Technology to Detect Levels of Stress Based on Voice Information

● Naoshi Matsuo ● Shoji Hayakawa ● Shouji Harada

As a monitoring service technology, Fujitsu has developed a way to process voice signals in order to detect the level of a person's stress based on their conversation, and also developed prototypes of a terminal–center linking system using smartphones, which are increasingly becoming widespread, to evaluate the system's operation. This technology is intended to be used to detect problems between customers and employees at an early stage so as to improve customer satisfaction and to monitor elderly people living alone and people driving vehicles in order to ensure safety, security, etc. This technology to detect the level of stress analyzes changes in the tone of a person's voice, that is, in the pitch and level of their voice, compared with a normal situation. Such changes occur when a person is under stress and has a dry throat. As a result of examining the detection accuracy in an evaluation experiment with simulated conversations in which the subjects were put under stress, it was found that this system can detect the level of stress with an accuracy of 90% or higher. This paper describes the technology to detect the level of stress based on the pitch and intensity of a voice and the configurations of the two prototype systems developed.

1. Introduction

Fujitsu is working to develop a monitoring service that can be applied in households, offices, hospitals and automobiles. The market for this monitoring service in Japan is expected to grow; for example, it is forecast to have a size of 11.6 billion yen in 2015 in the area of keeping an eye on elderly people who live alone, with the number of such households estimated to be over 6 million.[1), 2)] Against this backdrop, Fujitsu is developing technology to analyze telephone conversations as one of the discreet monitoring services applied in a domestic or office environment. When a person receives a nuisance call or has a problematic phone call with clients, the person may be put under stress by the interlocutor's accusations, unreasonable demands, inappropriate language or information (that cause stress to the person receiving the call). We consider that this situation may be relieved quickly if a stressful situation is detected through the person's voice and a third person, who is calm and nearby, is alerted to give that person some support, as shown in **Figure 1**.

We are engaging in research on technologies to detect the level of stress a person is experiencing by analyzing voice pitch and intensity, and catching typical keywords associated with stressful situations, as well as through aggregating data on stress level that is necessary for the development of the technologies. We have also developed some prototypes of the system that link a terminal (using smartphones) and center (FUJITSU Intelligent Society Solution: SPATIOWL). We are now experimenting with the system to identify levels of stress from voice data of telephone conversations.

2. Characteristics of voices under stress

It is believed that when a person is under stress the vocal organs such as vocal cords and related muscles are affected, and so is the mucus density in the respiratory tract.[3)–6)] It is therefore possible that the person's vocal characteristics, such as pitch and intensity levels, may fluctuate due to the affected vocal cord motions.

Figure 2 depicts the differences in these voice characteristics, comparing a normal voice (conversation with a friend) and stressed voice (simulated

problematic telephone conversation). The images represent the sound "ah" extracted and analyzed from recorded conversations. A person's voice is composed of sounds of various pitches. In Figure 2, the voice pitch is shown along the x-axis, and the y-axis indicates the sound intensity. The results show that when a person is under stress the changes in his or her voice's harmonic structure caused by the vibration of the vocal cords do not extend to the high pitched areas of the voice and there is less vibration of the vocal cords.

3. Technology to detect level of stress

3.1 Estimation of level of stress through the analysis of voice pitch and intensity

The results shown in Figure 2 are evidence that suggests stress affects the vocal cord motions. Based on this, we used the pitch variance and intensity level variance to determine the level of such characteristics, closely representing the vocal cord motions, in detecting level of stress.

Here, the pitch refers to the intervals of the wave peaks along the x-axis, while the levels are a square of amplitude, of the analyzed segment, as indicated in Figure 2. To calculate the variance, the average pitch and level are calculated for every conversation unit. Then, as shown in the variance formula, the sum of squares of the difference between the pitch and level values and the mean is calculated, and in addition a calculation is performed to obtain the average of the conversation units.

$$\text{Pitch variance} = \frac{\sum (\text{pitch of the analyzed segment} - \text{average pitch})^2}{\text{Number of the analyzed segment}}$$

$$\text{Level variance} = \frac{\sum (\text{level of the analyzed segment} - \text{average level})^2}{\text{Number of the analyzed segment}}$$

This means that the variance indicates the dispersion of these elements from the average values; hence, the larger the value is, the greater the fluctuation of the pitch and levels, and thus there is a greater degree of vocal cord motions. Conversely, a smaller variance value means that the pitch and levels do not change very much, indicating the voice, and thus the vocal cord movements, are small.

