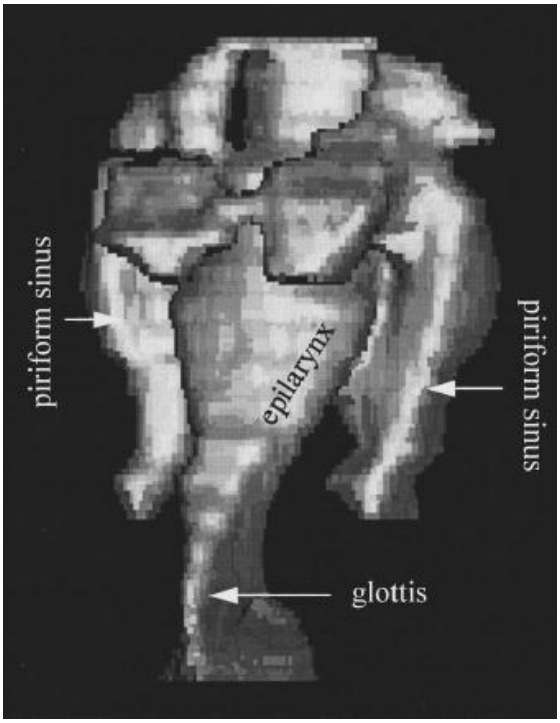


Figure 5.2 Anterior-lateral view of a vocal tract airway [108]

anterior-lateral view of a vocal tract airway. The epilarynx tube is the narrow portion located above the glottis. On the top of epilarynx tube, the structure is widened by a pharynx and two piriform sinuses [108].

A strong interaction between the glottal source and the vocal tract can exist. The interaction is mainly influenced by the laryngeal vestibule in the supraglottis (epilarynx tube). Variations in the cross-sectional area of supreglottis can result in the change of inertive reactance of supraglottal vocal tract, which enhances the pressures of the vocal folds and the glottal flow. The ratio of glottal area and cross-sectional area of epilarynx tube is a parameter influencing the interaction because of the changes in the supraglottal impedances. The harmonic distortion for the spectrum of glottal source is produced from the contributions of the interaction.

5.3 PHYSICAL PARAMETERS

5.3.1 PARAMETERS FOR THE VOCAL FOLDS AND THE VOCAL TRACT

A method which fits the two-mass model to real speech is proposed for classifying speech under stress. Some of the physical parameters characterizing the vocal folds and the vocal tract are estimated. The two-mass vocal fold model was original proposed by Ishizaka and Flanagan to simulate the process of speech production [93]. We propose three types of feature parameters extracted from the two-mass model: stiffness, vocal tract length, and cross-sectional area of the entrance of the vocal tract. In the following sections, we will define these parameters and describe their characteristics.

- Stiffness

The stiffness parameters are related to muscle tension in the vocal folds. Generally, the stiffness of the vocal folds is considered to depend mainly on two muscles: the cricothyroid muscle (CT) and the thyroarytenoid muscle (TA) [96]. In the two-mass model, coupling stiffness k_c is relative to the tension in the TA muscle, so a high k_1 value and a low value for k_c represent the contraction of the CT muscle and relaxation of the TA muscle, which are described in 3.1.

Stiffness parameters are the main factors relating to fundamental frequency, and they can also determine the amplitude of the glottal area and glottal volume velocity [97], so source excitation is significantly influenced by the degree of stiffness. During the production of speech, the natural frequency of the vocal folds is determined by both their mass and stiffness. However, in order to simplify the estimation algorithm, only the stiffness parameters are estimated, with mass fixed as a

constant.

- Vocal tract length and cross-sectional area

One physical source of the inter-speaker variability is the differences in vocal tract length (VTL), which influence spectral frequencies of the generated speech signals. Physical differences in VTL are more marked between male and female speakers. VTL can vary from approximately 13 cm for adult females to over 18 cm for adult males, and differences in VTL influence spectral formant frequency, which varies by as much as 25% among adult speakers. Due to the variation caused by differences in VTL, it is necessary to estimate a speaker's vocal tract length in speaker dependent systems.

The two-mass model is connected to a four tube model representing the vocal tract [93]. The tube model is constructed using a transmission line analogy involving n cylindrical, hard-walled sections. The elemental values of the model are determined by cross-sectional areas $A_1 \cdots A_n$, and cylinder lengths $l_1 \cdots l_n$. The total length of the vocal tract is defined as L_{VT}. The details related to the vocal tract are described in 3.3.

5.3.2 RELATIONSHIP BETWEEN PHYSICAL PARAMETERS AND ACOUSTIC PARAMETERS

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

When airflow propagates 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:

$$P_1 - P_{22} = \frac{1}{2}\rho\frac{U_g^2}{A_{g2}^2}\left[2N(1-N)\right], \tag{5.1}$$

where P_1 is the pressure at the inlet of the vocal tract. Here the parameter N is defined as $N = A_{g2}/A_1$, where A_1 is the area of the entrance to the vocal tract. N denotes the difference in area between the outlet of the vocal folds and the inlet of the vocal tract, which is significant to the acoustic interaction between the glottal source and the vocal tract [93]. Since glottal area A_{g2} does not change significantly during the oscillation of the vocal folds, A_1 is the parameter relating to the acoustic interaction.

In 3.3, Equation (3.19) shows that airflow velocity U_g depends on both the stiffness of the vocal folds and area of the entrance to the vocal tract A_1. Therefore, it is our assumption that parameters k_1, k_2, k_c, A_1 related to velocity have an impact on acoustic interaction. In this paper, experiments are performed to represent the presence of this interaction by showing the relationship between physical and acoustic parameters. Due to the presence of these interactions, changes in the oscillation of the vocal folds affect the distribution of formants, and different shapes of the vocal tract (length and area) also influence the glottal source. Table 5.1 lists and describes the physical and acoustic parameters.

Table 5.1 Physical and acoustic parameters

Parameter	**Variable**
Physical	k_1, k_2, k_c, A_1, A_2, A_3, L_{VT}
Acoustic	F_0, F_1, F_2, F_3

We first examine how stiffness parameters impact the distribution of formants. First, we fixed the shape of the vocal tract and examined how variation in the stiffness parameters of the vocal folds affects the shift of formants. The vocal tract model was represented by a standard tube configuration for the each vowel [94]. In order to reduce the number of parameters to be estimated, and simplify the proposed method, typical values were adopted for the configuration of the tube model. Therefore, as typical values, the length chosen for the vocal tract was $L_{VT} = 16cm$, with each element $l_i = 4cm$, was fixed. When a specific stiffness is checked, the other stiffness parameters are fixed at typical values. We changed stiffness parameters k_1 (20—240kdyn/cm), k_2 (2—40kdyn/cm) and k_c (2.5—70kdyn/cm) to examine variation in formants. Formant estimation is based on modeling vocal tract frequency response using linear prediction coding (LPC) techniques. It estimates formant frequencies from the all pole model of the vocal tract transfer function.

Figure 5.3 shows the relationship between the stiffness parameters and different formants. It shows that k_2 does not significantly influence formants, but that first and second formants will shift their location to a lower frequency with the increase of k_1, althrough there is no significant change in the third formant (F_3). A similar phenomenon occurs for k_c. When k_c decreases, F_1 also has a tendency to shift to a lower frequency, while F_2 and F_3 are less influenced by the variation of k_c. Therefore, it is shown that stiffness

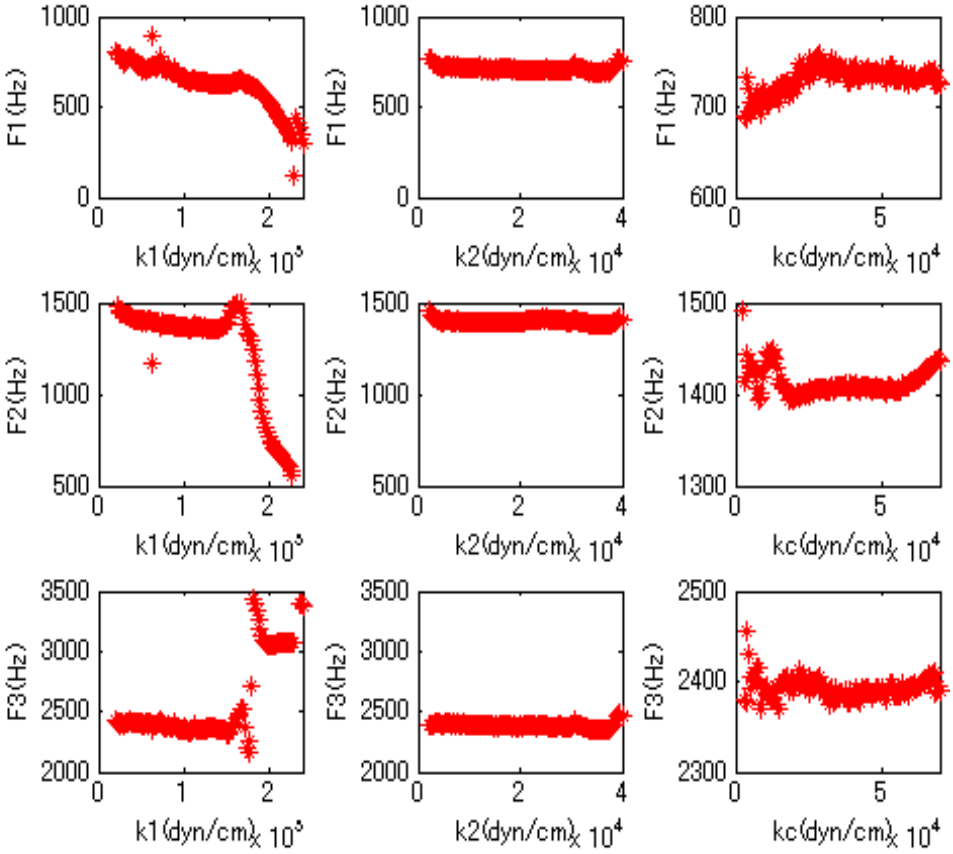


Figure 5.3 Impact of stiffness parameters in vocal folds on formants.

parameters k_1 and k_c can affect the distribution of formants, and that the first and second formants are easily affected by acoustic interaction.

Next, we fixed the configuration of the vocal folds and examined how variation of the cross-sectional area of the vocal tract impacts the fundamental frequency (F_0) of speech. Stiffness was fixed at typical values $k_1 = 80000\ dyn/cm, k_2 = 8000\ dyn/cm, k_c = 25000\ dyn/cm$ to check how the fundamental frequency changes with the area function. When checking the impact of a specific area, other areas and VTL were fixed at typical values. When considering VTL, all the cross-sectional areas were fixed at typical values. We then change cross-sectional area or VTL to examine their impact on F0. The variation range for VTL was 13 to 19cm, and cross-sectional area of VT ranged 0.1 to 20 cm^2. The algorithm for estimation of the fundamental frequency of speech is YIN [109]. It is based

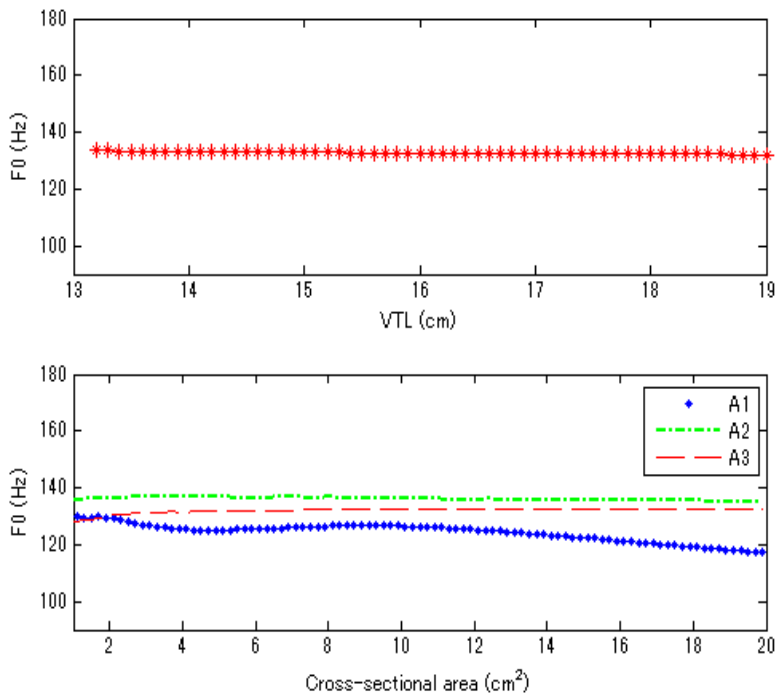


Figure 5.4 Impact of vocal tract length and cross-sectional area of vocal tract on fundamental frequency.

on the well-known autocorrelation method, with a number of modifications used to prevent error.

Figure 5.4 shows the relationship between the vocal tract parameters (vocal tract length and cross-sectional area A_1, A_2, A_3) and fundamental frequency. It shows that vocal tract length (VTL) has less impact on F_0, and only determines the distribution of formants. However, an increase in cross sectional area A_1 can cause F_0 to change significantly. While cross sectional area A_2 and A_3 also have an impact on F_0 to some extent, their influence is insignificant compared to A_1. Therefore, A_1 is the parameter which we believe is related to the acoustic interaction between the glottal source and the vocal tract.

Therefore, it is our conclusion that stiffness of the vocal folds k_1, k_c and cross-sectional area A_1 affect both the fundamental frequency and formants, and further, the interaction between the glottal source and the vocal tract.

The experimental results show that stiffness of the vocal folds and cross-sectional area A_1 have an impact on the interaction between the glottal source and the vocal tract. It is believed that the variations in acoustic interaction differ markedly between neutral and stressed speech [91], so stiffness and A_1 should be selected as parameters for representing stress.

In theory, Equation (5.1) shows that both the velocity of glottal airflow, and the difference between the area of the outlet of the vocal folds and the inlet of the vocal tract, have an impact on the pressure difference inside and outside of the glottis. So the two factors can cause variations in the airflow patterns in the glottis, and thus are likely to be effective to represent the presence of stress.

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

A_1 in the four-tube model is the area of the entrance to the vocal tract in the supraglottis. Narrowing A_1 facilitates phonation by decreasing the oscillation threshold pressure of the vocal folds [108]. Since glottal area A_{g2} does not change significantly during the oscillation of the vocal folds, A_1 is the main factor determining the pressure difference between the inside and outside of the glottis and has an impact on the acoustic interaction between the glottal source and the VT. Based on these considerations, we also make the assumption that A_1 is an effective parameter for stress classification.

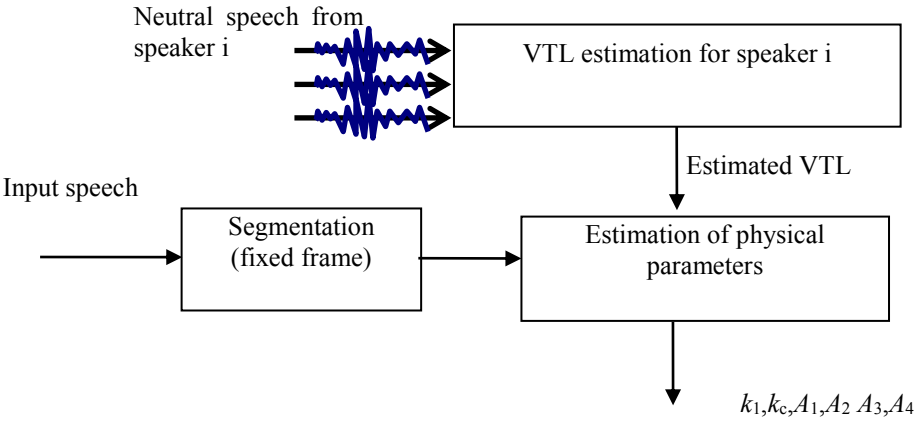


Figure 5.5 Block diagram showing the an outline of our method

5.4 ESTIMATION METHOD FOR PHYSICAL PARAMETERS

5.4.1 ESTIMATION FOR THE VOCAL TRACT LENGTH (VTL)

The goal of stress classification is to determine from speech data whether a specific person is under stress or not. When speech is input to the system, it is split into several frames, and further estimation of the physical parameters is performed for each frame. VTL for each speaker is first calculated, then the obtained VTL is input as a known parameter. Then the two-mass model is fit to each speech sample to simulate the vocal folds and the vocal tract. An outline of our method is shown in Figure 5.5.

In the first step, estimation of vocal tract length (VTL) is performed. Since vocal tract length (VTL) has no impact on the glottal source, it can be estimated separately. Because VTL varies with each speaker, all of the neutral speech data from each speaker is used to estimate the vocal tract length of that speaker. Here we mainly consider the neutral speech for each speaker in the database. During

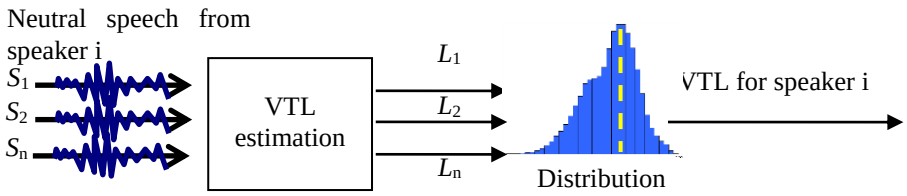


Figure 5.6 Block diagram showing the algorithm for VTL estimation. This method utilizes all of the neutral speech for each speaker.

