Mental Tension Detection in the Speech based on physiological monitoring

Michiaki Ariga, Yoshikazu Yano, Shinji Doki, and Shigeru Okuma

Abstract— **The focus of this paper is mental tension detection in speech to assist control the tension in day-to-day business such as conferences and operations in a call center. It is difficult to use classical techniques for mental tension detection in day-to-day business because those techniques require invasion body by electrodes or squirts and tied up by cables. In order to achieve a non-invasive, non-contact and low-restricting method, this proposed technique uses acoustic features in the speech. The technique uses the vocal tract model which represents the shape and the tightness of throat muscle. The Gaussian Mixture Model (GMM) classifies two mental tension states: high-tension and non-tension. The experiment result shows high recognition rate of mental tension detection.**

I. INTRODUCTION

In order to achieve cooperative human machine interfaces, it is required for machines to recognize the purposes and the conditions of operators. The detection of high tension state is widely required which causes adverse effects in some operations. Under high tension state, it is occurred several kinds of troubles at day-to-day businesses such as decrease of efficiency in conferences, error in judgments of operators in a call center, steering control failure on driving, and so on. In order to control the mental tension, it is required to detect high tension state.

There are classical techniques to detect mental tension such as a Skin Conductance Response (SCR) which is caused by differences of potential derived from mental sweating [1], heart rates which are measured by electrodes [2] and hormones in blood or saliva [3] [4]. These techniques compel electrodes invasion on a body and motion restriction by cables. These invasivenesses and the restrictions made it difficult to apply classical techniques in day-to-day businesses. Biochemical tests such as analysis of hormones need to take some period, therefore these lack quick responses. Additionally these sensors give a psychological burden on test subjects.

It is known that the mental tension changes utterances from those of non-tension [5] [6]. The non-tension state is defined as the state we are not in high-tension state. In this paper, the proposed method is mental tension detection in speech which is non-invasive, non-contact and low-restricting. Utterances shows rapid responses according to changing tension states. Additionally, heart rates is observed to confirm high-tension state.

Michiaki Ariga, Shinji Doki, and Shigeru Okuma are with Department of Electrical Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya City, Japan {ariga,doki,okuma}@okuma.nuee.nagoya-u.ac.jp

Yoshikazu Yano is with the school of Electrical Engineering, Aichi Institute of Technology, Yakusa-cho, Toyota City, Japan yoshiyano@aitech.ac.jp

II. MENTAL TENSION

In this section, the definition of the mental tension is described. Generally, mental tension becomes higher when we do some inexperienced tasks or we attention-getting jobs. Particularly, we become high tension by evaluator.

A. Responses against Mental High Tension

When feeling societal menace, people get high tension. The societal menace is occurred when we are against for making more mistakes, when we are evaluated by supervisor, or when we feel anxiety about bleak future.

Following three responses are emerged by mental tensions; subjective experiences, physiological responses and behavioral responses. Subjective experiences are the awarenesses that we find ourselves becoming tension high. Physiological responses appear on the physiological information emerged by autonomic nervous system. Behavioral responses are expressed in gestures such as quavers and milling around, which are recognized as habits.

As obtaining objective data easily, mental tension is defined as appearances of physiological responses are emerged.

B. Physiological Response

When we are stimulated at a certain situation and become tension high, physiological responses appear such as rising of heart rates, dilating peripheral blood vessels, rising in blood pressure, mental sweating and so on. Since physiological responses are controlled by autonomic nervous system, we cannot control the responses consciously. Though there are differences of degree among individuals in physiological responses, these responses appear in people under high tension. By monitoring physiological responses, high tension is detected subjectively.

III. PROPOSED TECHNIQUE

There are two typical features under high tension in speech. Former is tightness of throat muscles as physiological responses and latter is rapid utterance as behavioral responses. The proposed technique detects physiological responses, because appearances of physiological responses do not depend on individual characteristics, while those of behavioral responses depend on individual habits. The proposed system models each distribution of the utterances in high-tension and non-tension using voice quality features which are associate with the shape of vocal tract and tightness

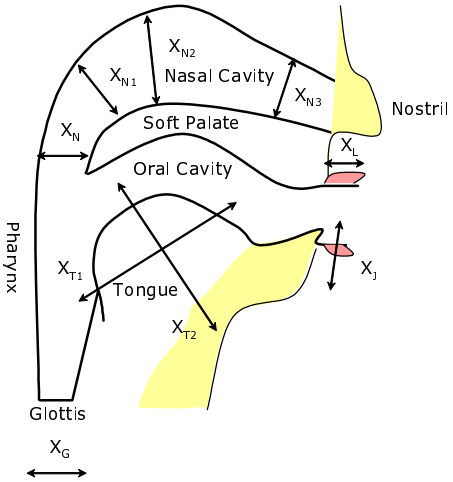

Fig. 1. Model of Vocal Tract

of throat muscle using vocal tract model. Additionally, the Gaussian Mixture Model is used for learning and recognizing states of tension.