Figure 3 shows the relationship between the pitch/level variance and level of stress, based on data obtained with simulated telephone conversations for male and female voices. As the results show, there is a distinct difference in the pitch/level variance between a normal conversation and one conducted under stressful circumstances, the variance being smaller in the latter case. This indicates the fact that the vocal cord motions

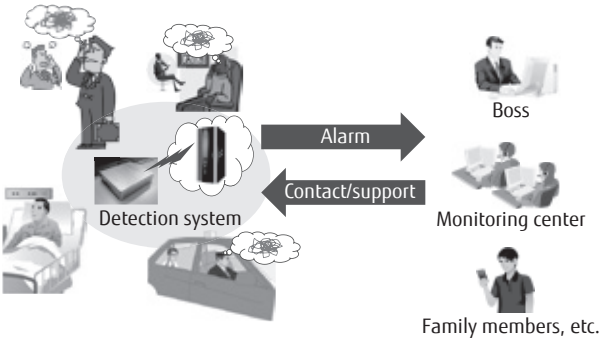


Figure 1
Monitoring service.

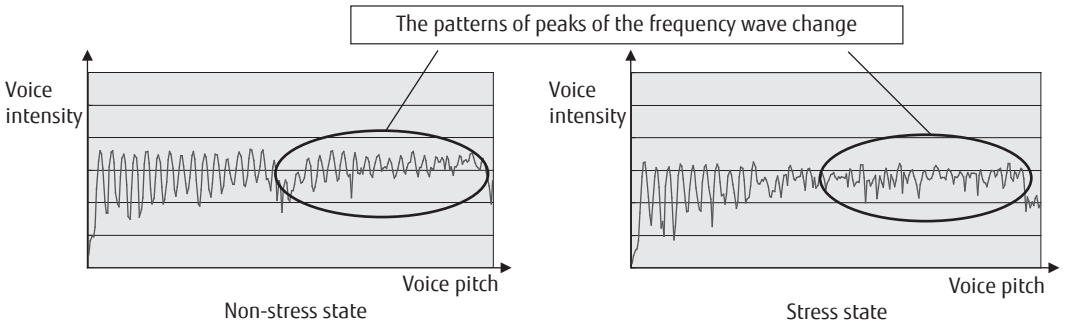


Figure 2
Relationship between the pitch and intensity of a voice.

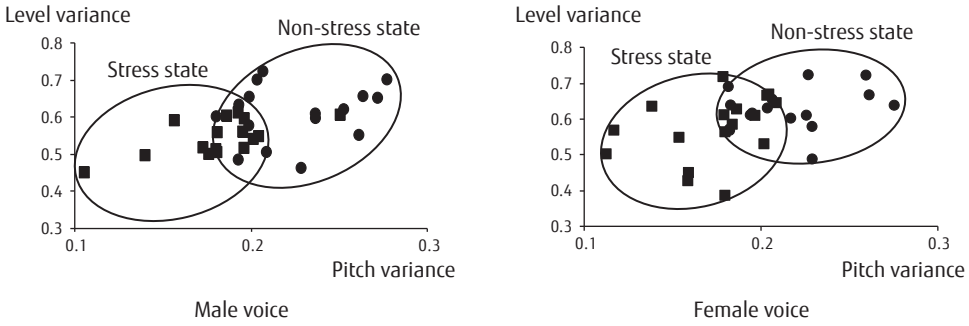


Figure 3
Relationship between pitch/intensity level variances and stress.

become restricted by stress, which is in agreement with the results seen in Figure 2.

Therefore, we can estimate the level of stress in terms of the voice pitch and intensity level variance as follows:

- Both pitch and level variance values are low: high level of stress
- Either the pitch or level variance value is high: low level of stress (i.e., closer to non-stress state)

3.2 Detection of typical keywords using voice recognition

We researched telephone conversations that were closely associated with stress, such as nuisance calls and fraud calls,[7), 8)] and we found some keywords related to money handling and misconduct that appeared frequently in such conversations. Therefore, we decided to apply the voice recognition technology to the stress detection system.