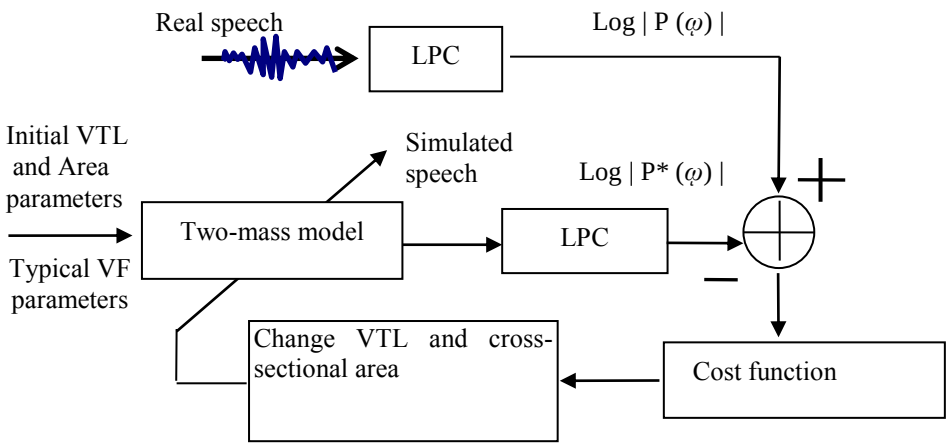


Figure 5.7 Block diagram showing the details of estimation of vocal tract length

VTL estimation, real speech coming from a database is analyzed using linear predictive coding (LPC) to reach the spectral envelope. The stiffness parameters are fixed at the typical values and are taken as an input $k_1 = 80000$ dyn/cm, $k_2 = 8000$ dyn/cm, $k_c = 25000$ dyn/cm. The two-mass model is then fit to the neutral speech of each speaker to estimate the parameters of vocal tract length and cross-sectional area. Nelder-Mead simplex method [103] is used for searching the optimal values for fitting. $L_{VT}(i,k)$ is estimated for the sample k from the speaker i. For each speaker i, the probability distribution $P_i(L_{VT}(i,k))$ of VTL for all neutral speech is calculated, and we choose the one with the highest probability as the estimated vocal tract length.

$$L_{VT}(i)^* = arg\,max_{L_{VT}(i,k)}\, P_i(L_{VT}(i,k)) \qquad (5.2)$$

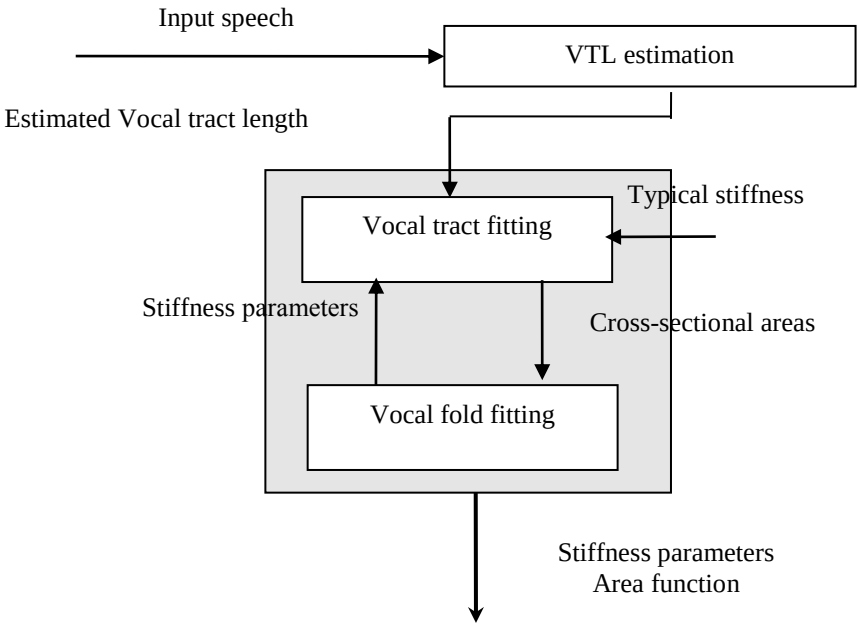


Figure 5.8 Structure of main fitting algorithm, which includes three parts: (1) estimation of VTL, (2) vocal folds fitting, (3) vocal tract fitting.

The algorithm for VTL estimation is shown in Figure 5.6. The detailed fitting procedure is the same as that used for vocal tract fitting described below, which is shown in Figure 5.7.

5.4.2 ITERATION ALGORITHM FOR FITTING

In the next step, the obtained VTL is input as a known parameter, and the two-mass model is fit to each speech sample for every speaker. Fitting the model to real speech poses a difficulty: estimation of too many parameters may make the fitting method unstable. The solution to this problem is to split the process into two main parts, so that the vocal folds (VF) and the vocal tract (VT) are fit with two different cost functions. However, the existence of interaction between VF and VT makes it

impossible to fit VF and VT separately, and changes in the stiffness parameters and in A_1 in the tube model can influence both formants and the glottal source. An alternative is to perform iteration when fitting the vocal folds and the vocal tract. So an iteration method is used for vocal fold and vocal tract fitting, which are accomplished as follows. Figure 5.8 shows the structure of the fitting algorithm.

For vocal tract fitting, stiffness parameters are fixed at typical values ($k_1 = 80000\,\text{dyn/cm}$, $k_2 = 8000\,\text{dyn/cm}$, $k_c = 25000\,\text{dyn/cm}$), and are taken as an input to vocal tract fitting. The parameters for the cross-sectional areas are then estimated. Next, the obtained areas are used as an input for vocal fold fitting, and the two-mass model is fit to estimate the new stiffness parameters. When current stiffness differs significantly from the typical value, the corresponding formants are also affected and some variations can occur. In such cases, vocal tract fitting needs to be performed again. We take iterations for the two parts until the results reach convergence.

The detailed structure of vocal tract fitting and vocal fold fitting is shown in Figures 5.9 and Figure 5.10. Vocal tract fitting includes two steps. First, real speech coming from a database is analyzed using linear predictive coding (LPC) to reach the spectral envelope. In the second step, a simulation is performed using the two-mass model to produce speech using an initial area function. The same spectral envelope is calculated from the simulated speech, and is compared with the one obtained in the first step to find the difference between them. The difference between the simulated spectrum and the target spectrum is represented by a cost function. The area function is then varied and glottal flow is simulated until the cost function reaches a minimum. Optimal values of the physical parameters are then estimated using the Nelder-Mead simplex method [103].

Here, we select A_1, A_2, A_3 and A_4, as variables in vocal tract fitting. Optimal values of the physical parameters are estimated using the Nelder-Mead simplex method, which is implemented to search

for the optimal physical parameters to minimize the cost function.

Vocal fold fitting uses the same process as vocal tract fitting, with the difference that the residual signal is obtained using LPC analysis, and the spectrum of the residual signal is available to construct the cost function 2 in Figure 5.10 for vocal fold fitting:

$$C_2 = \frac{\sum_{i=1}^{fs/2} \left| S^*(\omega_i) - S(\omega_i) \right|^2}{\sum_{i=1}^{fs/2} \left| S(\omega_i) \right|^2}, \tag{5.3}$$

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

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

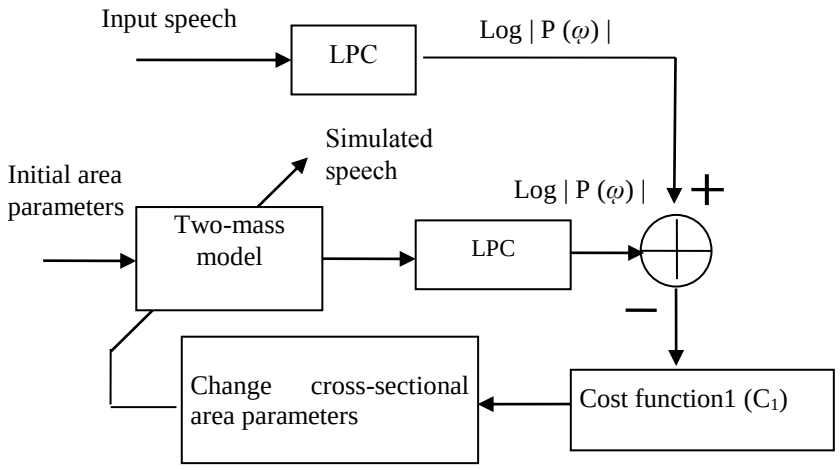


Figure 5.9 Block diagram showing the detailed structure of our vocal tract fitting method.

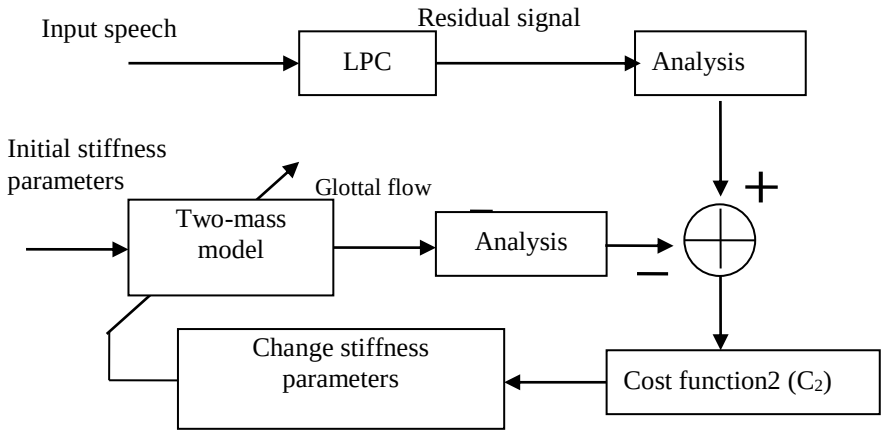


Figure 5.10 Block diagram showing the detailed structure of our vocal fold fitting method.

focused on the harmonic structure of the spectrum of residual signal. [110]. So in this study, obtaining the residual signal using LPC can work well for showing the harmonic structure of the spectrum.

5.4.3 COST FUNCTIONS

We utilized four different cost functions in order to compare their classification performance. Cost function 1 (C_1) in Figure 5.9 are defined as follows:

- *Formant* ($C_{F1\text{-}F2}$)

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

$$C_{F1-F2} = \alpha_1 \left(F_1^* - F_1\right)^2 + \alpha_2 \left(F_2^* - F_2\right)^2, \qquad (5.4)$$
$$\alpha_1 = 1/\overline{F_1}, \alpha_2 = 1/\overline{F_2}$$

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

- *RMS distance of spectral envelope* (C_{rms})

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

$$C_{rms} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left|\log P(\omega_i) - \log P^*(\omega_i)\right|^2},$$
$$P(\omega) = \frac{1}{|A(\omega)|^2} = \frac{1}{\left|\sum_{k=0}^{P} a_k e^{-j\omega k}\right|^2}, \qquad (5.5)$$

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

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

$$C_{I-S} = \frac{1}{N}\sum_{i=1}^{N}\frac{P(\omega_i)}{P^*(\omega_i)} - \log\frac{P(\omega_i)}{P^*(\omega_i)} - 1, \tag{5.6}$$

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

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

$$\begin{aligned} C_{E-F}^{(1)} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|\log P(\omega_i) - \log P^*(\omega_i)\right|^2} \qquad n = 1 \\ C_{E-F}^{(n)} &= \alpha_1\left(F_1^* - F_1\right)^2 + \alpha_2\left(F_2^* - F_2\right)^2 \\ &\quad + w_1(H_1^* - H_1)^2 + w_2(H_2^* - H_2)^2, \qquad n \geq 2 \end{aligned} \tag{5.7}$$

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

In order to describe the accuracy of the fitting method, we calculate the error in F_0, F_1, F_2, F_3 and F_4 between real and simulated speech with cost function $C_{E\text{-}F}$:

$$Err_{F0} = \left(F_0 - F_0^*\right)$$

$$Err_{F1} = \left(F_1 - F_1^*\right)$$

$$
\begin{aligned}
Err_{F2} &= \left(F_2 - F_2^*\right) \\
Err_{F3} &= \left(F_3 - F_3^*\right) \\
Err_{F4} &= \left(F_4 - F_4^*\right)
\end{aligned}
\tag{5.8}
$$

where the asterisk denotes the target value for real speech. Then we show the distributions of the error, as shown in Figure 5.11. Simulated results using these four cost functions are shown in Figure 5.12.

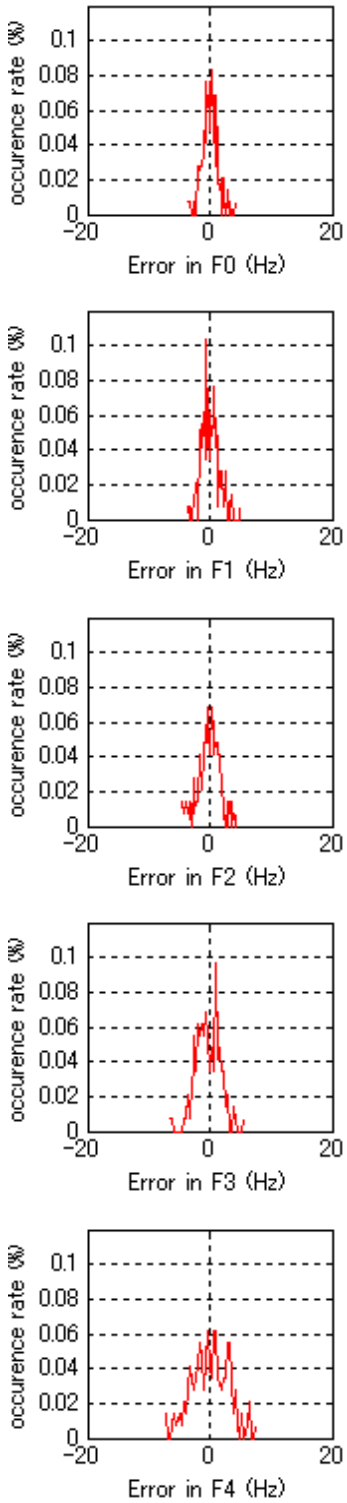


Figure 5.11 Error distribution of F_0, F_1, F_2, F_3 and F_4 between real and simulated speech. The cost function used is $C_{E\text{-}F}$,

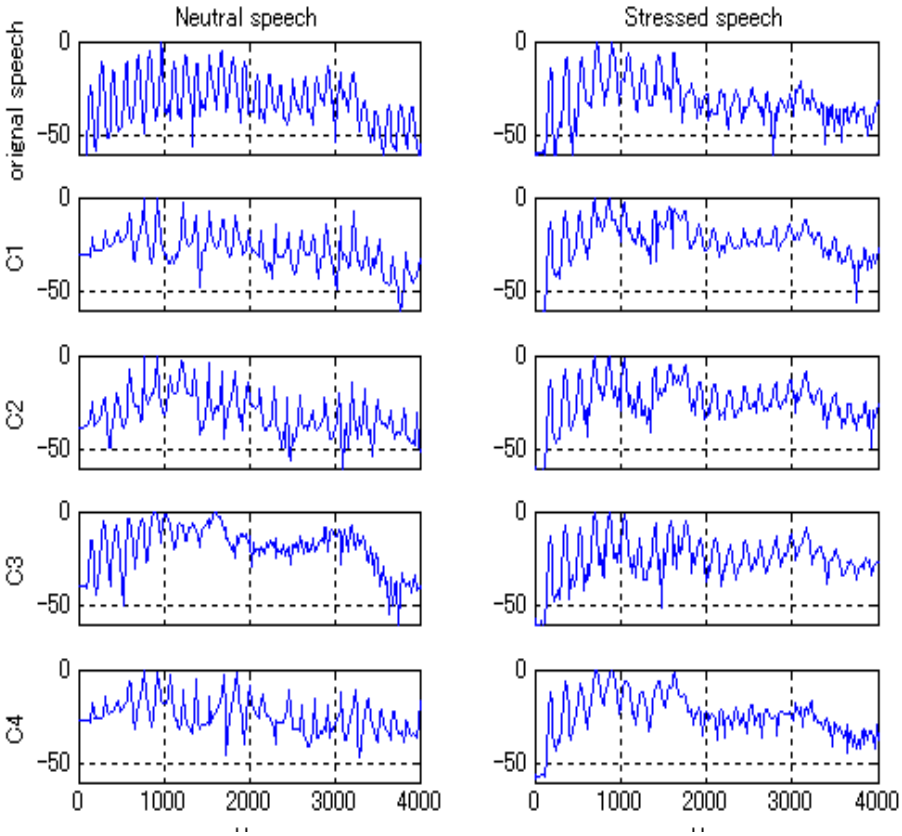


Figure 5.12 Simulation results of fitting for neutral and stressed speech. Spectrums for original speech (top) and for simulated speech with four cost functions (C_{F1-F2}, C_{rms}, C_{I-S} and C_{E-F}), under neutral (left column) and stressed (right column) conditions. In this figure, C1= C_{F1-F2}, C2= C_{rms}, C3= C_{I-S} and C4= C_{E-F}.

5.4.4 IMPROVED ALGORITHM FOR FITTING

Fitting the model to real speech poses a difficulty: the existence of interaction makes it impossible to fit VF and VT separately. The iteration method considers the interaction, but increase the complexity to perform the iteration.

VTL for each speaker is first calculated using the algorithm described in the section 2.2.1. When speech is input to the system, it is split into several frames, and further estimation of the physical

parameters is performed for each frame. The main structure is shown in Figure 5.13. Based on the analysis in 5.3.2, it is believed that stiffness parameters k_1, k_c and cross-sectional area A_1, affecting both the glottal source and formants, are related to the acoustic interaction between the glottal source and the vocal tract. L_{VT}, A_2, A_3, and A_4, however, do not influence the glottal source, thus having no impact on the interaction. Therefore, parameters k_1, k_c, and A_1 should be estimated together and selected as feature parameters for classification.

The detailed fitting method for estimation of physical parameters is shown in Figure 5.14. This method includes two steps. First, vocal tract fitting is performed with a typical vocal fold setting. The output of this part of the model is the estimated cross-sectional areas of the four-tube model: A_1, A_2, A_3, and A_4. Cost function 1 (C_1) is defined as the root mean square (RMS) distance between envelopes of the simulated and the original speech, and distribution of the first and second formant are also considered. We propose matching the spectral envelope initially in the first iteration, and then, in the second iteration, the characteristics of the formant are fully considered:

$$\begin{aligned} C_1^{(1)} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left|\log P(\omega_i) - \log P^*(\omega_i)\right|^2} \qquad \text{1st iteration} \\ C_1^{(2)} &= \alpha_1\left(F_1^* - F_1\right)^2 + \alpha_2\left(F_2^* - F_2\right)^2 \\ &\quad + w_1(H_1^* - H_1)^2 + w_2(H_2^* - H_2)^2, \qquad \text{2nd iteration} \end{aligned} \tag{5.9}$$

where F_1, F_2, H_1, H_2 refer to the frequency and amplitude of the first and second formant.