A. Voice Quality Feature

1) Vocal Tract Model: The vocal tract model is a model which corresponds the shape of vocal tract and tightness of throat muscle. The cross sectional view of throat is shown in Fig.1. The vocal tract is consisted of the nasal cavity, the oral cavity, the pharynx, the glottis, and so on. The human speech signal is given by the equation

$$X(\omega) = G(\omega)H(\omega) \tag{1}$$

where ω is frequency, $X(\omega)$ is speech signal, $G(\omega)$ is a vibration of vocal cords and $H(\omega)$ an impulse response.

The vocal tract model represents the shape of the throat tube and the resonance of the vocal tract. The voice quality features can extract the resonance of the vocal tract.

2) Cepstrum: The cepstrum $c(\tau)$ is one of the voice quality features and given by following equation

$$c(\tau) = \mathscr{F}^{-1} log|X(\omega)| = \mathscr{F}^{-1} log|G(\omega)| + \mathscr{F}^{-1} log|H(\omega)| \tag{2}$$

where τ is the quefrency which correspond to the frequency, $\mathscr{F}^{-1}$ is the symbol of inverse Fourier transform and $X(\omega)$ is short time spectrum amplitude. It is able to extract vocal tract property with analysing low-order cepstrum.

It is used Mel-Frequency Cepstrum Coefficient (MFCC) as a voice quality feature. The MFCC is commonly used for speaker and environment recognition applications. The MFCC is obtained from cepstrum using Mel Filter Bank, which is an aural scale of pitches.

The cepstrum represents the shape of vocal tract and the tightness of throat muscles. When the system uses the cepstrum with small number of utterances, it is required to use same content among all states of speech. The shapes of vocal tract are affected by pronunciations such as vowels and consonants. Thus, the cepstrum is affected by pronunciations. In order to avoid this affect, it is required to obtain a lot of number of utterances with various phoneme.

The cepstrum can be calculated from short period speech data (25ms) which is much shorter that for extracting than prosodic features such as the pitch. On the point of real time recognition, the results should be calculated with little latency. This system estimates high-tension state from 25ms utterance with small amount of calculation. The cepstrum can extract a lot of sample data easily because it can extract in short period. Thus, when the system learns, the burdens of speakers are smaller than the situation when prosodic features are used.

B. Model of High-tension and Non-tension Utterances

It is used the Gaussian Mixture Model (GMM) to model high-tension and non-tension utterances. The GMM is suited for text-independent applications such as speaker and environment recognition [7] [8]. The Gaussian mixtures represent the tension-dependent distribution of input data, which is the MFCC extracted by utterances in this paper. The GMM uses the likelihood which represents how suitable for extracted features of utterances are for the two tension models. The system selects the model as the output state which outputs larger value of likelihood (Fig.2). In Fig.2, the system recognizes output state as high-tension state.

The algorithm of the GMM is given below. A gaussian mixture model of high-tension state λ_H is a sum of M gaussians and given by the equation

$$p(x \mid \lambda_H) = \sum_{m=1}^{M} w_m b_m(x) \tag{3}$$

where x is a D-dimensional random vector, D is the number of Mel-Cepstral coefficients, $b_m(x)$, $m = 1,...,M$, are the gaussians and w_m, are the mixture weights. The non-tension state λ_N takes a similar way to calculate. Each component gaussian is a D-variate Gaussian function and given by the equation

$$b_m(x) = \frac{1}{\sqrt{(2\pi)^D}(|\sigma_m|)^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(x-\mu_m)^t \sigma_m^{-1}(x-\mu_m)\right\} \tag{4}$$

with mean vector μ_m and covariance matrix σ_m. The mixture weights satisfy the constrain that

$$\sum_{m=1}^{M} w_m = 1 \tag{5}$$

The Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all gaussians. For tension detection, each state is represented by a GMM and is referred to by model λ_H or λ_N.