We employed a word-spotting-type voice recognition technology[9)] to identify typical key words in conversational voice data between people and to examine level of stress using the number of detected key words. The keyword lists will vary when deployed in actual services depending on the contexts of target service use, such as office workers or elderly people living alone.

3.3 Detection of stress state

As stated above, the level of stress is inferred through analyzing the pitch and intensity levels of conversational voices, as well as by counting the number of typical keywords using the word-spotting-type voice recognition technology. The results are then weighted and aggregated, and compared with the threshold to judge whether the state is normal or stressful.

Adjustments are made to the voice pitch/level analysis, voice recognition processes, weighting, and threshold for post-aggregation judgement in advance using the data for preparatory adjustments.

4. Evaluation results

We recorded the simulated telephone conversation data several times when a subject talked to a friend to determine a non-stress state, and when a subject had a nuisance call to determine a stress state, and conducted the following experiment.

Subjects and their friends were recruited from the general public, so that the subject group represented well-distributed demographic attributes (age and gender). We employed a male actor to make the nuisance calls. The subjects were not informed of the details of the simulated nuisance call, but advised to hang up if they found the conversation upsetting.

A total of 73 persons recorded their telephone conversations, and the data obtained from them were then processed to detect stress states both from the nuisance calls and conversations with friends, which is shown below. The experimental result gave us a good insight into the establishment of basic technology for detecting stress levels from conversational voice data.

- Stress detection rate: approx. 91%
- Detection in error: 0%

5. System to detect levels of stress

We have developed a prototype of a terminal–center linking system, in view of practical applications for the system to detect levels of stress.

Before implementing the monitoring service, the challenge was to find ways to update software so as to be able to tune various parameters, such as the voice recognition keyword lists and the threshold value that are used when judging whether a state is normal or stressful, without causing disruptions to users' everyday routines. Also, a system must be put in place to centrally manage terminals and cover many users simultaneously. Furthermore, the detection process must be distributed between the terminals and the center according to the type of service and the performance level of the user's devices.

5.1 Blocks of software functionality

To realize a detection of level of stress with a wide variety of configuration patterns for a terminal–center process, we divided the software into functional blocks as shown in **Figure 4**, and defined processes of the blocks, API and data format for exchanging between blocks for implementation.

1) Extraction of the feature quantities

The system calculates feature quantities based on the pitch and intensity levels taken from voice signals, with a conversion process such as Fast Fourier Transform (FFT). The levels are obtained both for detecting level of stress and for use in voice recognition processes.

2) Detection of state of stress

The system detects the state of stress based on the voice pitch and intensity levels.

3) Voice recognition referencing

The system detects the keywords registered in the keyword list.

4) Determination of the level of stress

Based on the results from state of stress and keyword detection, the system judges whether the state is stressed or non-stressed.

5.2 Configuration of prototype system

As stated in the preceding section, software is divided into functional blocks, within which 10 different system configurations can be made. For this paper, we constructed the following prototype system as a simple example (**Figure 5**).

1) Configuration of state of stress detection mainly processed in terminals
2) Configuration of state of stress detection divided between terminal and center

Configuration 1) is suitable for a service in which many user terminals are connected to the monitoring center, such as in a home monitoring service. Compared to configuration 2), it may take longer to update the system with adjusted keyword lists and judgment threshold, but per-terminal data traffic is small, and there is less load on the center.

Configuration 2) is suited to providing a service, where corporate terminals may have certain cost restrictions to connect to the center such as an employee monitoring service for an office. This is because the processing load of the terminal is lower than that of configuration 1) even though the traffic per terminal is high and the center's processing load is high. The center analyzes the detected state of stress based on the feature quantities of the voice data, and the results may be used to educate call-operators in order to enhance customer satisfaction, or to adjust software according to the detected state. Procedures pertaining to each prototype system are as follows:

1) Configuration of state of stress detection mainly processed in terminals

a) Parameters such as keyword lists for voice recognition and threshold for state of stress detection are downloaded to terminals from the center. The keyword lists, for example, the contexts of money handling or corporate misconduct, are arranged according to the types of service use.