In the second step, A_2, A_3, and A_4 are fixed at obtained values, and A_1 is considered as the initial value for the next fitting. In the second fitting, k_1, k_c, and A_1 are selected as control parameters, with the cost function 2 (C_2) in Equation (5.3). Optimal values of the physical parameters are estimated using

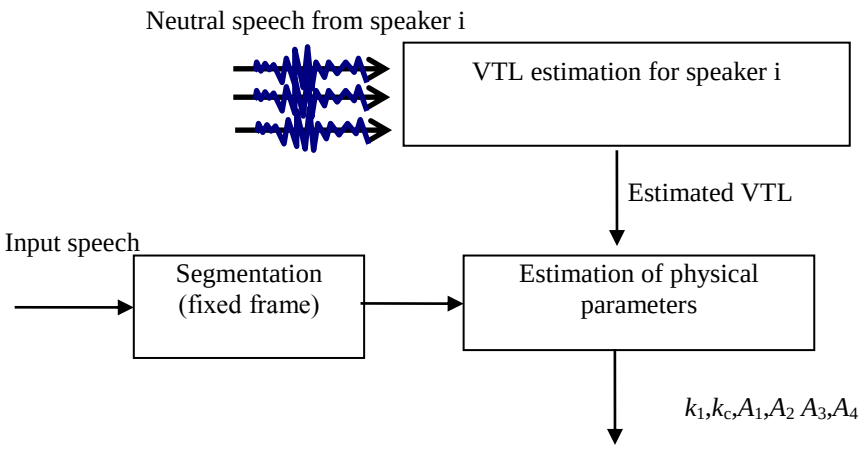


Figure 5.13 Main structure of the method

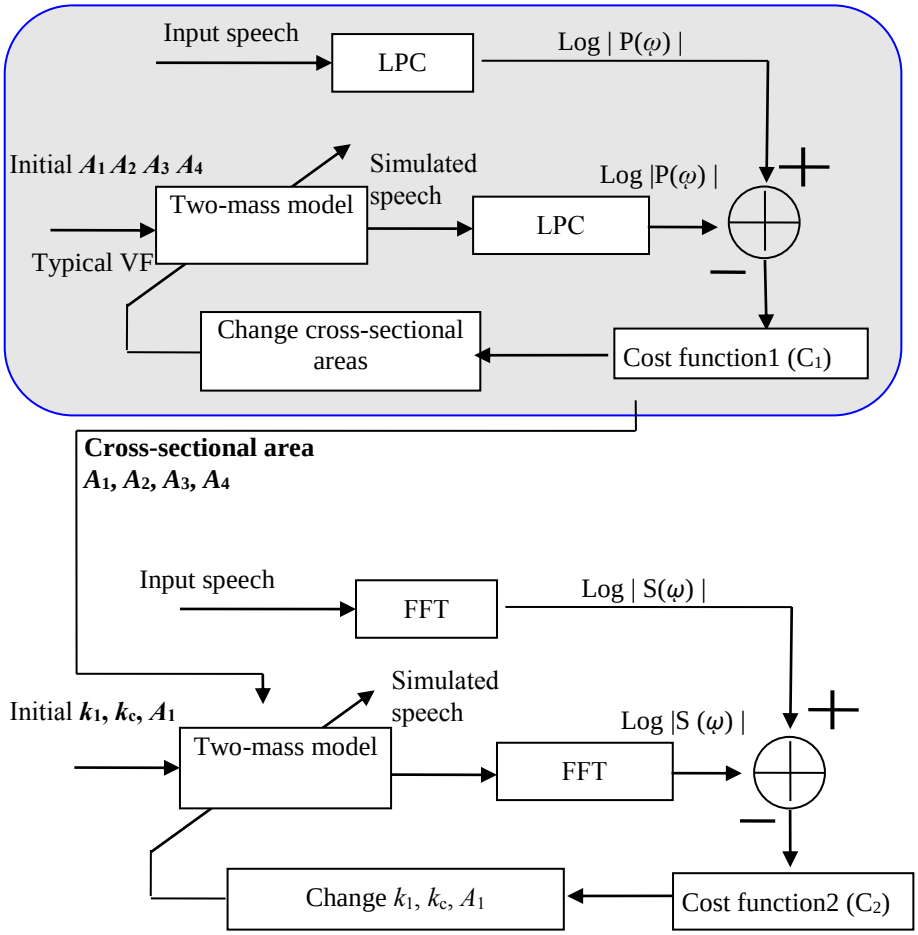


Figure 5.14 Detail of estimation of physical parameters

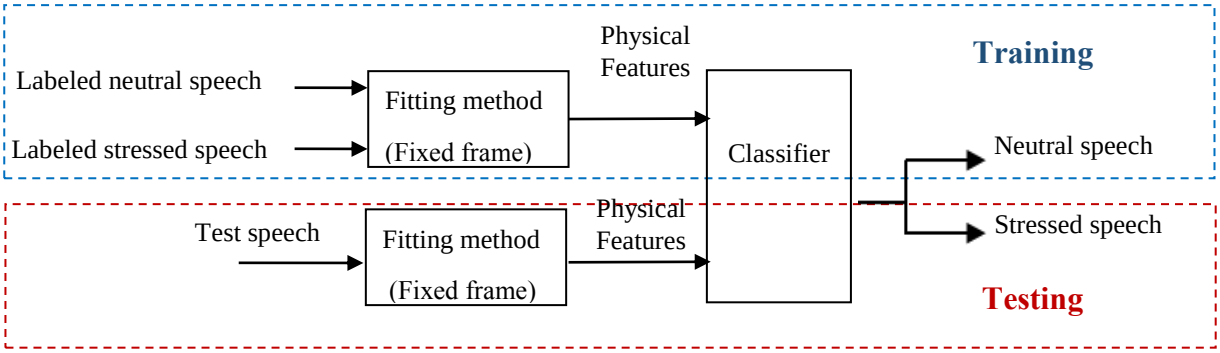


Figure 5.15 Block diagram of our classification method. A linear classifier is used for the training and testing process.

the Nelder-Mead simplex method, which is implemented to search for the optimal stiffness parameters resulting in minimizing the cost function.

5.5 CLASSIFICATION

Evaluation of the physical parameters is speaker dependent. The structure of the classification method is shown in Figure 5.15.

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

During testing, test speech is input into the system and split into frames, and the trained linear

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

$$j^* = argmin\left(\sum_{i=1}^{K} d_i(V_i, a_j)\right) \qquad j = N \quad or \quad S \tag{5.10}$$

5.6 EXPERIMENTAL EVALUATION

5.6.1 DATA SELECTION AND EXPERIMENT SETUP

In the experiments, we used a database collected by the Fujitsu Corporation containing speech samples from eleven subjects (four male and seven female) [110]. To simulate mental pressure resulting in psychological stress, the speakers performed three different tasks while having telephone conversations with an operator, in order to simulate a situation involving pressure during a telephone call. The three tasks involved (A) Concentration; (B) Time pressure; and (C) Risk taking. For each speaker, there are four dialogues with different tasks. In two dialogues, the speaker was asked to finish the tasks within a limited amount of time, and in the other dialogues there is relaxed chat without any task.

All of the data comes from telephone calls, so the sampling frequency was 8 kHz. Vowels /a/ /i/, /u/, /e/ and /o/ were used as samples. The experiments were conducted for each speaker, and all of the results were speaker dependent. The number of samples was different for each speaker. The range of the total number of samples is from 100 to 250 for each vowel, the total amount is about 450-700 for

each person. A K-fold cross-validation method was used in the classification experiments, in which K was set to four. Using this method, the data set was divided evenly into four subsets, and for each classification, one of the subsets was used as a test set and the other three subsets were combined to form a training set. The final result was obtained by calculating the average classification rate across four trials. The samples were analyzed with 12-order LPC and the frame size chosen to perform the experiment was 64ms, with 16ms for frame shift.

For configuration of the two-mass model, the following values were adopted, using the typical values for males: $m_{1M} = 1.25\times10^{-4}$ kg, $m_{2M} = 2.5\times10^{-5}$ kg, $l_{gM} = 0.014$ m, $d_{1M} = 0.0025$ m, $d_{2M} = 5\times10^{-4}$ m, $\zeta_{1M} = 0.1$, $\zeta_{2M} = 0.6$, $x_0 = 2\times10^{-4}$ m, $P_S = 500\ Pa$. The vocal tract model was represented by a tube model, and the number of elements was limited to four cylindrical sections of equal length. Typical values used for configuration for females were as follows: $m_{1F} = 4.56\times10^{-5}$ kg, $m_{2F} = 9.1\times10^{-6}$ kg, $l_{gF} = 0.01$ m, $d_{1F} = 1.79\times10^{-3}$ m, $d_{2F} = 3.6\times10^{-4}$ m, $\zeta_{1F} = 0.1$, $\zeta_{2F} = 0.6$, $x_0 = 2\times10^{-4}$ m, $P_S = 500\ Pa$.

Furthermore, the ranges for the control parameters were, $k_1 : 10-140$ kdyn/cm, $k_2 : 2-14$ kdyn/cm, $k_c : 4-45$ kdyn/cm, VTL: 13-19 cm, $A_1, A_2, A_3, A_4 : 0.2$ - 20 cm^2.

5.6.2 COMPARISON OF COST FUNCTIONS

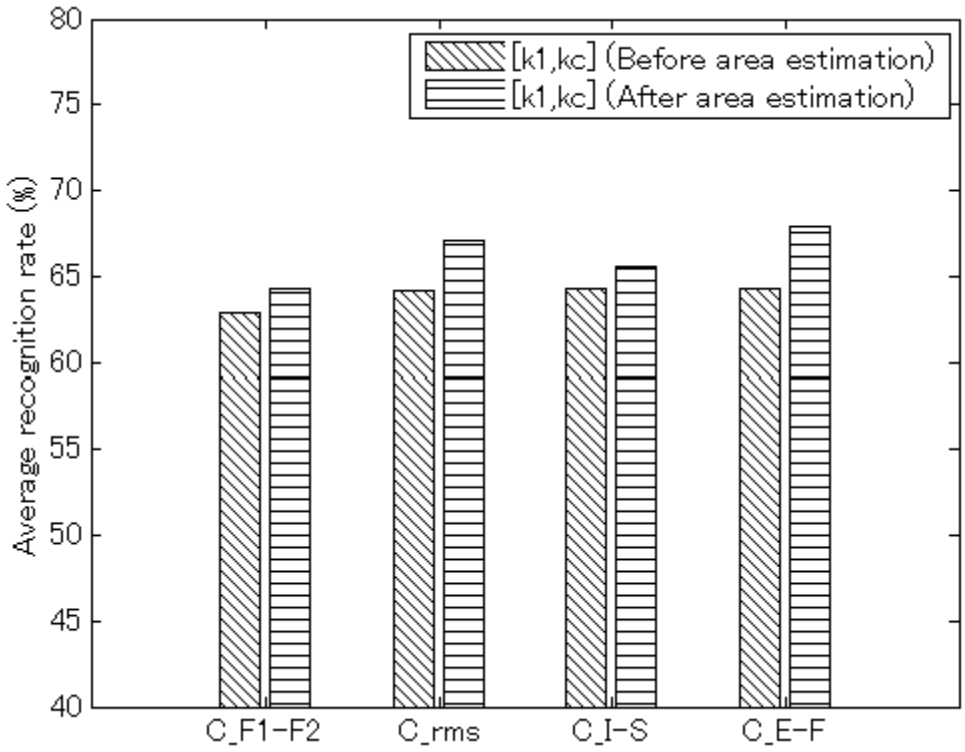


Figure 5.16 Average classification results of four cost functions $C_{F1\text{-}F2}$, C_{rms}, $C_{I\text{-}S}$ and $C_{E\text{-}F}$. The results for varied VF and fixed VT are the classification rate when the stiffness parameters are estimated with fixed VTL and cross-sectional area. Varied VF and varied VT denote that the parameters for stiffness and cross-sectional area are estimated by fitting the two-mass model to real speech.

In the first evaluation, we estimated the vocal tract length of all of the speakers, and two comparisons were made. First, we estimated the cross-sectional area function using the vocal tract fitting method with the four proposed cost functions, and then the shape of the vocal tract was fixed at the obtained values (length and area). We used $[k_1, k_c]$ to check classification performance for neutral and stressed speech using only the cost function for the vocal folds in equation (5.3). In the second comparison, we estimated stiffness parameters $[k_1, k_c]$ with varied vocal tract, so cost functions both for VF and VT were used to perform the fitting, and iteration was performed. Here varied VT denotes that the parameters for cross-sectional area are also estimated by fitting the two-mass model instead of being fixed as constants. Finally, the performance of cost functions $C_{F1\text{-}F2}$, C_{rms}, $C_{I\text{-}S}$ and $C_{E\text{-}F}$ was evaluated

using the classification rate of $[k_1, k_c]$. We used a linear classifier for classification, and the average classification rate for all of the speakers was calculated. The results are shown in Figure 5.16.

The results illustrate that classification performance is improved by nearly 4% when vocal tract values are variable. In this case, the cost functions for the vocal tract are used and formants are also considered, which results in more information about the frequency domain of the speech being available, making the estimated results more reliable. Furthermore, we compared the performance of different cost functions. Our results show that the stress classification rate for $C_{E\text{-}F}$ is higher than for the other cost functions. Since $C_{E\text{-}F}$ can match the rough shape of the spectral envelope, and also effectively catch the characteristics of F_1 and F_2, which have been proven to be sensitive to the interaction between the VF and VT, the classification of stressed speech is improved.

5.6.3 EVALUATION FOR PHYSICAL PARAMETERS

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

- ***Evaluation of vocal tract length (VTL) estimation***

First, a comparison was made to evaluate the vocal tract length estimation for each speaker. In this experiment, vowels /a/, /i/, /u/, /e/ and /o/ were selected as samples. However, the samples for /a/ and /e/ were not mixed together. The two vowels were first used for evaluation separately, and then the

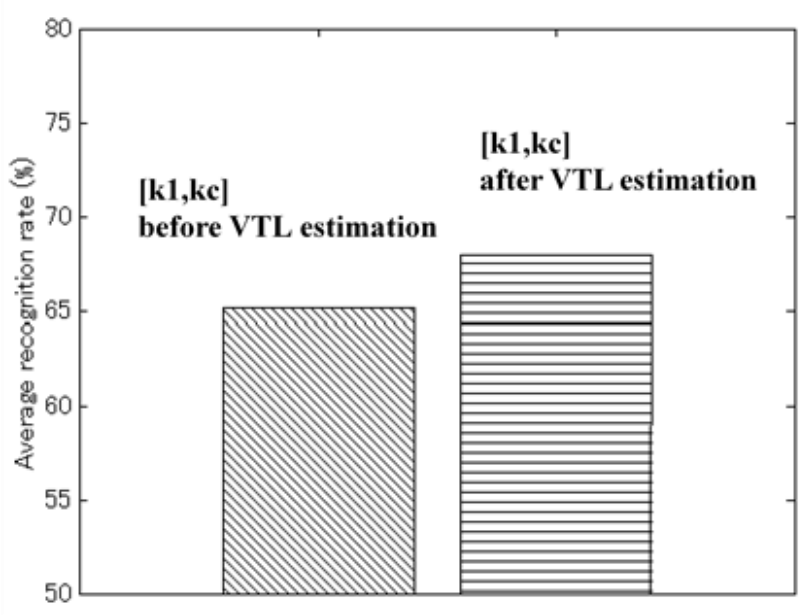


Figure 5.17 Comparison of performance of physical parameters k_1, k_c before and after VTL estimation.

average recognition rate for the two vowels was calculated to show the experimental results. The physical parameters were estimated using the proposed fitting method and the estimated parameters were used as features to perform the stress classification. The evaluation results for VTL estimation are shown in Figure 5.17. Features of physical parameters [k_1, k_c] were compared for their classification performance before and after VTL estimation. Our results show that performance of [k_1, k_c] is improved if the estimation of VTL is performed and average classification rate is increased by 3%. Since a speaker's vocal tract length is calculated from the neutral speech of that specific speaker, and used as a known value for the estimation of other physical parameters, improvement in classification can be achieved by improving the accuracy of

VTL estimation.

- ***Evaluation of stiffness parameters of the vocal folds***

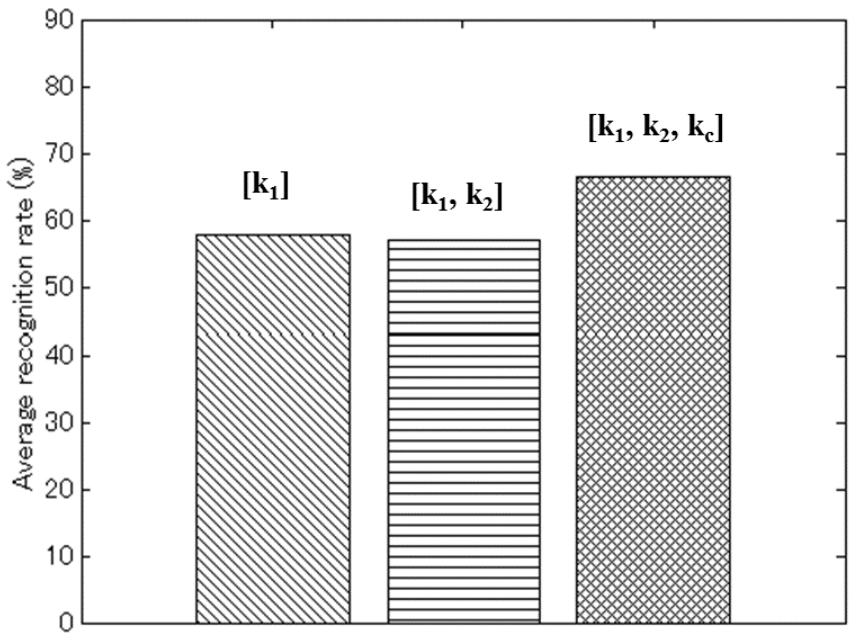


Figure 5.18 Illustration of classification results for physical parameters of the vocal folds. The performance of stiffness parameters k_1, k_c show their effectiveness for stress classification.

In this evaluation, we focused on the stiffness parameters of the vocal folds, and the effect of each stiffness parameter on stress recognition was then examined. The physical parameters $k_1, k_2, k_c, A_1, A_2, A_3, A_4$ were estimated from varied VF and varied VT values with estimated VTL, and other physical parameters were fixed at the typical values described in Section 5.5.1. We focused on the evaluation of k_1, k_2 and k_c. The classification performances of $\{[k_1]\}, \{[k_1, k_2]\}$ and $\{[k_1, k_2, k_c]\}$ for different speakers are shown in Figure 5.18. These results that stress classification performance is improved when k_c is considered. k_1 and k_c therefore, are the parameters which are effective in stress classification. However, average classification accuracy decreases when taking k_2 into account. It suggests that k_2 is not effective in the classification of neutral and stressed speech, therefore it is sufficient to select k_1 and k_c as feature parameters in further evaluations.

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

Next, we focused on each parameter of the cross-sectional area individually, and each area's impact on stress recognition was then examined separately. The parameters $k_1, k_2, k_c, A_1, A_2, A_3, A_4$ were estimated with varied VF and varied VT values. The parameter sets $\{[k_1, k_c]\}, \{[k_1, k_c, A_1]\}, \{[k_1, k_c, A_1, A_2]\}, \{[k_1, k_c, A_1, A_2, A_3]\}$ were also evaluated. Their performance is shown in Figure 18. Among the results, we first consider sets $\{[k_1, k_c]\}$ and $\{[k_1, k_c, A_1]\}$. The results show that stiffness $[k_1, k_c]$ is a better parameter for classifying stressed speech. When A_1 is taken into account, classification performance is further improved. This suggests that A_1 is an important parameter strongly related to stress. When A_1 is increasing, it indicates that the area in the supraglottis is broadening. This results in a decrease in the pressure difference inside and outside of the glottis, causing variation in the airflow pattern and further changes in the interaction around the false vocal folds. Considering the performance of sets $\{[k_1, k_c, A_1]\}, \{[k_1, k_c, A_1, A_2]\}, \{[k_1, k_c, A_1, A_2, A_3]\}$, we found that they have roughly the same classification accuracy. This illustrates that performance cannot be greatly improved by taking A_2 and A_3 into account, and that A_2 and A_3 probably have only a small effect on acoustic interaction. It appears that A_1 is sufficient to classify stressed speech from neutral speech, which agrees with the conclusion of our first evaluation.