An extracted feature vector line X from T frames of input speech is given by the following equation

$$X = \{x_1, x_2, \ldots, x_T\} \tag{6}$$

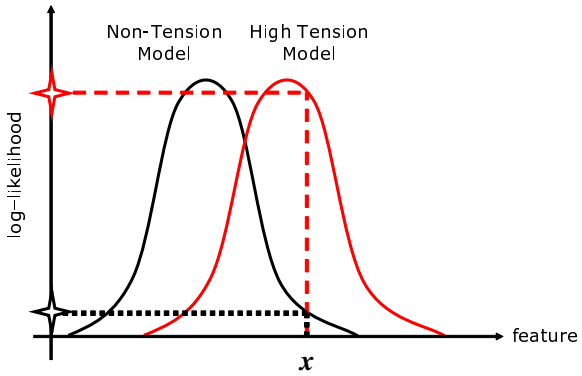

Fig. 2. Recognition using the fitness of the input vector

TABLE I
SPEECH DATABASE 1
BOLDFACED CONDITIONS ARE DEFFERENCE ONES FROM DATABASE 2

Speaker	:	**6 male students**
State	:	Non-Tension , High-Tension
Speech text	:	"The purpose of this study is,", etc... phrase on **briefing session** in Japanese
Speech length	:	Ave. 3sec
Environment	:	**Lecture room** with reverberation sound and fan noise
Sampling frequency	:	16kHz 16bit / monaural

On the extracted feature vector line X, the likelihood $P(X|\lambda)$ and log-likelihood $logP(X|\lambda)$ are given by the equations

$$P(X \mid \lambda) = \prod_{t=1}^{T} p(x_t \mid \lambda) \tag{7}$$

$$\log P(X \mid \lambda) = \sum_{t=1}^{T} p(x_t \mid \lambda) \tag{8}$$

A GMM is learned by maximizing likelihood in (7) and estimate parameters. EM algorithm is used for estimation the parameters.

According to (8), using the likelihood, the GMM has a capability to recognize tension state by input data. The system recognizes the suitableness of the feature distribution in unknown utterance for each tension model. On purpose log-likelihood is motivated to calculate easily. By recognizing tension state from time-series input utterances, the system selects the model as the output state which outputs larger value of log-likelihood in(8).

IV. EXPERIMENT OF TENSION DETECTION IN SPEECH

A. Experimental Condition

The purpose of this experiment is to clarify the system can classify the speeches into two mental tension states. It was investigated the recognition rate and the likelihood of utterances on high-tension and on non-tension.

It is prepared Speech database1 shown in Table I. Different conditions from the follow experiment are listed in boldface. Speech data are recorded from 6 male students. Experimentation procedures are mentioned to obtain high-tension and non-tension speeches. In order to get obvious high-tension utterances, we obtained the speeches in briefing session that speakers were assessed external observer. Those utterances are easy to distinguish between high-tension or non-tension utterances by listening. It is said in psychology that the speech is the situation we feel societal anxiety. The speech task is utilised at psychological experiment scene. Thus, all the utterances recorded on a briefing session were labeled as high-tension utterances. The length of phrases is from 1 second to 15 seconds (Ave. 3 seconds).

We obtained the non-tension speeches when the speaker calmed down adequately. In order to clear off the phonological affect which is mentioned in III-A.2 and to evaluate the affect of only mental tension, we used the same contents of non-tension utterances as that of high-tension ones. The speech database was obtained in a lecture room, there were reverberation sounds and background noises that were occurred by a projector fan. In order to arrange the same condition on recording high-tension utterances, we obtained non-tension utterances in the same room and same noise condition.

It is used 18 dimension MFCC for the voice quality feature. The condition of feature extraction is shown in Table II. The frame length for feature extraction is 25ms so that we obtain a lot of utterances from the speech database. The number of utterances is shown in Table III. The 80% of utterances are used for training recognition system, and the rests 20% are applied for verifying proposed system.

GMM is used for recognition system. The number of Gaussian mixtures is 32. The system was trained and recognized speaker independently.