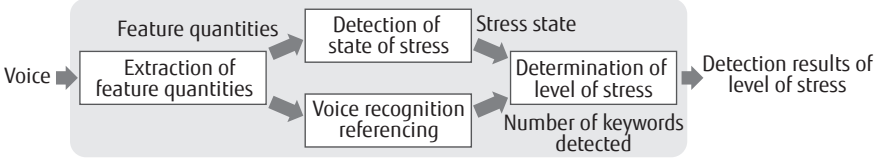
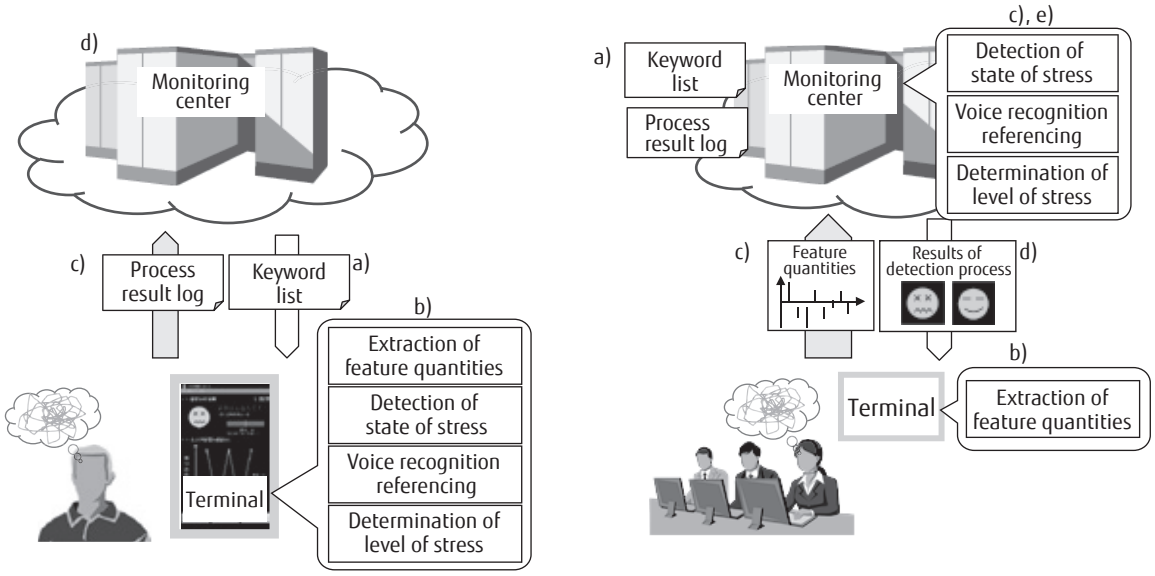


Figure 4
Function blocks of process to detect state of stress.



1) Stress state detection mainly processed in terminals

2) Processing of stress state detection divided between terminal and center

Figure 5
Example of system configurations.

Download requires access to the center. In order to prevent access concentration to the center, the terminal access is restricted, e.g., to once a day, and accessible time for each terminal is assigned randomly.

b) When a conversation is started on the terminal, a series of data processes take place, starting from the extraction of voice pitch and intensity levels, followed by detection of state of stress based on the feature quantities and keyword referencing through voice recognition, and judging level of state. If the result is judged to be a stressed state, the detection system sends an alert to the monitoring center or a third person such as a family member, who can provide calm support.

c) A log file with the results of stress state detection, voice recognition and determination of level of stress is uploaded from the terminal to the center, to be added to the record database of the detection system.

d) To improve the detection accuracy, the database is used to adjust the keyword lists and thresholds, although this process must be handled manually. The adjustments will be reflected in the parameter updates to be downloaded by the terminals.

2) Configuration of state of stress detection divided between terminal and center

a) The parameters of keyword lists and detection threshold are set up at the center.

b) When a conversation is started on the terminal, the terminal extracts the feature quantities (voice pitch and intensity).

c) The feature quantities are uploaded to the center, and processed through the detection of state of stress, voice recognition referencing, and determination of level of stress.

d) The evaluation results are downloaded to the terminals. If the subject is judged to be in a stressed state, just like in the case of configuration 1) above, the center will alert a third person such as a boss so that appropriate support can be given.

e) The results of detecting state of stress and voice recognition referencing conducted at the center, together with the level of stress determination results, are combined with the corresponding feature quantities and recorded in a log file.