A_2 and A_3 do affect F0 to some extent, which was illustrated in Figure 5.4, so they have some influence on acoustic interaction and, further, on stress classification, but we believe their influence is insignificant. The characteristics of the vocal tract also affect stress classification to some extent. Since A_2 and A_3 represent the shape of the vocal tract, $[k_1, k_c, A_1, A_2, A_3]$ can achieve some improvement in the recognition rate, but the increase is very small, which suggests that A_2 and A_3 are less important for stress classification than A_1.

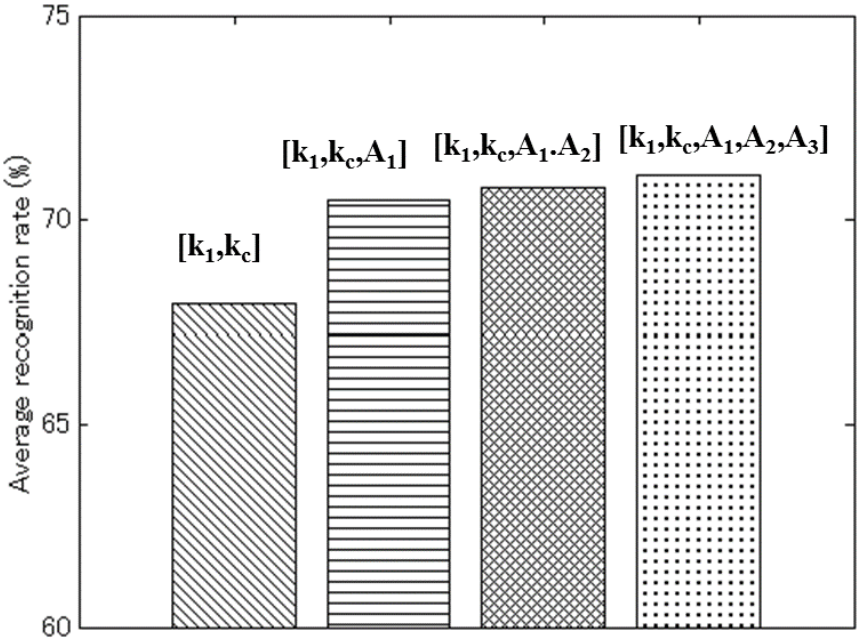


Figure 5.19 Classification results for physical parameters of the vocal tract. The performance of cross-sectional area parameter A_1 shows its effectiveness for stress classification.

5.6.4 EVALUATION FOR PROPOSED FEATURE PARAMETERS

As a result of our evaluation process, parameter set $[k_1, k_c, A_1]$ was proposed. Figure 5.20 shows the distribution results for k_1, k_c, and A_1 with an estimated VTL. These results show that the proposed parameters are effective for stress classification. The estimated values of the parameters are limited in range, and these ranges correspond to the actual range of human beings. As this distribution shows, stiffness and area of the entrance to the vocal tract are good indicators of stressed speech. Under stressed conditions, the value of k_1 becomes relatively large, k_c smaller, and A_1 increases compared with the same parameters under neutral conditions. This indicates that stress causes variation in the muscle tension of the vocal folds, and that the area at the entrance to the vocal tract in the supraglottis becomes wider when the speaker is under stress.

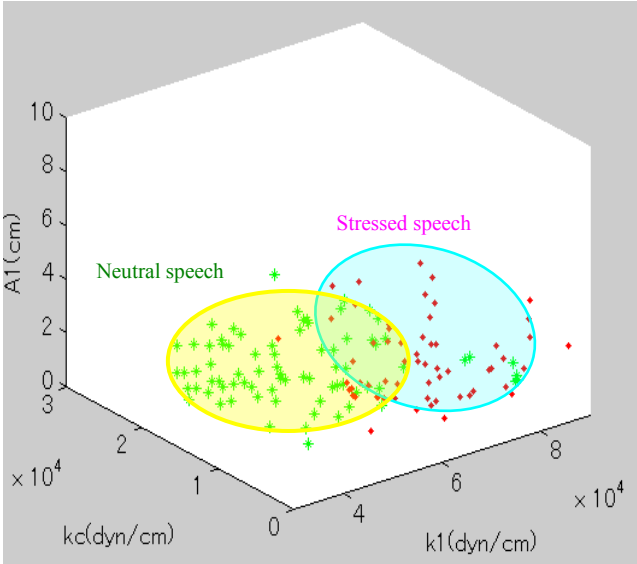


Figure 5.20 Distributions of estimated parameters k_1, k_c and A_1 for neutral and stressed speech.

We then compared the performance of proposed parameters $[k_1, k_c, A_1]$ with traditionally proposed features, namely [SFM, F0], [TEO] and [MFCC]. The results are shown in Figure 5.21. As our experimental results show, [SFM, F0], which characterizes the vocal folds, works well in classifying stressed speech. This shows that the characteristics of the vocal folds play a very important role in stress classification. MFCC, which represents vocal tract information, is also effective for stress classification, illustrating that the characteristics of the vocal tract also affect stress classification to some extent, which agrees with our previous results in Figure 5.19. The results shown in Figure 5.21 demonstrate that our proposed physical parameters outperform the features traditionally used for stress detection, which suggests that parameters estimated from a physical model are more effective at representing stress during phonation than traditional methods. Results show that $[k_1, k_c, A_1]$ has the best stress recognition performance of the physical parameter sets. This illustrates that stiffness of the vocal folds and the cross-sectional area at the entrance to the vocal tract in the supraglottis, are the factors which are most impacted when a speaker is under stress.

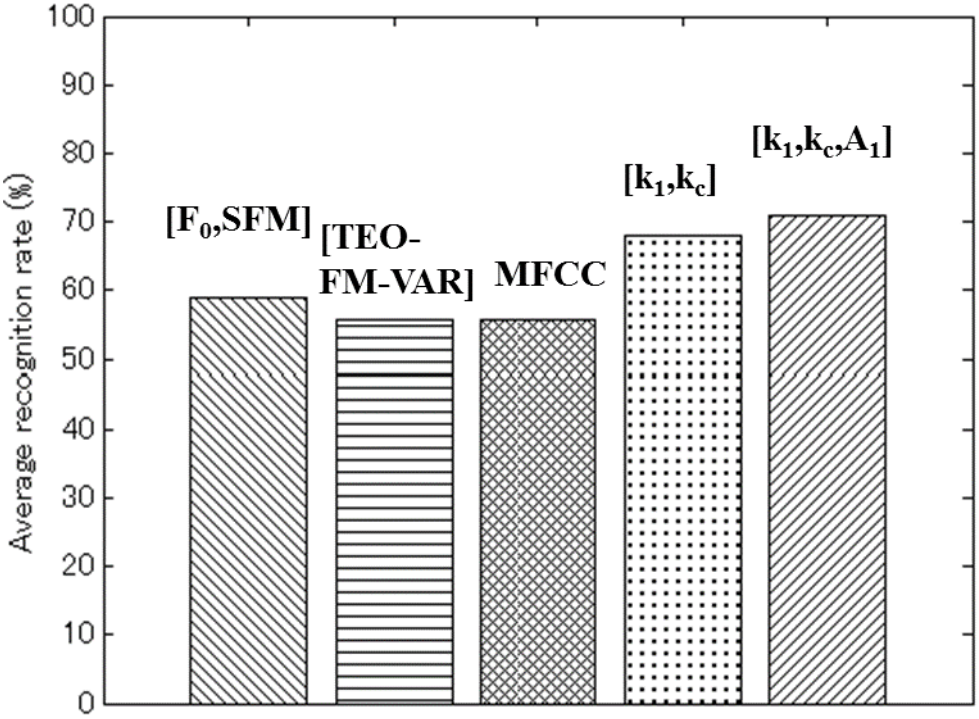


Figure 5.21 Performance of proposed features, compared with traditional methods

5.7 SUMMARY

In this chapter, we explored more effective features for the classification of neutral and stressed speech based on a physical model. To achieve this target, a two-mass model characterizing the properties of the vocal folds and the vocal tract was used to simulate speech production. Physical parameters including, stiffness of the vocal folds, vocal tract length, and cross-sectional area of the vocal tract were investigated and estimated using a method that fits the two-mass model to real data. Cost functions were used as targets to reach more reliable results. The obtained parameters were used as physical features to classify stressed speech. We concluded that two parameters: 1) stiffness of the

vocal folds, and 2) the area at the entrance to the vocal tract in the supraglottis, which is related to the velocity of glottal airflow and acoustic interaction between vocal folds and the vocal tract, are key indicators of stress during phonation. The average performance in the classification of speech under stress was improved by 10%-15% using the proposed features, compared to traditional methods of stressed speech classification.

CHAPTER 6: CLASSIFICATION BY MODELING THE AERODYNAMICS OF LARYNGEAL VENTRICLE

6.1 STRESS CLASSIFICATION BASED ON THE LARYNGEAL VENTRICLE

In the third chapter, we have discussed the aerodynamics in the glottis during the process of speech production. Teager suggested that speech production is a nonlinear process and proposed a nonlinear model. Based on the theory [56][57][79], it is believed that the airflow does not always propagate as a plane wave in the glottis and the vocal tract. The airflow coming from glottis is very unstable when it enter the ventral around the false vocal folds. Airflow separation occurs long the walls of the laryngeal ventricle. Some airflow separates from its body at the outlet of glottis and propagates along the wall of laryngeal ventricle, and then reattach to its body again at the entry into the false vocal folds. The separation will change the effective area into 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.

Cairns showed that the impact of airflow separation differs markedly between neutral and stressed speech [81]. In physiological systems, it is believed that changes in physical characteristics induced by stressful conditions affect airflow separation [61]. 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 previous chapters, we estimated parameters for the vocal folds and the vocal tract, based on a two-mass model [15], 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 chapter, 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 these physical parameters.

This chapter is organized as follows. In Section 6.2, we propose a physical model of the airflow aerodynamics in the ventricle and the false vocal folds. In Section 6.3, an equivalent circuit is proposed to explain the aerodynamics. And a model system for the voiced corresponding to the circuit is presented in Section 6.4. Experiments are performed in Section 6.5 to evaluate the proposed parameters and show their corresponding classification performance for classification of neutral and stressed speech. Conclusions are drawn in Section 6.6.

6.2 MODELING AIRFLOW AERODYNAMICS

The traditional two-mass model was proposed by Ishizaka and Flanagan to simulate the process of speech production [93]. Each vocal fold is represented by two mass-spring-damper systems. The laryngeal part can be depicted using the traditional two-mass model to represent the mechanism of the vocal folds.

$$m_1 \frac{d^2 x_1}{dt^2} + r_1 \frac{dx_1}{dt} + s_1(x_1) + k_c(x_1 - x_2) = F_1, \tag{6.1}$$

$$m_2 \frac{d^2 x_2}{dt^2} + r_2 \frac{dx_2}{dt} + s_2(x_2) + k_c(x_2 - x_1) = F_2, \tag{6.2}$$

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. In this equation, S_i are the equivalent tensions with non-linear relations given by

$$s_i(x_i) = k_i\left(x_i + \eta x_i^{\,3}\right), \tag{6.3}$$

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

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

$$r_1 = 2\zeta_1\sqrt{m_1 k_1} \qquad r_2 = 2\zeta_2\sqrt{m_2 k_2}, \tag{6.4}$$

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

The two-mass model is connected to a four tube model representing the vocal tract [93]. The tube model is constructed using a transmission line analogy involving n cylindrical, hard-walled sections. The elemental values of the model are determined by cross-sectional areas $A_1 \cdots A_n$, and cylinder lengths $l_1 \cdots l_n$. The total length of the vocal tract is defined as L_{VT}. The tube model can be represented by an equivalent circuit, in which the inductances $L_n = \rho l_n / 2A_n$, the capacitances are $C_n = l_n \cdot A_n / \rho c^2$, and the resistances $R_n = \left(S_n / A_n^2\right)\sqrt{\rho\mu\omega/2}$ where c is the velocity of sound. Here, the tube model has been limited to four cylindrical sections of equal length, $n = 4$. In this study, the model is limited to only vowel articulation (as vowels were the subject of the experiments) and modal voice production. These assumptions greatly simplify the modeling of the vocal tract and the glottal source.

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

Figure 6.1 shows a sketch of our traditional model, and the structure of our proposed model is presented in Figure 6.2, in which the laryngeal ventricle and false vocal folds (fvf) are modeled to show the airflow patterns between the vocal folds and the vocal tract.

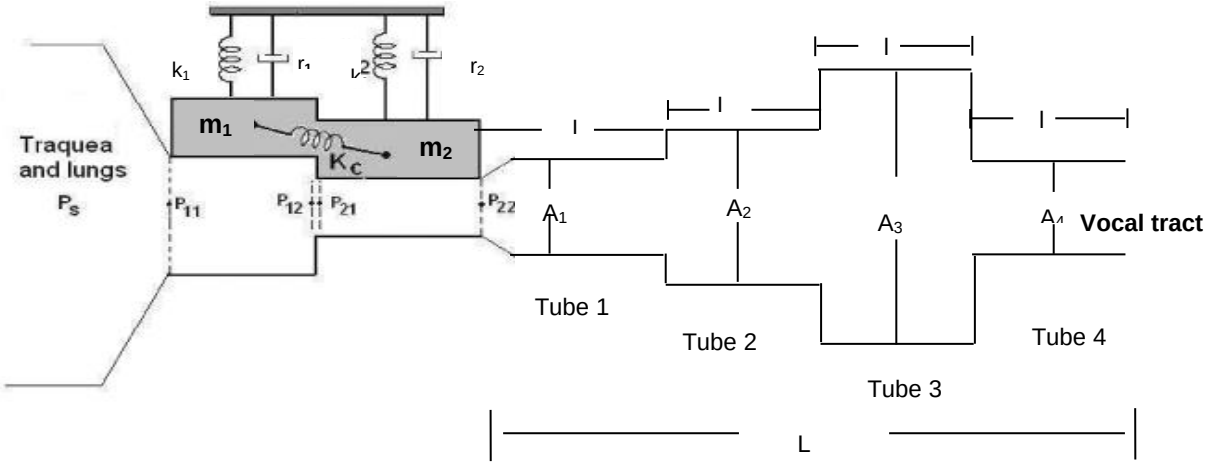


Figure. 6.1 The traditional two-mass model

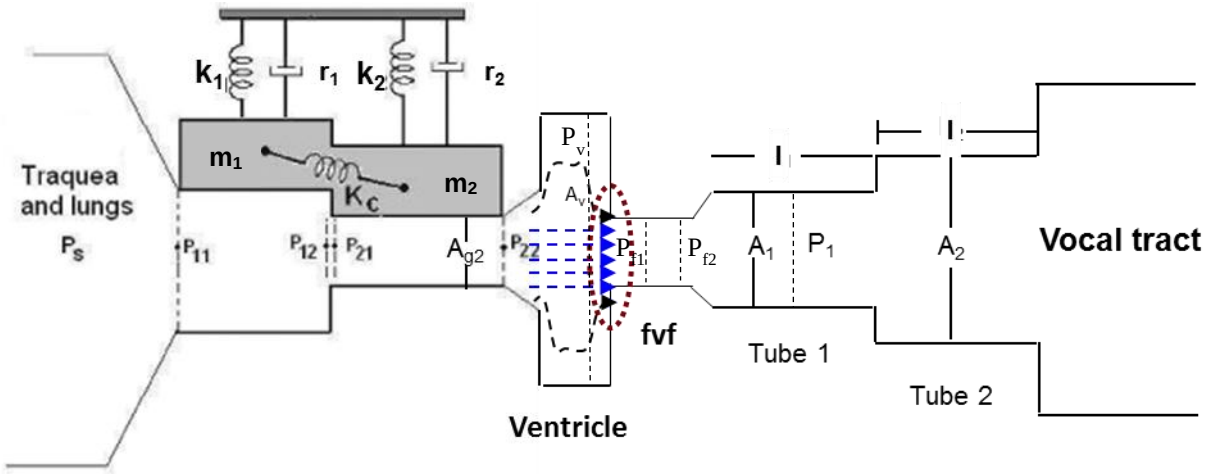


Figure. 6.2 The modified two-mass model

6.2.1 PRESSURE DROP AT THE GLOTTIS

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

$$P_s - P_{11} = \frac{\rho U_g^2}{2A_1^2}, \tag{6.5}$$

where ρ is the air density, and U_g is the volume velocity of glottal airflow. A_{g1} is the cross-sectional lower glottal area, which can be represented 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. 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:

$$P_s - P_{11} = (1.00 + 0.37)\frac{\rho U_g^2}{2A_{g1}^2}, \tag{6.6}$$

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

$$P_{i1} - P_{i2} = \frac{12\mu d_i l_g^2 U_g}{A_{gi}^3}, \qquad i = 1,2 \tag{6.7}$$

where μ is the air viscosity coefficient, and d_1 d_2 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:

$$P_{21} - P_{12} = \frac{\rho U_g^2}{2}\left(\frac{1}{A_{g1}^2} - \frac{1}{A_{g2}^2}\right), \tag{6.8}$$

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

6.2.2 PRESSURE DROP AROUND LARYNGEAL VENTRICLE AND FALSE VOCAL FOLDS

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:

$$P_{22} - P_E = -\frac{\rho}{2}\cdot\frac{2}{A_{g2}A_E}\left(1 - \frac{A_{g2}}{A_E}\right)U_g{}^2, \tag{6.9}$$

where A_E is the area at the entrance to the laryngeal ventricle, and P_v 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 $A_{g2} = A_E$.

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_v 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:

$$P_v - P_{f1} = \frac{\rho}{2}\left(\frac{1}{A_f^2} - \frac{1}{A_v^2}\right)U_g{}^2, \tag{6.10}$$

where A_f is the area of the false vocal folds. Since the false vocal folds do not vibrate during the process of phonation, A_f 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:

$$P_{f1} - P_{f2} = 12\frac{\mu l_f^2 d_f}{A_f^3}U_g, \tag{6.11}$$

where l_f and d_f 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.

$$P_{f2} - P_1 = -\frac{\rho}{2}\cdot\frac{2}{A_f A_1}\left(1 - \frac{A_f}{A_1}\right)U_g{}^2, \tag{6.12}$$

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

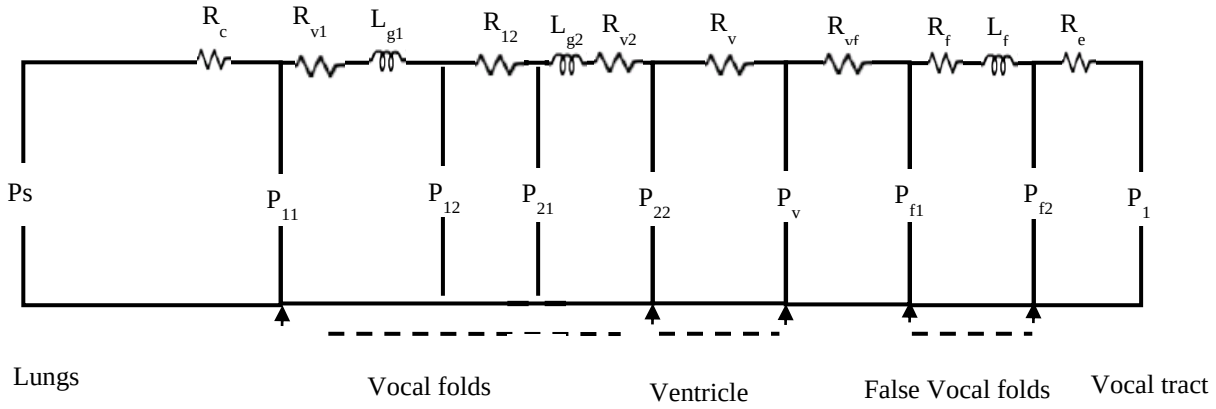


Figure 6.3 Equivalent circuit for the glottis, laryngeal, and false vocal folds.