TABLE II
FEATURE PROPERTY

Feature	:	MFCC in 18 dim.
Frame length	:	25 msec
Frame interval	:	10 msec

TABLE III
FRAME NUMBERS OF EACH SPEAKER

	Non-Tension	High-Tension
Speaker 1	41062	14555
Speaker 2	70316	21203
Speaker 3	57489	15663
Speaker 4	79305	23499
Speaker 5	40715	13655
Speaker 6	74026	24473

B. Result 1 : Long Time Analysis

We experimented in order to clarify the capability to classify the utterances into two degrees of mental tension. The way to calculate the recognition rate of high-tension

TABLE IV

RECOGNITION RATE OF 6 SPEAKERS A FRAME

	Non-Tension	High-Tension
Speaker 1	85.5%	78.9%
Speaker 2	83.0%	76.7%
Speaker 3	73.8%	71.3%
Speaker 4	78.0%	76.5%
Speaker 5	88.7%	84.3%
Speaker 6	81.9%	78.6%

TABLE V

RECOGNITION RATE OF 6 SPEAKERS A SECOND

	Non-Tension	High-Tension
Speaker 1	98.8%	98.1%
Speaker 2	98.5%	92.3%
Speaker 3	99.3%	94.3%
Speaker 4	99.9%	99.1%
Speaker 5	99.3%	95.6%
Speaker 6	99.1%	96.1%

speech is shown in next equation:

$$R_{Tension}(\%) = \frac{Frame_{Tension}}{Frame_{All}} \times 100. \quad (9)$$

where $R_{Tension}$ is high-tension recognition rate, $Frame_{Tension}$ is the number of frames in which the likelihood of high-tension utterances is larger than that of non-tension utterances, and $Frame_{All}$ is the number of all frames. The recognition rate of non-tension speech is calculated alike. Calculating and comparing likelihood a frame, the proposed system selected the state in which the system outputted a larger value of likelihood. Thus, the system have recognized which model is similar to the input utterance.

The result of recognition experiment a frame is shown in Table IV. The recognition rate of 6 speakers is 80.8%. In this paper, since we have obtained the speech with an obviously different state of the high-tension and non-tension, the recognition rate is expected nearly 100.0%.

Two reasons were offered why the recognition rate is not high. One reason is considered that there are some similar vocal features among high-tension and non-tension utterances. It is considered as another reason that it is too short window length for some utterances to estimate the shape of vocal tract. This is because some utterances include the utterances of consonant. The utterances of consonant cannot extract the state of the throat tuba.

The state of mental tension in autonomic nervous system does not change in short time. It is able to analyse autonomic nervous system function in at least from 9 to 15 seconds. Since we recorded from 1 to 15 seconds utterances, the recognition rate in speech a second is investigated.

The result of recognition a second is shown in Table V. In the case of 1 second, the frame interval is 10 ms so that 100 frames mean 1 second. The system calculated the sum log-likelihood of 100 frames for 1 second analysis. The result shows that the recognition rates is 98.4% a second. It was found from the result that it was reduced the effect

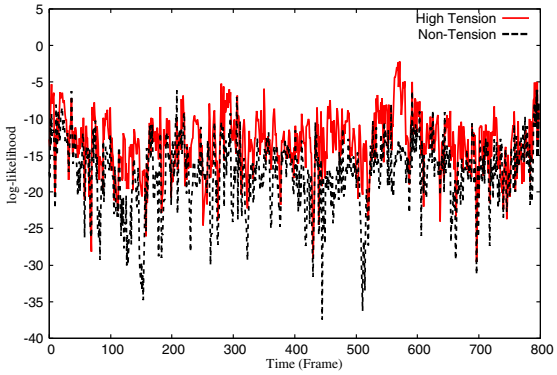

Fig. 3. Temporal transitional likelihood in high-tension utterance

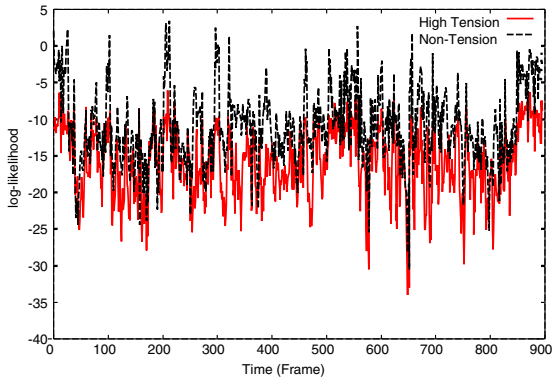

Fig. 4. Temporal transitional likelihood in non-tension speech

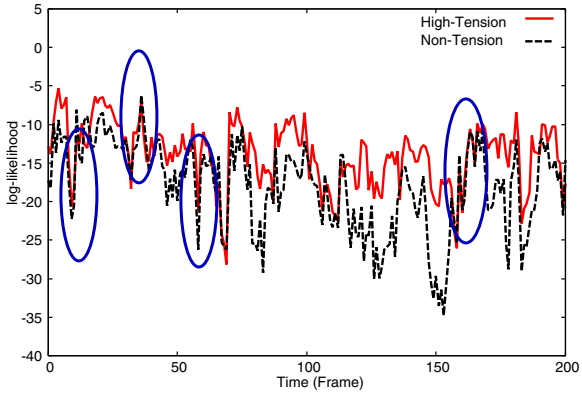

Fig. 5. Detail temporal transitional likelihood from 1 to 200 frame in high-tension utterance

of consonant to recognise in long period. Since recognition rates of every speaker are high, MFCC extracts the feature of mental tension speaker independently.