Also like 1), the record database is used to update the parameters, replacing less frequent keywords with more frequent ones, and adjusting the detection threshold. However, unlike configuration 1), the

feature quantities are also recorded in the database, making it possible to analyze the stressed state and evaluate the effectiveness of the subsequent adjustments. The adjustments in the center thus made can be applied to process c) above without delay.

5.3 Operational verification of prototype systems

To verify the system operability, the above-stated prototype systems were implemented on smartphones (CPU clock frequency 1 GHz; RAM 512 Mbytes) as terminals and a PC was used to process the center functionality.

With the recorded voice data as an input, the evaluation test yielded results indicating the same accuracy levels in detecting stressed state for configurations 1) and 2).

We also conducted a test on actual live conversations, and analyzed the resource usage of the smartphones for running the detection processes in parallel with conversations. The results are as follows:

1) CPU usage: approx. 53%
   RAM usage: approx. 31 Mbytes
2) CPU usage: approx. 16%
   RAM usage: approx. 18 Mbytes

The experiment suggested that the monitoring service system could be configured according to the service types and terminal performance configurations.

6. Conclusion

We developed a technology to detect state of stress from voice data in order to help create a monitoring service, and verified that the system prototypes operated effectively. We will further advance the system so that it can be commercialized for personal and corporate services in collaboration with other divisions. We will also look into the possibilities of integrating different types of information, such as images and vibration, so that the technology can be applied in various situations, such as in a noisy in-vehicle environment of a fast-driving car. We will also aim to expand the technology to see if it can also be adapted to house noises and mechanical noises and detect malfunctions of appliances in houses and plants.

This project was partly sponsored by the Core Research for Evolutional Science and Technology (CREST) funding by the Japan Science and Technology Agency (JST).

References

1) Seed Planning, Inc.: Market size for monitoring services (in Japanese).
http://www.seedplanning.co.jp/press/2011/2011092801.html
2) Cabinet Office of Japan: Population forecast of Lone Seniors (in Japanese).
http://www8.cao.go.jp/kourei/whitepaper/w-2013/zenbun/pdf/1s2s_1.pdf
3) G. Zhou et al.: Nonlinear feature based classification of speech under stress. IEEE Trans. on Speech and Audio Processing, Vol. 9, No. 3, pp. 201–216, 2001.
4) T. Kitamura: MRI observations of larynx deformations through emotional changes. The Twenty-Third General Meeting of the Phonetic Society of Japan, submitted papers, pp. 45–50, 2009 (in Japanese).
5) H. Nakagawa et al.: Lubrication mechanism of the larynx during phonation: an experiment in excised canine larynges. Folia Phoniatr Logop, Vol. 50, No. 4, pp. 183–194 (1998).
6) R. E. Witt et al.: Effects of surface dehydration on mucosal wave amplitude and frequency in excised canine larynges. Otolaryngol Head Neck Surg, Vol. 144, No. 1, pp. 108–113 (2011).
7) National Consumer Affairs Center of Japan: Information on Monitoring (protecting seniors, children and people with impairment from troubles) (in Japanese).
http://www.kokusen.go.jp/mimamori/index.html
8) National Police Agency: Fraudulent phone crime - incidence and amount of damage (Fraudulent phone crime combat website) (in Japanese).
http://www.npa.go.jp/safetylife/seianki31/1_hurikome.htm
9) Animo Limited and Fujitsu Laboratories Limited: Japan's first keyword-type voice search software "VoiceTracking/KeywordFinder" launched in the market (in Japanese).
http://pr.fujitsu.com/jp/news/2009/02/20-1.html

Naoshi Matsuo
Fujitsu Ltd.
Mr. Matsuo currently engages in R&D of sound processing, voice recognition processing and non-verbal information processing.

Shouji Harada
Fujitsu Ltd.
Dr. Harada currently engages in R&D of voice recognition processing and non-verbal information processing.

Shoji Hayakawa
Fujitsu Ltd.
Dr. Hayakawa currently engages in R&D of sound processing and non-verbal information processing.