The effective area of the ventricle into the false vocal tract A_V 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 used as an indicator for stress classification.

6.3 MODEL FOR VOICED SOUND

6.3.1 EQUIVALENT CIRCUIT

According to the pressure difference relations from equations (6.5) – (6.12), an equivalent circuit is generated to represent the relationship with acoustic impedance elements, which is shown in Figure 6.3

In this figure, U_g is the current which is continuous. The elements of the circuit are given as:

$$R_c = 1.37 \frac{\rho}{2} \frac{|U_g|}{A_{g1}^2}$$

$$R_{v1} = 12 \frac{\mu l_g^2 d_1}{A_{g1}^3}$$

$$L_{g1} = \frac{\rho d_1}{A_{g1}}$$

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

$$R_{v2} = 12\frac{\mu l_g^2 d_2}{A_{g2}^3}$$

$$L_{g2} = \frac{\rho d_2}{A_{g2}}$$

$$R_v = -\frac{\rho}{2}\cdot\frac{2}{A_{g2}A_v}\left(1 - \frac{A_{g2}}{A_v}\right)\left|U_g\right| \tag{6.13}$$

$$R_{vf} = \frac{\rho}{2}\left(\frac{1}{A_f^2} - \frac{1}{A_v^2}\right)\left|U_g\right|$$

$$R_f = 12\frac{\mu l_f^2 d_f}{A_f^3}$$

$$L_f = \frac{\rho d_f}{A_f}$$

$$R_e = -\frac{\rho}{2}\cdot\frac{2}{A_f A_1}\left(1 - \frac{A_f}{A_1}\right)\left|U_g\right|,$$

We here calculate the total acoustic impedance of the glottis, ventricle, and false vocal folds, Z_g

$$Z_g = \frac{\rho}{2}\left|U_g\right|\left\{\frac{1.37}{A_{g1}^2}+\frac{1}{A_{g2}^2}-\frac{1}{A_{g1}^2}-\frac{2}{A_{g2}A_v}\left(1-\frac{A_{g2}}{A_v}\right)+\frac{1}{A_f^2}-\frac{1}{A_v^2}-\frac{2}{A_f A_1}\left(1-\frac{A_f}{A_1}\right)\right\}+\left(R_{v1}+R_{v2}\right)$$
$$+\, j\omega\left(L_{g1}+L_{g2}\right)$$

$$=\frac{\rho}{2}\left|U_g\right|\left\{\frac{0.37}{A_{g1}^2}+\frac{1-2\frac{A_{g2}}{A_v}\left(1-\frac{A_{g2}}{A_v}\right)}{A_{g2}^2}+\frac{1}{A_f^2}-\frac{1}{A_v^2}-\frac{2}{A_f A_1}\left(1-\frac{A_f}{A_1}\right)\right\}+\left(R_{v1}+R_{v2}\right)+j\omega\left(L_{g1}+L_{g2}\right)$$

$$=\frac{\rho}{2}\left|U_g\right|\left\{\frac{0.37}{A_{g1}^2}+\frac{1-2\frac{A_{g2}}{A_v}\left(1-\frac{A_{g2}}{A_v}\right)}{A_{g2}^2}-\frac{1}{A_v^2}+\frac{1-2\frac{A_f}{A_1}\left(1-\frac{A_f}{A_1}\right)}{A_f^2}\right\}+\left(R_{v1}+R_{v2}\right)+j\omega\left(L_{g1}+L_{g2}\right)$$

$$=\frac{0.19\rho}{A_{g1}^2}+\frac{\rho\left[0.5-\frac{A_{g2}}{A_v}\left(1-\frac{A_{g2}}{A_v}\right)\right]}{A_{g2}^2}+\left[-\frac{0.5\rho}{A_v^2}\right]+\frac{\rho\left[0.5-\frac{A_f}{A_1}\left(1-\frac{A_f}{A_1}\right)\right]}{A_f^2}+\left(R_{v1}+R_{v2}\right)+j\omega\left(L_{g1}+L_{g2}\right)$$

$$=R_{k1}+R_{k2}+R_{y1}+R_{y2}+\left(R_{v1}+R_{v2}\right)+j\omega\left(L_{g1}+L_{g2}\right),$$

(6.14)

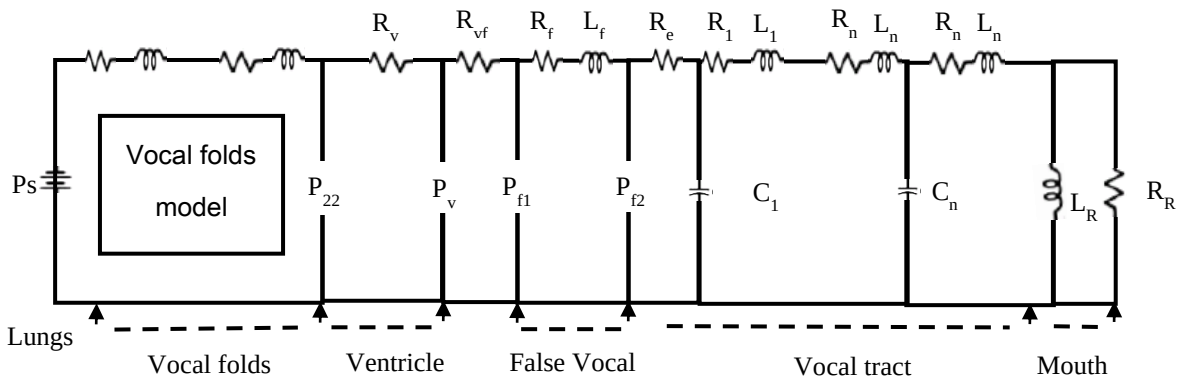


Figure 6.4 Equivalent circuit of model for the synthesis of voiced sounds.

where $R_{k1} = \frac{0.19\rho}{A_{g1}^2}$, $R_{k2} = \frac{\rho\left[0.5 - \frac{A_{g2}}{A_v}\left(1 - \frac{A_{g2}}{A_v}\right)\right]}{A_{g2}^2}$, $R_{y1} = \left[-\frac{0.5\rho}{A_v^2}\right]$, $R_{y2} = \frac{\rho\left[0.5 - \frac{A_f}{A_1}\left(1 - \frac{A_f}{A_1}\right)\right]}{A_f^2}$

6.3.2 NETWORK MODEL FOR SPEECH PRODUCTION

An equivalent circuit representing the system for generating voiced sounds is proposed, which is shown in Figure 6.4. In this figure, the trachea, bronchi, and lungs are neglected. From the left to the right, the vocal folds, laryngeal ventricle, the false vocal folds, and the vocal tract are simulated by the two-mass model. A constant air pressure in the lungs can approximated the subglottal pressure and the vocal tract is simulated using a transmission line analogy involving n cylindrical, hard-walled sections. The elemental values of the model are determined by cross-sectional areas $A_1 \cdots A_n$, and cylinder lengths $l_1 \cdots l_n$. The total length of the vocal tract is defined as L_{VT}. The tube model can be represented by an equivalent circuit, in which the inductances $L_n = \rho l_n / 2A_n$, the capacitances are $C_n = l_n \cdot A_n / \rho c^2$, and the resistances $R_n = \left(S_n / A_n^2\right)\sqrt{\rho\mu\omega/2}$ where c is the velocity of sound. Here, the tube model has been limited to four cylindrical sections of equal length, $n = 4$. The model is terminated in a radiation load equal to that of a circular piston in an infinite baffle. $L_n = \left(8\rho/3\pi\right)\sqrt{\pi A_n}$, $R_R = 128\rho c / 9\pi^2 A_n$, where A_n is the area of the mouth.

Considering the Figure 6.4, we can propose the differential equations related to the volume velocities of the system according to the equivalent circuit:

$$R_c U_g + R_{v1} U_g + L_{g1}\frac{dU_g}{dt} + R_{12} U_g + R_{v2} U_g + L_{g2}\frac{dU_g}{dt} + R_E U_g$$

$$R_v U_g + R_{vf} U_g + R_f U_g + L_f \frac{dU_g}{dt} + R_e U_g + L_1 \frac{dU_g}{dt} + R_1 U_g + \frac{1}{C_1}\int_0^t \left(U_g - U_1\right) dt - P_s = 0$$

$$
\begin{aligned}
&(L_1 + L_2)\frac{dU_1}{dt} + (R_1 + R_2)U_1 + \\
&\frac{1}{C_2}\int_0^t (U_1 - U_2)dt + \frac{1}{C_1}\int_0^t (U_1 - U_g)dt = 0, \\
&(L_2 + L_3)\frac{dU_2}{dt} + (R_2 + R_3)U_2 + \\
&\frac{1}{C_3}\int_0^t (U_2 - U_3)dt + \frac{1}{C_2}\int_0^t (U_2 - U_1)dt = 0, \\
&(L_3 + L_4)\frac{dU_3}{dt} + (R_3 + R_4)U_3 + \\
&\frac{1}{C_4}\int_0^t (U_3 - U_L)dt + \frac{1}{C_3}\int_0^t (U_3 - U_2)dt = 0, \\
&(L_4 + L_R)\frac{dU_L}{dt} + R_4 U_L - L_R \frac{d(U_R)}{dt} \\
&\frac{1}{C_4}\int_0^t (U_L - U_3)dt = 0, \\
&L_R \frac{d}{dt}(U_R + U_L) + R_R \cdot U_R = 0
\end{aligned}
\tag{6.15}
$$

6.4 ESTIMATION METHOD

Fitting the model to real speech poses a difficulty because the existence of interaction makes it impossible to fit the vocal folds (VF) and vocal tract (VT) separately. Based on the pressure distribution discussed above, it is believed that stiffness parameters k_1, k_c and cross-sectional areas A_V, A_1, determining volume velocity U_g, are related to the acoustic interaction between the glottal source and the vocal tract. A_2, A_3 and A_4, however, are not directly related to the glottal source, and thus have less impact on the interaction, as we showed in chapter 5. Therefore, parameters k_1, k_c, A_1 and A_V, should be estimated together and selected as feature parameters for stress classification.

The detailed fitting method for estimation of the physical parameters is shown in Figure 6.5. This method includes two steps. First, vocal tract fitting is performed with a typical vocal fold setting. The

outputs of this part of the model are the estimated cross-sectional areas of the four-tube model: A_1, A_2, A_3, and A_4. Cost function 1 (C_1) is defined as the root mean square (RMS) distance between envelopes of the simulated and the original speech, and distribution of the first and second formant are also considered. We propose matching the spectral envelope initially in the first iteration, and then, in the second iteration, the characteristics of the formant are fully considered:

$$
\begin{aligned}
C_1^{(1)} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left|\log P(\omega_i) - \log P^*(\omega_i)\right|^2} && \text{1st iteration} \\
C_1^{(2)} &= \alpha_1\left(F_1^* - F_1\right)^2 + \alpha_2\left(F_2^* - F_2\right)^2 \\
&\quad + w_1(H_1^* - H_1)^2 + w_2(H_2^* - H_2)^2, && \text{2nd iteration}
\end{aligned}
\tag{6.16}
$$

In the second step, A_2, A_3, and A_4 are fixed at obtained values, and A_1 is considered as the initial value for the next fitting. In the second fitting, k_1, k_c, A_V and A_1 are selected as control parameters, and cost function 2 is defined as:

$$
C_2 = \frac{1}{N}\sum_{i=1}^{N} \left|\log S(\omega_i) - \log S^*(\omega_i)\right|^2, \tag{6.17}
$$

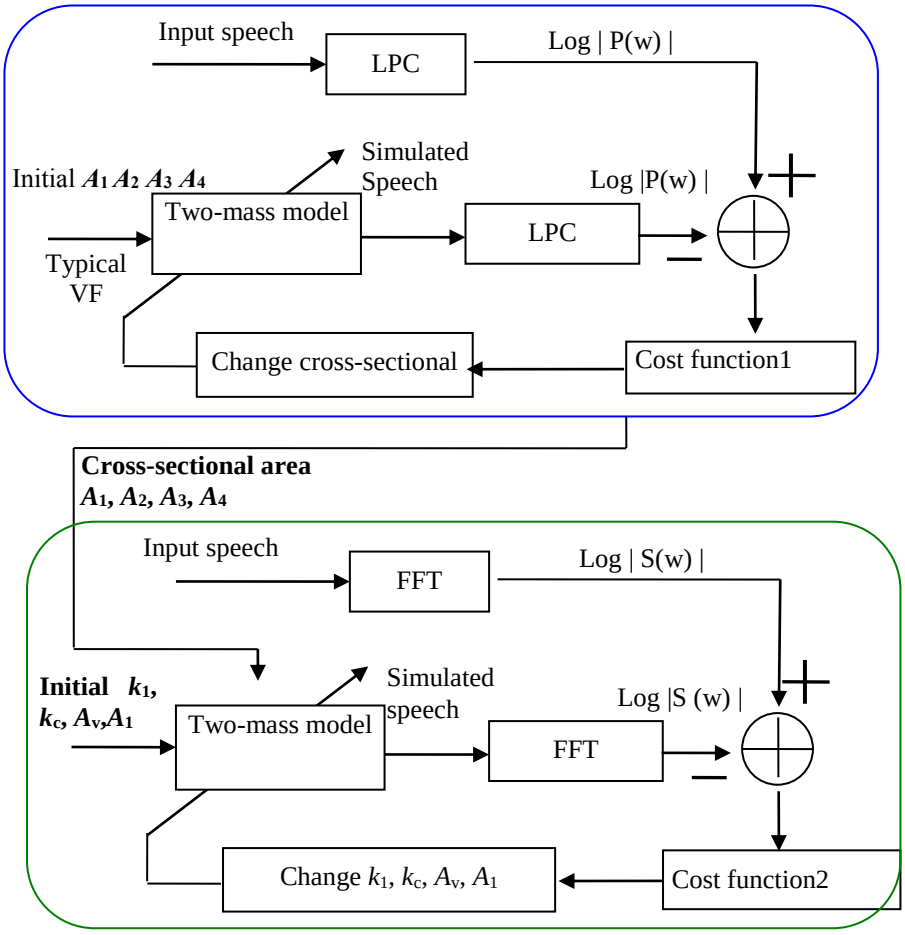


Figure 6.5 Method for estimation of physical parameters

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, which is implemented to search for the optimal stiffness parameters which minimize the cost function.

6.5 EXPERIMENTAL EVALUATION

6.5.1 CONFIGURATION FOR MALE AND FEMALE SPEAKER

For the configuration of the two-mass model, the following values were adopted, using typical values for males: $m_{1M} = 1.25\times10^{-4}$ kg , $m_{2M} = 2.5\times10^{-5}$ kg , $l_{gM} = 0.014$ m , $d_{1M} = 0.0025$ m , $d_{2M} = 5\times10^{-4}$ m , $\zeta_{2M} = 0.6$, and $x_0 = 2\times10^{-4}$ m . The vocal tract model was represented by a standard tube configuration for the vowel /a/ [94], 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 $A_1 = 8\times10^{-5}$ m^2, $A_2 = 4\times10^{-5}$ m^2, $A_3 = 3\times10^{-4}$ m^2, and $A_4 = 8\times10^{-4}$ m^2 . For the configuration for females, the typical values were as follows: $m_{1F} = 4.56\times10^{-5}$ kg , $m_{2F} = 9.1\times10^{-6}$ kg , $l_{gF} = 0.01$ m , $d_{1F} = 1.79\times10^{-3}$ m , $d_{2F} = 3.6\times10^{-4}$ m , $\zeta_{2F} = 0.6$, $x_0 = 2\times10^{-4}$ 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 $A_1 = 4.85\times10^{-5}$ m^2 , $A_2 = 2.4\times10^{-5}$ m^2 , $A_3 = 1.8\times10^{-4}$ m^2, and $A_4 = 4.85\times10^{-4}$ m^2.

As for the false vocal folds, the typical values for the length and the thickness are set to $l_f = 0.008$ m , $d_f = 0.0025$ m. Since the false vocal folds do not vibrate during the phonation, the cross-sectional area between the two false vocal folds is a constant. So the gap separating the false vocal folds is set ranging from 0.0023 m to 0.0083 m for female, and from 0.004 m to 0.0068 m for male. We here choose 0.006 m and 0.0055 m for female and male speakers. Hence, the cross-sectional area of the false vocal folds A_f can be determined.

6.5.2 DATABASE AND EXPERIMENTAL SETUP

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

The segments with the Japanese vowels /a/, /i/, /u/, /e/, /o/ were used as samples. The experiments were conducted for each speaker, and all of the results were speaker dependent. Here, we used samples from the eleven subjects to show the average classification performance. 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.

6.5.3 EVALUATION OF PHYSICAL PARAMETERS

6.5.3.1 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 (F_1, F_2, F_3), 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 6.6 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.

$$\mathrm{D} = \sqrt{\frac{1}{f(b)} \sum_{\omega_i \in B(b)} \left(10\log_{10}\left|S^*(\omega_i)\right| - 10\log_{10}\left|S(\omega_i)\right|\right)^2}, \tag{6.18}$$

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. $S(\omega)$ and $S^*(\omega)$ are the power spectrums of the residual signals for simulated and real speech, respectively. The sub-bands are described in Table 6.1.