C. Result 2 : Short Time Analysis

In order to clarify the similarity of high-tension and non-tension utterances, we analysed the temporal transitional log-likelihood a frame. The results are shown in Fig.3 and 4. The log-likelihood of high-tension utterances using each high-tension and non-tension GMM is shown in Fig.3, and that of non-tension speech is shown in Fig.4. The results show that average log-likelihood of the high-tension state in the high-tension utterances is larger than that of the non-tension one. Another result shows that log-likelihood of the non-

TABLE VI
SPEECH DATABASE 2
BOLDFACED CONDITIONS ARE DEFFERENCE ONES FROM DATABASE 1

Speaker	:	**One male and one female students**
State	:	Non-Tension , High-Tension
Speech text	:	"The purpose of this study is,", etc... phrase of **speech and conversation** in Japanese
Speech length	:	Ave. 3sec
Environment	:	**Shield room** with reverberation sound and exhaust fan noise
Sampling frequency	:	16kHz 16bit / monaural

TABLE VII
FRAME NUMBERS OF UTTERANCES EACH SPEAKER

	Non-Tension	High-Tension
Speaker7	3828	5440
Speaker8	4459	4430

tension state in the non-tension utterances is larger than that of high-tension one.

The detail of likelihoods in frames from 1 to 200 in Fig.3 is shown in Fig.5. This shows that there are some frames at the same value of the likelihood. This means that there are some utterances that is similar among two models. In this paper, since we used the non-tension speech, in other words neutral speech, the non-tension state of throat is considered an obviously different from that of high-tension. Since there are some frames in neutral utterances like this, there are larger number of non-tension utterances, such as angry utterances, that are similar state of vocal tract to some high-tension utterances.

V. CONFIRMATION OF HIGH TENSION BY HEART RATES

A. Experimental Condition

In order to back up high-tension state in speech with physiological information, heart rates are obtained with speech.

The heart rates are obtained using Biopac System, model MP-100. It is prepared prepared Speech database 2 shown in Table VI. Different conditions from speech database 1 are listed in boldface. Utterances are recorded by different speakers from former experiment. The shield room is used for recording, there were reverberation sounds and exhaust fan noises.

Experimentation procedures are mentioned to obtain high-tension and non-tension speeches. In order to get obvious high-tension utterances, we have obtained the utterances in the speech task. The speakers were assessed externally by recording the video. The speakers were given 10 minutes to prepare for the speech with anticipatory anxiety. Thus, all the utterances in speech period were labeled as high-tension utterances. Before explanation the detail of the experiment, the non-tension speeches were obtained from conversations with unfamiliar experimenter. All the utterances in conversation period labeled as non-tension utterances. It was

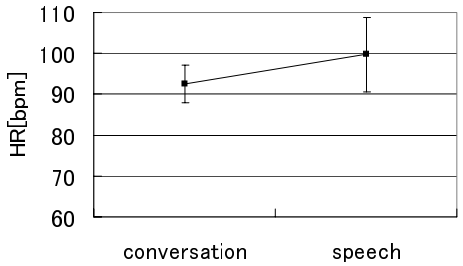

Fig. 6. Average heart rates in each period of speaker 7

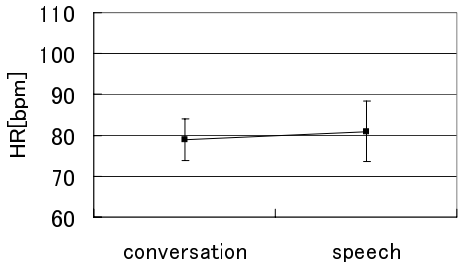

Fig. 7. Average heart rates in each period of speaker 8

TABLE VIII
RECOGNITION RATES OF EACH SPEAKER A FRAME

	Non-Tension	High-Tension
Speaker7	63.5%	69.5%
Speaker8	69.8%	68.4%

avoided obtaining same contents of non-tension utterances as that of high-tension ones to prevent speakers from feeling conversation as pseudo task and from becoming high-tension state. The number of samples is shown in Table VIII. The 80% of utterances are used for training recognition system, and the rests 20% are applied for verifying proposed system.