Table 6.1 Sub-bands for the spectrum

	Sub-Band						
	1	2	3	4	5	6	7
Frequency band(Hz)	0-1000 (Hz)	500-1500 (Hz)	1000-2000 (Hz)	1500-2500 (Hz)	2000-3000 (Hz)	2500-3500 (Hz)	3000-4000 (Hz)

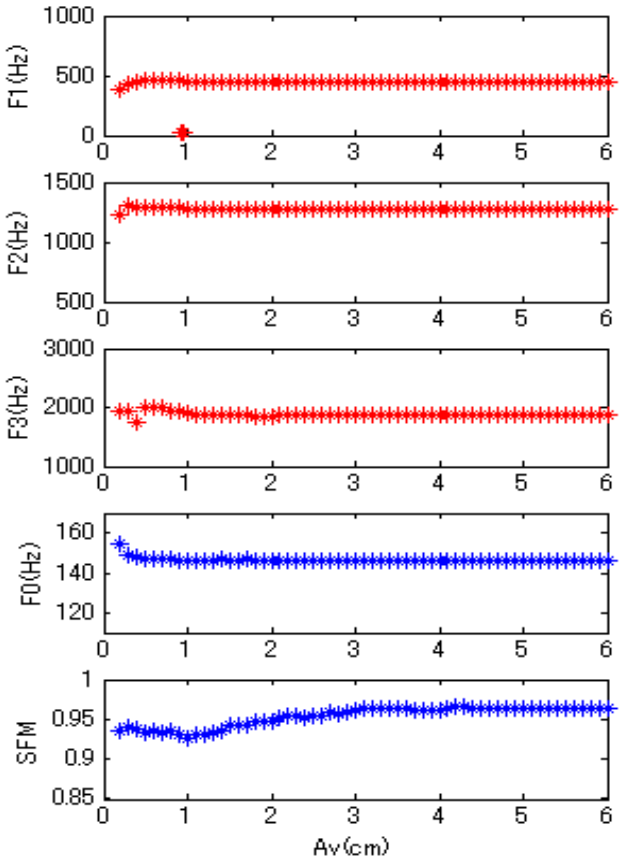


Figure 6.6 Impact of A_V on acoustic parameters

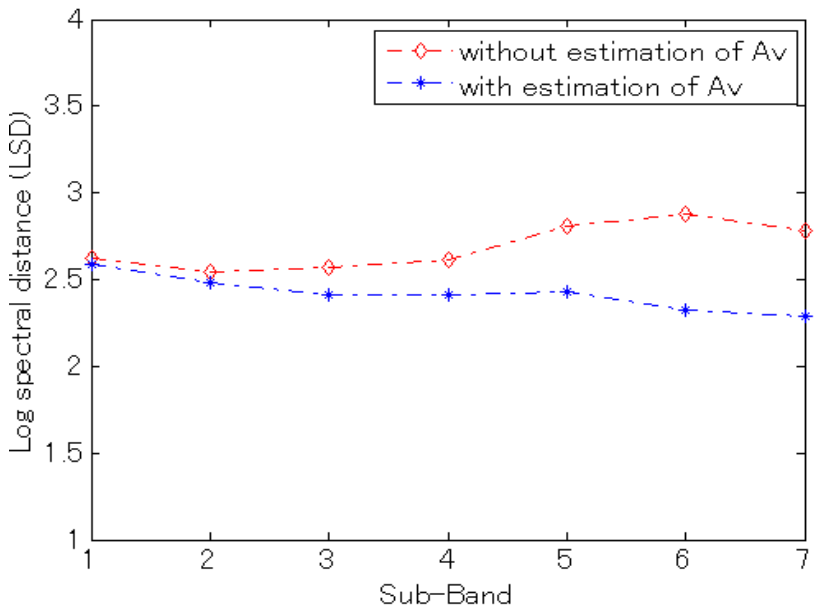


Figure 6.7 LSD to evaluate impact of A_V on spectrum simulation

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 6.7, 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 A_V, spectral distortion decreases significantly. This indicates that the estimation of A_V can improve simulation accuracy in the high frequency bands.

6.5.3.2 Evaluation of physical parameters

- Evaluation under vowel dependent condition

In this section, we compared the performance of physical parameter sets, [k_1, k_c], [k_1, k_c, A_1] and [k_1, k_c, A_1, A_V], to evaluate the effectiveness of proposed parameter A_V. Samples of the individual vowels /a/, /i/, /u/, /e/, /o/ were selected respectively for vowel-dependent experiments, and the average classification rate was then calculated. The parameters k_1, k_c, A_1, and A_V were estimated from real speech with the obtained vocal tract length for each speaker. Figure 6.8 compares the classification rates of parameter sets [k_1, k_c], [k_1, k_c, A_1], [k_1, k_c, A_V] and [k_1, k_c, A_1, A_V] Comparing these results, we can see that parameter sets A_1 and A_V achieve better performance under the vowel dependent condition, in which individual vowels are considered separately. The performance of [k_1, k_c] is improved by 5% when A_V is considered because A_V represents the airflow variations in the laryngeal ventricle. A_1 is also effective for stress detection because, when considering only one vowel, the shape of the vocal tract does not change significantly, so A_1 only represents acoustic interaction, thus improving classification performance.

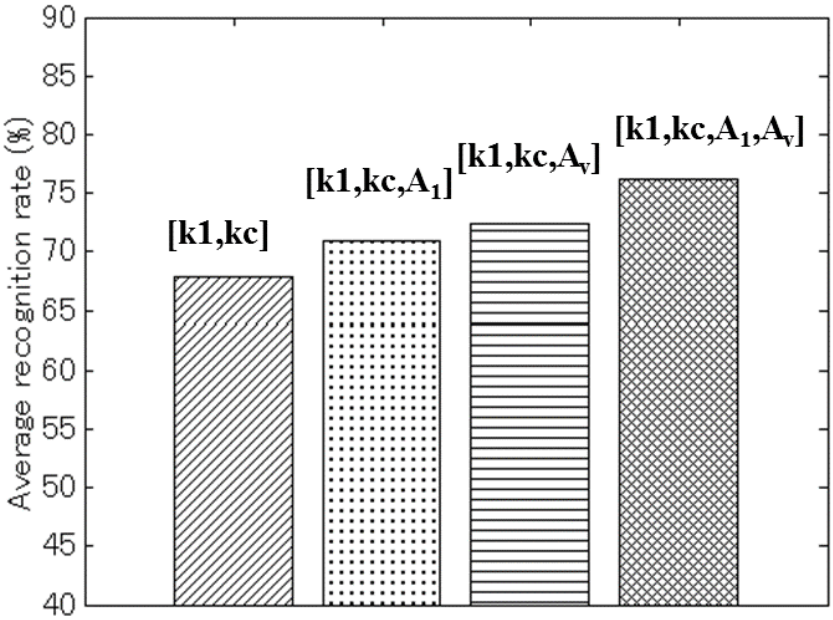


Figure 6.8 Evaluation under vowel-dependent condition.

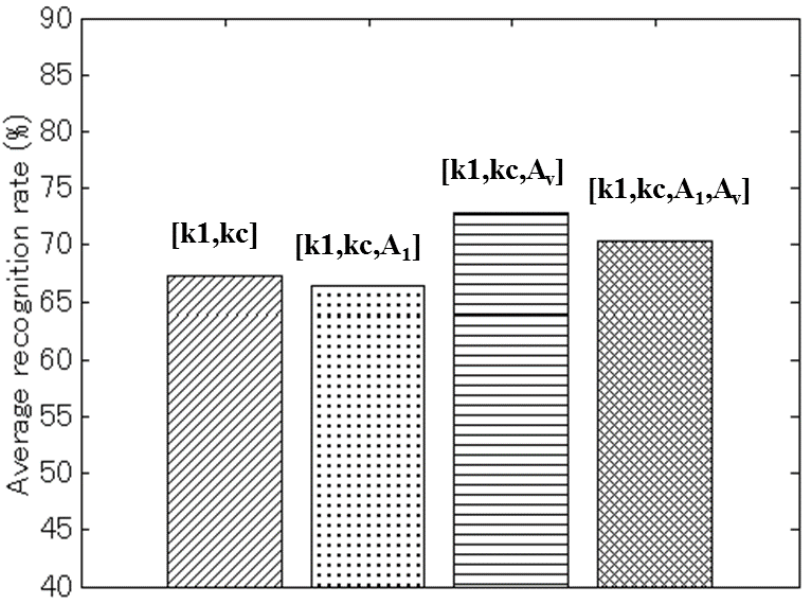


Figure 6.9 Evaluation under vowel-independent condition.

- Evaluation under vowel-independent condition

In this evaluation, speech 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. Experiments were conducted for each speaker, and all of the results were speaker dependent.

Results show that the classification performance decreases when A_1 is considered. A_1 determines the shape of the vocal tract, so it is not effective under the vowel-independent condition. When A_V is taken into account, classification performance is improved. The results are shown in Figure 6.9. Since the samples selected in the experiment are mixed data from all the vowels, the results show that A_V 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 A_V is an essential parameter strongly related to stress. Larger A_V 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.

6.6 SUMMARY

In this chapter, 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 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.

CHAPTER 7: COMPARISON OF PERFORMANCE USING DIFFERENT CLASSIFIERS

7.1 INTRODUCTION

Researchers have been focusing on the classification of speech under stress. In recent years, much research has been performed with the objective of automatically recognizing stressed speech from the neutral speech. Speech recognition systems are modeled using statistical approaches, such as Hidden Markov Models (HMM), Artificial Neutral Networks (ANN), Gaussian Mixture Models (GMM) and Bayesian classifiers, and these are also the methods usually applied for stress classification.

Busch used a Bayesian classifier to separate neutral, loud, angry, and Lombard speech [69]. Hansen investigated features of speech based on Mel-frequency cepstrum coefficients (MFCC), and stress classification was performed using the separability of distance metrics and a neutral network classifier [63]. HMMs can model acoustic and temporal variability in speech signals using a Markov process. Gaussian Mixture Models (GMM) are used to model the state of an HMM representing a probability distribution, because GMMs are an effective parametric model for generating robust mathematical frameworks. In 2003, Schuller propose a classification method using an HMM with instantaneous features and a GMM with global statistical properties [111], but its shortcomings were limited energy and pitch. In another study, the use of a combining algorithm and an N-dimensional

Hidden Markov Model (HMM) was proposed for stressed speech classification [70]. Results showed that stress classification can improve the robustness of speech recognition systems. Bou-Ghazale and Hansen developed perturbation models of neutral-to-stressed speech using an HMM framework [71]. Their model could simulate different speaking styles by perturbing the neutral training data. Their results showed that classification performance can be improved using this technique. An additional method using a conventional approach for re-training reference models has been put forward to improve the robustness of speech recognition systems [72].

Artificial neural networks (ANN) have been successfully applied to improve speech recognition performance [112] [113], solving many problems, but shortcomings still exist. For example, convergence is too slow, optimal topologies need to be covered, and they are easy to overfit [114] [115]. Recently, methods have been proposed to balance overfitting and improve classification performance. Yoshitomi developed a framework combining an ANN and a Hidden Markov Model (HMM) to classify emotional speech and facial expressions, which achieved some improvement [116].

Support Vector Machines (SVM) construct N-dimensional hyperplanes to classify test samples into two classes [117]. It is a classifier which estimates decision surfaces directly rather than modeling a probability distribution from the training data. SVMs often achieve better classification performance than other traditional classifiers because they require a smaller amount of training data. In 2001, Yu proposed a Support Vector Machine (SVM) based emotion recognition system, which achieved better recognition performance [118]. Other works have introduced an SVM combined with a GMM for emotional speech recognition. MFCC features extracted from speech are used to train a GMM, and then the GMM's feature vector is generated as an input for the SVM [119].

In 1987, an extended method called a multi-style training model was first used. However, during evaluation it was shown that multi-style training does not work well under the speaker-independent condition when there is a limited amount of training data [73]. Sanjay used a reliable GMM framework to model the TEO feature TEO-CB-Auto-Env for stress classification. In this method, the training data size and number of mixtures were pre-determined to achieve the best classification performance [74].

In our previous chapters, we described how linear classifiers have become the dominant application for classifying stressed speech. In this chapter, speech features are modeled using GMMs and SVMs as statistical classifiers, and the stress classification performance of various speech features is measured. This chapter is organized as follows. In Section 7.2, physical features of speech production are modeled using a Gaussian Mixture Model (GMM). We then perform stress classification based on a Support Vector Machine (SVM) in Section 7.3. Experiments are performed in Section 7.4 to evaluate different classifiers and methods, and their corresponding stress classification performances are then shown. Conclusions are drawn in Section 7.5.

7.2 CLASSIFICATION BASED ON GAUSSIAN MIXTURE MODEL (GMM)

A Gaussian Mixture Model (GMM) is a parametric probability density function represented by a weighted sum of the density of a Gaussian component, given by the equation:

$$p(\mathbf{x}|\boldsymbol{\lambda}) = \sum_{i=1}^{M} \omega_i g(\mathbf{x}|\boldsymbol{\mu}_\mathbf{i}, \Sigma_i), \tag{7.1}$$

where x is a D-dimensional vector, and $\boldsymbol{g}(\boldsymbol{x}|\boldsymbol{\mu_i}, \Sigma_{\boldsymbol{i}})$ is the component Gaussian density, with ω_i representing the mixture weight. Each component follows the Gaussian distribution:

$$g(\mathbf{x}|\boldsymbol{\mu}_\mathbf{i}, \boldsymbol{\Sigma}_\mathbf{i}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_\mathbf{i}|^{D/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{i})^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{i})\right\} \tag{7.2}$$

with mean μ_i and covariance Σ_i. The weights satisfy the constraint that $\sum_{i=1}^{M} \omega_i = 1$

The completed GMM is composed of the mean vector, covariance matrices and mixture weights for each component. Parameters are represented by the equation:

$$\lambda = \{\boldsymbol{\omega}_\mathbf{i}, \boldsymbol{\mu}_\mathbf{i}, \boldsymbol{\Sigma}_\mathbf{i}\} \qquad i = 1, \ldots\ldots, M \tag{7.3}$$

The GMM is trained using the given training data to estimate parameters in the model, according to some pre-determined criteria. A well-known estimation method is Maximum Likelihood (ML). Given a sequence of training data $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{\boldsymbol{T}}\}$, the likelihood of a GMM can be represented by:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} p(\mathbf{x}_\mathbf{i}|\lambda), \tag{7.4}$$

It is hard to calculate the maximum because the expression is a non-linear function of the parameter λ. Another well-known estimation method, called expectation-maximization (EM), can also be used for parameter estimation. Using this method, parameters are estimated iteratively. The principle of the EM method is to estimate a new model λ' with an initial model λ, which makes $p(\mathbf{x}|\lambda') \geq p(\mathbf{x}|\lambda)$. The new model can be taken as the initial model for the next iteration. Iterations are performed and

repeated until a convergence is achieved. The estimation formula for each EM iteration can be represented as a:

- Mixture weight

$$\boldsymbol{\omega}_{\mathbf{i}}^{'} = \frac{1}{T}\sum_{t=1}^{T}\Pr\left(i\middle|\mathbf{x_t},\lambda\right), \tag{7.5}$$

- Mean

$$\boldsymbol{\mu}_{\mathbf{i}}^{'} = \frac{\sum_{t=1}^{T}\Pr\left(i\middle|\mathbf{x_t},\lambda\right)x_t}{\sum_{t=1}^{T}\Pr\left(i\middle|\mathbf{x_t},\lambda\right)}, \tag{7.6}$$

- Variance

$$\boldsymbol{\sigma}_{\mathbf{i}}^{'\mathbf{2}} = \frac{\sum_{t=1}^{T}\Pr\left(i\middle|\mathbf{x_t},\lambda\right)\mathbf{x}_t^2}{\sum_{t=1}^{T}\Pr\left(i\middle|\mathbf{x_t},\lambda\right)} - \mu_i^{'2}, \tag{7.7}$$

The a posteriori probability for component *i* is given by:

$$\Pr\left(i\middle|\mathbf{x_t},\lambda\right) = \frac{\omega_i g\left(\mathbf{x_t}\middle|\boldsymbol{\mu}_\mathbf{i},\boldsymbol{\Sigma}_\mathbf{i}\right)}{\sum_{k=1}^{M}\omega_k g\left(\mathbf{x_t}\middle|\boldsymbol{\mu}_\mathbf{k},\boldsymbol{\Sigma}_\mathbf{k}\right)}, \tag{7.8}$$

We modeled physical speech features using Gaussian Mixture Models. Two GMMs, one for neutral and one for stressed speech, were trained. In order to increase the amount of training data, the GMMs were trained using the training data of the three male speakers, and the testing set was generated using the testing data of the three male speakers and all of the data from the four female speakers.

7.3 CLASSIFICATION BASED ON SUPPORT VECTOR MACHINE (SVM)

A Support Vector Machine (SVM) is a supervised statistical learning method based on the Vapnik-Chervonenkis (VC) dimension introduced by Vladumir Vapnik and Alexey Chervonenkis of Bell Lab [120].

The basic principle of SVM is to map the input space to a high-dimensional space through non-linear transformation and thus achieve the optimal linear classification plane according to minimal structure criterion in the new space. This non-linear transformation is completed by defining the proper inner-product function.

Suppose the linear separable sample set $(\mathbf{x_i}, \mathbf{y_i}), i = 1...n \quad x \in R^n, y \in \{+1, -1\}$, and y are class labels). The general form of linear discriminant functions is $g(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + \mathbf{b}$ and the classification plane equation is $\mathbf{w} \cdot \mathbf{x} + \mathbf{b} = 0$. We then normalize the discriminant function for all the samples in the two classes to make $\|g(\mathbf{x})\| \geq 1$, so that for the nearest sample to the classification plane, $\|g(\mathbf{x})\| = 1$, making the classification margin equal $2/\|w\|$. Requiring all samples to be correctly classified by the classifier, this becomes:

$$y_i[(\mathbf{w} \cdot \mathbf{x}) + \mathbf{b}] - 1 \geq 0, i = 1,2...n, \tag{7.9}$$

So the classification plane which meets the above mentioned conditions and achieves the minimum $\|w\|$ is the optimal classification plane.

The optimal classification plane problem can be considered as solving the constrained optimization problem restricted by Equation (7.9); thus the minimum value for the function is:

$$\phi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\| = \frac{1}{2}(\mathbf{w}\cdot\mathbf{w}), \tag{7.10}$$

This can be reduced to the following optimization problem, which can be resolved using the Lagrange Multiplier Method:

$$\max_{\alpha_i} \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j), \tag{7.11}$$

subject to $0 \le \alpha_i \le C$ and to the constraint $\sum_{i=1}^{l} a_i y_i = 0$

For non-linear problems, all the points should be considered to be mapped into the high-dimensional space in order to be linearly separable in this space, and only inner-product is calculated in the high-dimensional space. There is no need to know the non-linear transformation form, so by avoiding complicated calculation in high-dimensional transformations, the problem is greatly simplified. According to the Hilbert-Schmidt principle, kernel function selected can be used as the inner-product function only if one operation can satisfy the Mercer condition. So, the inner-product should be defined in the high-dimension space after mapping, to construct the maximum-margin hyperplane. In the process of mapping into the high-dimension space, the inner-product is replaced by a nonlinear kernel function:

$$\max_{\alpha_i} \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j k(\mathbf{x_i}, \mathbf{x_j}), \tag{7.12}$$

Some kernel functions include:

(1) Polynomials: Polynomial mapping is a popular method for non-linear modeling. The second kernel is usually preferable as it avoids problems with the hessian becoming zero.

$$k(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i} \cdot \mathbf{x_j})^d, \; k(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i} \cdot \mathbf{x_j} + 1)^d, \tag{7.13}$$

(2) Gaussian Radial Basis Functions: Radial basis functions most commonly assume a Gaussian form, where σ denotes the width of the RBF kernel.

$$k(\mathbf{x_i}, \mathbf{x_j}) = \exp\left(-\frac{\left\| \mathbf{x_i} - \mathbf{x_j} \right\|^2}{2\boldsymbol{\sigma}^2} \right) \tag{7.14}$$

(3) Linear functions: These are actually the inner product from the original space.

$$k(\mathbf{x_i}, \mathbf{x_j}) = \mathbf{x_i} \cdot \mathbf{x_j} \tag{7.15}$$

SVM is two-class classifier, but for multiple-class pattern recognition problems a combination scheme involving multiple SVM classifiers can be applied, such as the 1-a-r multiple-class classification method [121]. Stress classification is a typical two-class problem, in which neutral and stressed classes are created.

When the number of training samples is small, a well-known problem called “the curse of dimensionality” occurs. This can lead to the risk of overfitting of the training data and can result in weakening the generalization capabilities of the classifier. Conventional classification methods, such as the Gaussian Maximum Likelihood algorithm can also be applied to small sample size problems, but small training samples will result in unstable variance matrices, reducing classification performance. In the study of stress classification, it is difficult to acquire enough training samples

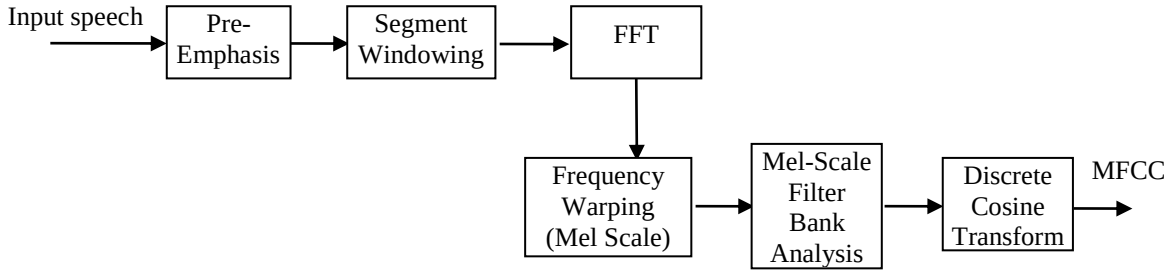


Figure 7.1 Process of MFCC extraction

from telephone communication to train the classifier, which results in the typical small sample size problem. Therefore SVM, which is a large margin based classifier with robust generalization ability, is proposed to address the small sample size problem.

7.4 METHODS AND EVALUATION

7.4.1 EXPERIMENTAL CONDITIONS

We compared the speech classification performance of several classifiers using traditional acoustic features ([SFM, F0], [TEO] and [MFCC]) with the proposed method using physical features. Fundamental frequency was obtained using the YIN method, which was proposed in [107]. YIN is based on the well-known autocorrelation method, with a number of modifications to prevent error. SFM is calculated from the spectrum of the residual signal analyzed by 12-order LPC. MFCCs have been widely applied for speech recognition because of their effectiveness in representing variations in the spectrum. Figure 7.1 shows the process for the calculation of the MFCC. Classifiers are trained using the feature vectors of16 MFCCs.

$[TEO-FM-VAR]$ 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 [61].

Speech segments with the Japanese vowels /a/ /i/, /u/, /e/ and /o/ were used as samples. The experiments were conducted for each speaker, and all of the results were speaker dependent. The samples for all the vowels were not mixed together. The vowels were first used for evaluation separately, and then the average recognition rate for the vowels was calculated to show the experimental results. 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 number of samples is about 450-700 per 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 the samples used for training and the rest used for testing. K was set to 4. The samples were analyzed with 12th-order LPC and the frame size chosen to perform the experiment was 64ms, with 16ms for frame shift.

7.4.2 EVALUATION OF CLASSIFIERS

GMMs are widely used as parametric models of the probability distribution of features. For stress classification, two GMM models were trained, one for neutral speech and the other for stressed speech. The data set for each speaker was divided evenly into four subsets, and for each classification one of the subsets was used as a test set and the other three subsets were combined to form a training set. The final result was obtained by calculating the average classification rate across

four trials using a K-fold cross-validation method. In order to increase the amount of training data, the GMMs were trained using a training set from three male speakers. The testing set of three male speakers and all of the data from the female speakers were combined to generate the testing data used in this experiment.

We first performed an experiment to find the number of mixtures which would result in the best performance of proposed feature $[k_1, k_c, A_1]$. Table 7.1 shows that the best performance is obtained when the number of mixtures equals four. When we increased the number of mixtures, the classification rate decreased, and it also made the GMM more complicated. Therefore, the number of mixture components of the GMM was set to four.

Table 7.1 Classification rates with different numbers of mixtures

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

SVMs use binary classification based on statistical learning theory, in which samples are mapped into a higher dimensional space using a kernel function to obtain a hyperplane with a maximum margin. There are many kernel functions, but only one function can satisfy the maximum margin with minimal classification error. Therefore, kernel selection is performed to evaluate the effect of different kernels on classification accuracy. We consider linear, polynomial (d = 3) and redial basis function (RBF) kernels ($\sigma = 2$) corresponding to the feature $[k_1, k_c, A_1]$. Table 7.2 shows the results in terms of stress classification accuracy.

Table 7.2 Effect of kernel selection on classification accuracy

	Kernels		
	Linear	**Polynomial**	**RBF**
Classification rate (%)	70.27	69.52	73.29

The results in the Table 7.2 show that the RBF kernel achieves the best classification performance. We also make a comparison of classification rates with different RBF widths, and the results are shown in Table 7.3. The results indicate that the performance of the classifiers is closely tied to the parameter settings. Note that performance can be improved by choosing different widths for the RBF kernel.

Table 7.3 Effect of RBF width on classification performance

Width of RBF kernel	Classification rate (%)
1	70.46
1.2	71.15
1.4	71.55
1.5	71.24
1.8	73.22
2	73.29
2.5	72.21
3.5	68.72
5	66.29

In the next evaluation, the features for [SFM,F0] [TEO-FM-VAR], [MFCC], $[k_1, k_c]$, and $[k_1, k_c, A_1]$ were modeled using GMMs with four mixture components and SVM with an RBF kernel. Classification performance is compared in Figure 7.2. We can see that better performance is achieved by the GMM compared to the linear classifier. However, the increase in the classification rate is small because of the lack of training data. If we increase the size of the training data significantly, major gains in classification rates should be achieved. We can also see that the SVM achieves even better performance than the GMM. SVMs are widely used as classifiers for two-class

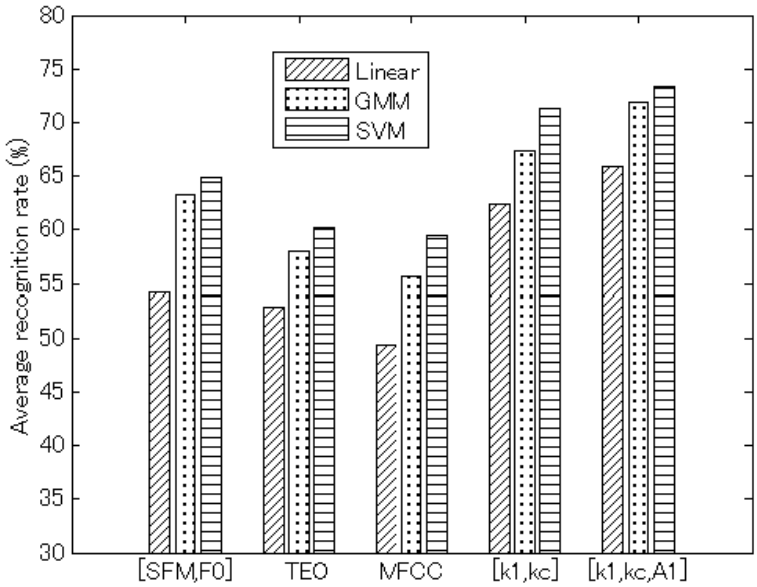


Figure 7.2 Classification performance for different classifiers

problem, which makes them suitable for separating stressed and neutral speech. Furthermore, the small size of the training sample shows the advantage of using SVMs for speech classification. According to these results, the SVM is the best classifier for the classification of speech under stress.

7.5 SUMMARY

In this chapter, speech features were modeled using Gaussian Mixture Models (GMM) and Support Vector Machines (SVM). A comparison is made of different classifiers to determine their performance in stressed speech classification. Results show that SVM outperforms the standard GMM and linear classifiers, because SVM can better solve the small sample size problem, which often occurs in stressed speech classification tasks. Therefore, the improvement of the classification rates for the proposed physical features can best be achieved by using GMMs and SVMs as statistical classifiers.

CHAPTER 8: CONCLUSION

In this dissertation, the classification of speech under stress was performed based on a physical model. We believe that the physical process of speech production plays a predominant role in the classification of stressed speech. The presence of stress can cause changes in the physical system, resulting in variations in the aerodynamics in the glottis and the vocal tract, causing stressed speech to be produced. Therefore, a physical model representing speech production can be used to characterize airflow patterns, and physical parameters of the model can be investigated to provide a better understanding of the characteristics of the physical speech production system. Methods were proposed to estimate the physical parameters of a two-mass model, and the results presented here could provide valuable insights into the classification of stressed speech. In this chapter, we draw our conclusions:

- Physical changes occur in the vocal folds when a speaker is under stress. A two-mass model can be fitted to real speech to estimate the physical parameters characterizing the behavior of the vocal folds. Results show that stress causes higher tension in the muscle of the vocal folds, lower subglottal pressure from the lungs and the vocal folds become more viscous than under neutral conditions;

- Physical changes occur in the vocal tract when a speaker is under stress. Physical parameters characterizing the shape of the vocal tract, such as vocal tract length and cross-sectional area were proposed as possible features for evaluation. Results showed that classification performance is improved by estimating these vocal tract parameters. Stress also results in

changes in the shape of the vocal tract, and the area at the entrance to the vocal tract become wider when a speaker is under stress;

- The laryngeal ventricle and false vocal folds are important areas related to the production of stressed speech. The two-mass model can be modified to model aerodynamics in the laryngeal ventricle. A physical parameter representing the effective area of laryngeal ventricle was explored and evaluation results show it is effective for stress classification;

- Improvement in the classification performance of physical parameters can be achieved by using GMMs and SVMs as statistical classifiers. Due to the small sample size of stress classification problems, an SVM is probably a better option.

We believe that physical characteristics and aerodynamics of the glottis and the vocal tract play an important role in the production of stressed speech, and are effective for stress classification. The results of this dissertation reinforced the importance of physical characteristics as an effective method for stress classification. Airflow patterns in the glottis, ventricle, and the vocal tract related to stress could also be understood.

Although we have improved the performance for stress classification, a number of problems still remain to be solved in the future. We have modeled the aerodynamics in the glottis and vocal tract using the two-mass model, however, more complicated phenomenon, such airflow separation and vortex formation should be further modeled. Database should be extended to include more speakers to evaluate the classification performance. Currently, the works we have done are mainly to classify the stress under workload, so in the future the proposed method can be applied to other stress tasks, such as emotion recognition. Although much work still remains to be done, the results are rewarding for research as a new perspective for speech analysis is promoted.

BIBLIOGRAPHY

[1] A. Mehrabian, “Communication without words,” *Psychol. Today,* vol. 2, no. 4, pp. 53–56, 1968.

[2] L.C. Nygaard, M.S. Sommers, D.B. Pisoni. “Speech perception as a talker-contingent process,” *Psychological Science,* vol. 5, no. 1, pp. 42-46, 1994.

[3] A.W.K. Gaillard and C. J. E Wientjes, “Mental load and work stress as two types of energy mobilization,” *Work and Stress*, vol. 8, no. 2, pp. 141–152, 1994.

[4] M. Frankenhaeuser, “A psychobiological framework for research on human stress and coping,” *Dynamics of stress: Physiological, psychological and social perspective* New York: Plenum Press, pp. 101-116, 1986.

[5] U. Lundberg, “Methods and applications of stress research,” *Technology and Health Care*, vol. 3, pp. 3-9, 1995.

[6] H.Ursin, and H.R. Eriksen, “The cognitive activation theory of stress. Psych oneuroendocrinology,” vol. 29, pp. 567-592, 2004.

[7] W.B. Cannon, “The emergency function of the adrenal medulla in pain and the major emotions,” *American Journal of Physiology*, vol. 33, pp. 356-372, 1914.

[8] R. McCarty and K. Pacak, “Alarm phase and general adaptation syndrome,” *G. Fink (Ed.), Encyclopedia of Stress,* New York: Academic Press, vol. 1, pp. 126-130, 2000.

[9] U. Lundberg, M. Frankenhaeuser, “Pituitary-adrenal and sympathetic-adrenal correlates of distress and effort,” *Journal of Psychosomatic Research*, vol. 24, pp. 125-130, 1980.

[10] T. G. Pickering, “Ambulatory blood pressure monitoring,” *New England Journal of Medicine*, vol. 354, no. 22, pp. 2368-2374, 2006.

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

[12] S. Bou-Ghazale， J H L. Hansen, “A comparative study of traditional and newly proposed feature for recognition of speech under stress,” *IEEE Trans on Speech and Audio Processing*, vol. 8, no. 4, pp. 429 – 442, 2000.

[13] E. Lombard， ”Le Signe de l’Elevation de la Voix,” *Ann. Maladies Oreille， Larynx.Nez， Pharynx*, vol. 37, pp. 101 – 119, 1911.

[14] H.J.M Steeneken, J.H.L. Hansen, “Speech under stress conditions : overview of the effect on speech production and on system performance,” *Proc ICASSP. USA：IEEE Press，* pp. 2079 – 2082, 1999.

[15] J.H.L. Hansen, “Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition,” *Speech communication*, vol. 20, no. 1, pp. 151-173, 1996.

[16] R. Fernandez, R.W. Picard, “Modeling drivers’ speech under stress,” *Speech Communication*, vol. 40, no.1, pp. 145-159, 2003.

[17] V Varadarajan, J.H.L Hansen, I. Ayako. “UT-SCOPE–a corpus for speech under cognitive/physical task stress and emotion,” *The Workshop Programme Corpora for Research on Emotion and Affect* Tuesday, 23rd May pp. 72, 2006.

[18] C.A. Simpson, “Speech Variability Effects on Recognition Accuracy Associated With Concurrent Task Performance by Pilots,” *Technical report, Psycho-Linguistic Research Associates,* 1985.

[19] J. H. L. Hansen, C. Swail, A. J. South, R. K. Moore, H. Steeneken, E. J. Cupples, T. Anderson, C. R. A. Vloeberghs, I. Trancoso, and P.Verlinde, “The Impact of Speech Under Stress on Military Speech Technology,” *NATO Research & Technology Organization RTO-TR-10,* vol. AC/323(IST)TP/5 IST/TG-01, Mar. 2000

[20] K. Prahallad, A. Black, R. Mosur, “Sub-Phonetic Modeling for Capturing Pronunciation, Variation in Conversational Speech Synthesis,” *Proceedings of the 31th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’06)*, Toulouse, 2006.

[21] S. Casale, A. Russo, S. Serrano, “Multi-style classification of speech under stress using feature subset selection based on genetic algorithms,” *Speech Communication*, , vol. 49, no. 10, pp. 801-810, 2007.

[22] J. H. L. Hansen, S. Bou-Ghazale, R. Sarikaya, “Getting started with SUSAS: A speech under simulated and actual stress database,” *Proc. Eurospeech.* vol. 97, no. 4, pp. 1743-1746, 1997.

[23] P.K. Rajasekaran, G.R. Doddington, J.W. Picone, “Recognition of Speech under Stress and in Noise,” *Proceedings of the 11th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’86),* Tokyo, pp. 733–736, 1986.

[24] N. Matsuo, N. Washio, S. Harada, A. Kamano, S. Hayakawa, K. Takeda. “A study of psychological stress detection based on the non-verbal information,” *IEICE Technical Report*, IEICE-SP2011-35, pp. 29-33, 2011. (In Japanese).

[25] T. Murry, E. J. Nelson, E. W. Swenson. “Speech Intelligibility During Exercise at Normal and Increased Atmospheric Pressures,” *ADA749321*, 1972.

[26] H. Montague, “Voice Control Systems for Airborne Environments,” *SCOPE ELECTRONICS INC RESTON VA*, no. 6205-0377, 1977.

[27] J. W. Glenn, R. N. Gordon, G. Moschetti, “Voice Initiated Cockpit Control and Interrogation（VICCI）System Test for Environmental Factors,” *SCOPE ELECTRONICS INC RESTON VA*, 1971.

[28] A. C. Busch, D. Eldredge, “Duration and Intensity of Vocalic Elements as Physical Correlates of Acoustic Stress,” *Perceptual and Motor Skills*, vol. 23, no. 3, pp. 801-802, 1966.

[29] J. Nicholson, K. Takahashi, R. Nakatsu, “Emotion recognition in speech using neural networks,” *Neural Computing & Applications*, vol. 9, no. 4, pp. 290-296, 2000.

[30] T. Moriyama, S. Ozawa “Emotion recognition and synthesis system on speech,” *IEEE International Conference on Multimedia Computing and Systems. Italy：IEEE Press*, vol. 1, pp. 840 - 844, 1999.

[31] L. S. Chen, T. S. Huang, T. Miyasato, R. Nakatsu, “Multimodal human emotion/expression recognition,” *Proceedings. Third IEEE International Conference on Automatic Face and Gesture Recognition. Japan：IEEE Press*, pp. 366-371, 1998.

[32] Y. Chen, “Cepstral domain talker stress compensation for robust speech recognition,” *IEEE Trans on Acoustics, Speech and Signal Processing*, vol. 36, no. 4, pp. 433 – 439, 1988.

[33] E. G. Bard, C. Sotillo, A. H. Anderson, M. M. Taylor. “The DCIEM map task corpus: spontaneous dialogue under sleep deprivation and drug treatment,” *ICSLP 96. USA, IEEE Press*, pp. 1958 – 1961, 1996.

[34] R. Van Bezooijen, “The Characteristics and Recognizability of Vocal Expression of Emotions,” *Foris, Netherlands*, 1984.

[35] F. J. Tolkmitt, K.R. Scherer, “Effect of experimentally induced stress on vocal parameters,” *J. Exp. Psychol. [Hum. Percept.]*, vol. 12, no. 3, pp. 302-313, 1986.

[36] K. E. Cumming, M.A. Clements. “Application of the analysis of glottal excitation of stressed speech to speaking style modification,” *ICASSP’93. USA : IEEE Press*, pp. 207 – 210, 1993.

[37] C. E. Williams, and K. N. Stevens, "On Determining the Emotional State of Pilots During Flight: An Exploratory Study," *Aerospace Medicine 40*, pp. 1369-1372, 1969.