Other conditions are same as former experiment.

B. Result 1 : Analysis of Heart Rates

The average and standard deviation heart rates of each period are shown in Fig.6 and 7. It was significant for speaker7 at less than 5% level of significance between heart rates of conversation period and those of speech period using t-test. On the other hand, no significant differences were observed in heart rates of speaker8. It is not clarified for both speakers that utterances of speech period were higher tension than those of conversation period.

It is considered that speaker8 was already high tension because of talking with unfamiliar experimenter. Thus, the speaker felt the conversation as one of a task.

C. Result 2 : Recognition Rate of Utterances

The results of recognition is shown in Table VIII. The results show that the recognition rates are 67.8% a frame.

We offer three reasons why the recognition rate is not so high than that of former experiment. First reason is that they were already high-tension state in conversation period. Second reason is considered small number of samples. Third

reason is the differences of contents among high-tension and non-tension utterances. It is also considered that it was not enough to use only heart rates as biomark. It is known that the analysis of heart rate variability such as respiratory sinus arrhythmia can obtain the biomarker which measures autonomic activities[9]. It is necessary to clarify what physiological information has an obvious difference of two tension states.

VI. CONCLUSIONS

In this paper, we proposed a detecting method of mental tension in the speech which use discontiguous sensor based on physiological information. Using Mel-Frequency Cepstrum Coefficient (MFCC) for voice quality feature, we made high-tension and non-tension models which are derived from the shape of vocal tract and the tightness of throat muscle. We used Gaussian Mixture Model (GMM) for learning that can recognise using distribution shape of vocal features.

The result shows that the proposed detecting technique can detach the high-tension speeches from the non-tension ones. We used the non-tension speech, in other words neutral speech. Since there are various non-tension speeches such as angry voices, we should investigate the capability to detach the high-tension speeches from various non-tension speeches.

We could not back up high-tension state in speech with physiological information. Non-tension utterances must be recorded from calmed down person. It is necessary to clarify how to obtain day-to-day physiological information and utterances. Additionally, we will investigate the heart rate variances which associate strongly with autonomic activities.

A further direction of this study will be to clarify what feature of speech is extracted from high-tension speeches and is detached from feature of non-tension speeches. Additionally, we will clarify what physiological information is useful to classify some discrete degrees of tension. The GMM can use for multi-classification and is suitable for modeling utterances by various degrees of tension.

REFERENCES

[1] P. Rani, N. Sarkar, C.A. Smith and L.D. Kirby, "Anxiety detecting robotic system - towards implict human-robot collaboration," Robotica, vol.22, pp.88-95, Dec. 2004

[2] P. Rani, N. Sarkar, C.A. Smith and L.D. Kirby, "Online stress detection using psychophysiological signal for implicit human-robot cooperation", Robotica, vol.20, no.6, pp.673-686, 2002

[3] E. Masahiro, "Effect of Psychological Stress on Pituitary Hormone Secretion," Journal of health science, vol.4, pp159-164, 1982

[4] Andreassi, J. L., *Psychophysiology: Human behavior and physiological response*, Hillsdale, NJ: Erlbaum, 1995.

[5] L.J.M. Rothkrantz, P.Wiggers, J.W.A. van Wees, and R.J. van Vark, "Voice Stress Analysis," Lecture Notes in Computer Science, vol.3206, pp.449-456, Oct. 2004

[6] J.H.L. Hansen, "Analysis and Compensation of Speech under Stress and Noise for Environmental Robustness in Speech Recognition," Speech Communications, Special Issue on Speech Under Stress, vol. 20(2), pp.151-170, Nov. 1996

[7] A.Reynolds and C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech, Signal Processing, vol.3, no.1, pp.72-83, Jan. 1995.

[8] Masaki Ida and Satoshi Nakamura, "HMM Composition-based Rapid Model Adaptation Using a Priori Noise GMM Adaptation Evaluation on AURORA2 Corpus," Proceedings of ICSLP2002, pp. 437-440, 2002.

[9] P.G.Katona and F. Jih, "Respiratory sinus arrhythmia: noninvasive measure of parasympathetic cardiac control," J. Appl. Physiol. Vol 39, 801-805, 1975