[38] C. E. Williams, K.N. Stevens. “Emotions and speech: some acoustic correlates,” *the Journal of the Acoustical Society of America*, vol. 52, no. 4, pp. 1238 – 1250, 1972.

[39] L.A. Streeter, N.H. Macdonald, W. Apple, R. M. Krauss, and K. M. Galotti, “Acoustic and Perceptual Indicators of Emotional Stress,” *J. Acoust. Soc. Am*, vol. 73, no. 4, pp. 1354-1360, 1983.

[40] M. H. L Hecker, K. N Stevens, G. von Bismark, and C. E. Williams, “Manifestations of Task-Induced Stress in the Acoustic Speech Signal,” *J. Acoust. Soc. Am*, vol. 44, pp. 993-1001, 1968.

[41] G. R. Griffin, C, E.Williams, “Effects of Different Levels of Task Complexity on Three Vocal Measures,” *Aviation, space, and environmental medicine*, 1987.

[42] D. B. Pisoni, R. H. Bernacki, H. C. Nusbaum, M. Yuchtman, “Some acousticphonetic correlates of speech produced in noise,” *ICASSP’85, USA : IEEE Press*, pp. 1581 - 1585, 1985.

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

[44] J. H. L. Hansen, S. Bou-Ghazale, “Robust speech recognition training via duration and spectral-based stress token generation,” *IEEE Trans on Speech and Audio Processing*, vol. 3, no. 5, pp. 415 – 421, 1995.

[45] J. H. L. Hansen, B. D. Womack, “Feature analysis and neural network based classification of speech under stress,” *IEEE Trans on Speech and Audio Processing*, vol. 4, no. 4, pp. 307 – 313, 1996.

[46] Z. S. Bond and T. J. Moore, “A note on loud and lombard speech,” *in Int. Conf. Speech Language Processing ’90*, pp. 969–972, 1990.

[47] J. H. L. Hansen, “Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition,” *Speech Communication*, vol. 20, no. 1, pp. 151-173, 1996.

[48] J. C. Junqua, “The influence of acoustics on speech production: A noise-induced stress phenomenon known as lombard reflex,” *Speech Communication*, vol. 20, no. 1–2, pp. 13–22, 1996.

[49] B. J. Stanton, L. H. Jamieson, and G. D. Allen, "Robust recognition of loud and Lombard speech in the fighter cockpit environment," *in Proc. Int. Conf. Acoustic, Speech, Signal Processing '89*, pp. 675–678, 1989.

[50]C. Baber and J.M. Noyes, "Automatic speech recognition in adverse environments," *Human Factors*, vol. 38, no. 1, pp. 142-155, 1996.

[51] M.E. McCauley, R.W. Root and F.A. Muckier, "Training evaluation of an automated air intercept controller training system," *Naval Training Equipment Center Report NAVTRAEQUIPCEN 8 1 -C-0055- 1, Orlando, FL*, 1982.

[52] M.L. Hecker, K.N. Stevens, G. von Bismarck and C.E. Williams "Manifestations of task induced stress in the acoustic speech signal," *J. Acoust. Soc. Am*, vol. 44, pp. 993-1001, 1968.

[53] C. Baber, B. Mellor, R. Graham, J. M. Noyes, and C. Tunley, "Workload and the use of automatic speech recognition: The effects of time and resource demands," *Speech Communication*, vol. 20, no. 12, pp. 37-54, 1996.

[54] E. G. Bard, C. Sotillo, A. H. Anderson, H. S. Thompson, and M. M. Taylor, "The DCIEM map task corpus: Spontaneous dialogue under sleep deprivation and drug treatment," *Speech Communication*, vol. 20, pp. 71–84, 1996.

[55] I. R. Murray, C. Baber, and A. South, "Toward a definition and working model of stress and its effects on speech," *Speech Communication*, vol. 20, pp. 3-12, 1996.

[56] D. B. Paul, "A speaker-stress resistant HMM isolated word recognizer," *in Proc. Int. Conf. Acoustic, Speech, Signal Processing '87*, pp. 713–716, 1987.

[57] R. Ruiz, E. Absil, B. Harmegnies, C. Legros, and D. Poch, “Time- and spectrum-related variabilities in stressed speech under laboratory and real conditions,” *Speech Communication*, vol. 20, pp. 111–130, 1996.

[58] J. Whitmore and S. Fisher, “Speech during sustained operations,” *Speech Communication*, vol. 20, pp. 55-70, 1996.

[59] A. Kamano, N. Washio, S. Harada, N. Matsuo, “A study of psychological suppression detection based on non-verbal information,” *IEICE Technical Report IEICE-SP2010-64*, pp. 107-110, 2010 (in Japanese).

[60] *Proc. Int. Conf. Acoustics, Speech, Signal Processing ’99: Special Session on Speech Under Stress*, vol. 4, Mar. 1999, pp. 2079–2098.

[61] S. E. Bou-Ghazale and J. H. L. Hansen, “Generating stressed speech from neutral speech using a modified CELP vocoder,” *Speech Communication*, vol. 20, pp. 93–110, Nov. 1996.

[62] G. Fant, “Acoustic Theory of Speech Production,” *Walter de Gruyter*, 1970.

[63] H. K. Dunn, “Methods of measuring vowel formant bandwidths,” *J Acoust Soc Am* vol. 33, no. 12, pp. 1737–1746, 1961.

[64] D. Y. Wong, J. D. Markel, A. H. Gray, “Glottal inverse filtering from the acoustic speech waveform,” *IEEE Trans Acoust Speech Signal Process*, vol. 27, no. 4, 350-355, 1979.

[65] J. F. Kaiser, “Some observations on vocal tract operation from a fluid flow point of view,” *Vocal fold physiology: biomechanics, acoustics, and phonatory control*, pp. 358-386, 1983.

[66] H. M.Teager, “Some observations on oral air flow during phonation”, *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 28, no. 5, pp. 599-601, 1980.

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

[68] J. F. Kaiser, “On Teager's Energy Algorithm and Its Generalization to Continuous Signals,” *in Proc. 4th IEEE Digital Signal Processing Workshop*. *New Paltz, NY*, 1990.

[69] P. Maragos， J. F. Kaiser， T. F. Ouatieri, “Energy separation in signal modulation with application to speech analysis,” *IEEE Transaction Signal Processing*, vol. 41, pp. 3024 – 3051, 1993.

[70] G. Zhou, J. H. L. Hansen， J. F. Kaiser, “Classification of speech under stress based on features derived from the nonlinear teager energy operator,” *ICASSP’98, USA, IEEE Press*, pp. 549 – 552, 1998.

[71] G Zhou， J. H. L. Hansen， J. F. Kaiser, “Nonlinear feature based classification of speech under stress,” *IEEE Trans on Speech and Audio Processing*, vol. 9, no. 3, pp. 201-206, 2001.

[72] G Zhou， J. H. L. Hansen， J. F. Kaiser, “Linear and nonlinear feature speech analysis for stress classification,” *ICSLP’98, .Australia: ASSTA Publication*, pp. 840 – 843, 1998.

[73] J. H. L. Hansen， B. D. Womack. “Feature analysis and neural network based classification of speech under stress,” *IEEE Trans on Speech and Audio Processing*， vol. 4, no. 4, pp. 307 – 313, 1996.

[74] L. N. Tin, S. W. Foo, L. C. De Silva, "Speech based emotion classification," *Electrical and Electronic Technology. Proceedings of IEEE Region 10 International Conference on TENCON, Singapore: IEEE Press*, pp. 297 – 301, 2001.

[75] E. Erzin, A. E. Cetin, Y. Yardimci, "Sub-band analysis for robust speech recognition in the presence of car noise," *ICASSP'95, USA: IEEE Press*, pp. 417 -420, 1995.

[76] R. Sarikaya, J. N. Gowday, "Wavelet based analysis of speech under stress," *Southeast Con'97 Engineering New Century. USA: IEEE Press*, 92 – 96, 1997.

[77] R. Sarikaya, J. N. Gowday. "Sub-band based classification of speech under stress," *ICASSP'98*, vol.1, pp. 569 – 572, 1998.

[78] R. Sarikaya, J. H. L. Hansen, "High resolution speech feature parametrization for monophone-based stressed speech recognition," *IEEE Signal Processing Letters, v*ol. 7, no. 7, pp. 82 – 185, 2000.

[79] A. C. Busch, D. Eldredge, "Duration and Intensity of Vocalic Elements as Physical Correlates of Acoustic Stress," *Perceptual and Motor Skills*, vol. 23, no. 3, pp. 801-802, 1966.

[80] B. D. Womack, J. H. L. Hansen, "N-D HMMs for combined stress speech classification and recognition," *IEEE Trans. Speech Audio Proc.*, Sept. 1996.

[81] S. E. Bou-Ghazale, J. H. L. Hansen, "Stress perturbation of neutral speech for synthesis based on hidden Markov models," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 201–216, 1998.

[82] R. Lippmann, E. A. Martin, and D. B. Paul, “Multi-style training for robust isolated-word speech recognition,” *in Proc. Int. Conf. Acoustic, Speech, Signal Processing ’87*, pp. 705–708, 1987.

[83] B. D. Womack and J. H. L. Hansen, “Classification of speech under stress using target driven features,” *Speech Communication*, vol. 20, pp. 131–150, 1996.

[84] S. A. Patil, J. H. L. Hansen, “Detection of speech under physical stress: Model development, sensor selection, and feature fusion”, 2008.

[85] R. Daniloff, G. Schuckers, and L. Feth, “The physiology of speech and hearing: An introduction,” *Englewood Cliffs, NJ: Prentice-Hall*, 1980.

[86] D. B. Fry, “The physics of speech,” Cambridge: Cambridge University Press, 1979.

[87] Scherer, R. Klaus, “Methods of research on vocal communication: Paradigms and parameters,” Handbook of methods in nonverbal behavior research, pp. 136-198. 1982.

[88] Zemlin, R. Willard, “Speech and Hearing Science, Anatomy and Physiology.” 1968.

[89] S. M. Teager, “Evidence for nonlinear production mechanisms in the vocal tract,” *in Speech Production and Speech Modeling*. vol. 55, pp. 241–261, 1989.

[90] C. Z. Wei, et a1, “Computational aeroacoustics of phonation Part: Effects of flow parameters and ventricular folds,” *J Acoust Soc Am*, vol. 112, no. 5, pp. 2147, 2002.

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

[92] M. Krane, M. Barry, and T. Wei, “Unsteady behavior of flow in a scaled-up vocal folds model,” *J. Acoust. Soc. Am*, vol. 122, pp. 3659–3670, 2007.

[93] K. Ishizaka, J. L. Flanagan, “Synthesis of voiced sounds from a two-mass model of the vocal cords,” *Bell.Syst.Tech Journal*, vol. 51, pp. 1233-1268, 1972.

[94] J. L. Flanagan, “Speech analysis, synthesis, and perception,” Springer-Verlag, New York, 1972.

[95] “Vocal folds,”[Online] Wikipedia, Available: http://en.wikipedia.org/wiki/Vocal_folds.

[96] T. Haji, K. Mori, K. Omori, “Experimental studies on the viscoelasticity of the vocal fold,” *Acta oto-laryngologica*, vol. 112, no.1, pp. 151-159, 1992.

[97] R. J. B. Hemler, G. H. Wieneke, J. Lebacq, “Laryngeal mucosa elasticity and viscosity in high and low relative air humidity,” *European archives of oto-rhino-laryngology*, vol. 258, no. 3, pp. 125-129, 2001.

[98] C. Lucero, “Chest- and falsetto-like oscillations in a two-mass model of vocal folds,” *J. Acoust. Soc. Am*, pp. 3355-3399, 1996.

[99] B. K. Finkelhor, I. R. Titze, P.L. Durham, “The effect of viscosity change in the vocal folds on the range of oscillation,” *J. Voice*, vol. 1, pp. 320-325, 1988.

[100] T. Kaneko, H. Asano, J. Naito, N. Kobayashi, K. Hayashi and T. Kitamura, “Biomechanics of the vocal cords-on damping ratio,” *J. Jpn. Soc Bronchoesophagol*, vol. 25, pp. 133-138, 1972.

[101] N. Isshiki, “Functional surgery of the laryn,” Kyoto University, Kyoto, Japan, pp. 62-67, 1977.

[102] P. H. Dejonckere, and J. Lebacq. “Damping coefficient of oscillating vocal folds in relation with pitch perturbations,” *Speech Communication*, vol. 3, pp. 89-92, 1984.

[103] D. Kincaid, W. Cheney, “Numerical analysis: mathematics of scientific computing,” *3rd ed. (Brook/Cole, Pacific Grove, CA)*, pp. 722-723, 2002.

[104] “Vocal tract” [Online] Wikipedia, Available: http://en.wikipedia.org/wiki/Vocal_tract.

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

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

[107] I. R. Titze, “Acoustic Interpretation of Resonant Voice. *J Voice*, vol. 15, pp. 519-528, 2001.

[108] I. R. Titze, B. H. Story, “Acoustic interactions of the voice source with the lower vocal tract,” *The Journal of the Acoustical Society of America*, vol. 101, pp. 2234, 1997.

[109] A. De Cheveigne, H Kawahara, “YIN, a fundamental frequency estimator for speech and music,” *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917-1930, 2002.

[110] A. Kamano, N. Washio, S. Harada, N. Matsuo, “A study of psychological suppression detection based on non-verbal information,” *IEICE Technical Report, IEICE-SP2010-64*, pp. 107-110, 2010 (in Japanese).

[111] B. Schuller, G. Rigoll, M. Lang, “Hidden Markov model-based speech emotion recognition,” *Proceedings of the IEEE ICASSP Conference, Hong Kong*, pp. 1-4, 2003.

[112] J. S. Bridle, L. Dodd, "An Alphanet approach to optimizing input transformations for continuous speech recognition," *Proc. ICASSP91*, pp. 277-280, 1991.

[113] Robinson, J. Anthony, "Dynamic error propagation networks," Diss. University of Cambridge, 1989.

[114] G. D. Cook, A. J. Robinson, "The 1997 ABBOT system for the transcription of broadcast news," *Proceedings of the 1998 Broadcast News Transcription and Understanding Workshop*, 1998.

[115] N. Ström, "Acoustic modeling improvements in a segment-based speech recognizer," *Proceedings of IEEE ASRU Workshop*, 1999.

[116] Y. Yoshitomi, S.I. Kim, T. Kawano, and T. Kitazoe, "Effect of Sensor Fusion for Recognition of Emotional States Using Voice, Face Image and Thermal Image of Face," *Proceedings of the ninth IEEE International Workshop on Robot and Human Interactive Communication*, pp.173-183, 2000.

[117] V. N. Vapnik, "Statistical learning theory. 1998," 1998.

[118] F. Yu, E. Chang, Y.Q. Xu, and H.Y. Shum, "Emotion Detection from Speech to Enrich Multimedia Content," *Proceedings of IEEE Pacific Rim Conference on Multimedia*, pp.550-557, 2001.

[119] H. Hu, M. Xu, and W. Wu, "GMM supervector based SVM with spectral features for speech emotion recognition," *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, pp. 413-416, 2007.

[120] O. Edgar, F. Robert, G. Federico. “Training support vector machines: an application to face detection,” *IEEE Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico*. Pp. 130-136, 1997.

[121] C. W. Hsu, C. J. Lin. “A comparison of methods for multiclass support vector machines,” *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415-425, 2002.

PUBLICATIONS

Journal papers

[1] Xiao Yao, Takatoshi Jitsuhiro, Chiyomi Miyajima, Norihide Kitaoka, and Kazuya Takeda, "Classification of speech under stress based on modeling of the vocal folds and vocal tract", EURASIP Journal on Audio, Speech, and Music Processing, July, 2013, doi:10.1186/1687-4722-2013-17.

[2] Xiao Yao, Takatoshi Jitsuhiro, Chiyomi Miyajima, Norihide Kitaoka, and Kazuya Takeda, "Classification of speech under stress by physical modeling", Acoustical Science and Technology. 2013 (to appear).

International conference proceedings

[1] Xiao Yao, Takatoshi Jitsuhiro, Chiyomi Miyajima, Norihide Kitaoka, and Kazuya Takeda, "Classification of speech under stress by modeling the aerodynamics of the laryngeal ventricle", 14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013), Lyon, France, (8/25-29), August 2013. (to appear).

[2] Xiao Yao, Takatoshi Jitsuhiro, Chiyomi Miyajima, Norihide Kitaoka, and Kazuya Takeda, "Estimation of vocal tract parameters for the classification of speech under stress", Proceedings of 37th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013), Vancouver, pp.7532-7536, 2013.

[3] Xiao Yao, Takatoshi Jitsuhiro, Chiyomi Miyajima, Norihide Kitaoka, and Kazuya Takeda, "Classification of Stressed Speech Using Physical Parameters Derived from Two-Mass Model", 13th Annual Conference of the International Speech Communication Association (INTERSPEECH 2012), Portland, Oregon, USA, (9/9-13), pp. Sept. 2012.

[4] Xiao Yao, Takatoshi Jitsuhiro, Chiyomi Miyajima, Norihide Kitaoka, and Kazuya Takeda, "Physical characteristics of vocal folds during speech under stress"，Proceedings of 37th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012), Kyoto, pp.4609-4612, 2012.

[5] Xiao Yao, Takatoshi Jitsuhiro, Chiyomi Miyajima, Norihide Kitaoka, and Kazuya Takeda, “An analysis of the speech under stress using the two-mass vocal fold model”, The 3rd International Workshop on Spoken Dialogue Systems Technology (IWSDS 2011), Granada, Spain, (9/1-3), pp.53-58, Sept. 2011.

Domestic conference proceedings

[1] Yao Xiao, Takatoshi Jitsuhiro, Chiyomi Miyajima, Norihide Kitaoka, Kazuya Takeda, " Classification of speech under stress using physical features based on two-mass model," 電子情報通信学会 技術報告 (IEICE Technical Report), vol.112, no.450, SP2012-115, pp.47-52, March 2013.

[2] Xiao Yao, Takatoshi Jitsuhiro, Chiyomi Miyajima, Norihide Kitaoka, and Kazuya Takeda, “Evaluation for vowel-independent classification of speech under stress based on interaction between the vocal folds and the vocal tract”, 2012 Autumn Meeting, Acoustic Society of Japan (ASJ), Shinshu University, Nagano, 1-2-19, pp.269-272, Sept. 2012.

[3] Xiao Yao, Takatoshi Jitsuhiro, Chiyomi Miyajima, Norihide Kitaoka, and Kazuya Takeda, “Detection for stressed speech based on two-mass model”, 2012 Spring Meeting, ASJ， Kanagawa University, Yokohaba. 1-7-2, March. 2012.

[4] Yao Xiao, Takatoshi Jitsuhiro, Chiyomi Miyajima, Norihide Kitaoka, Kazuya Takeda, "On the use of the two-mass vocal cord model in characterizing the stress speech," 電子情報通信学会 技術報告 (IEICE Technical Report), vol.111, no.97, SP2011-36, pp.35-40, June 2011