Emotion Voice Analysis System Connected to the Human Brain

Shunji Mitsuyoshi, Kouichi Shibasaki
AGI.Japan.Inc. / SGI.Japan.Ltd.
Yebisu Garden Place Tower 31F 4-20-3
Ebisu Shibuya-ku, 150-6031
Tokyo Japan
shunji@agi-web.co.jp

Yasuto Tanaka,
Makoto Kato,
Tsutomu Murata,
Tetsuto Minami,
Haruko Yagura
NICT
588-2, Iwaoka, Nishi-ku,
Koube-shi, Hyougo, Japan
ytanaka@po.nict.go.jp

Fuji Ren
Tokushima University
770-8503,1-1 Minamijousanjima-chou,
Tokushima-shi, Tokushima, Japan
ren@is.tokushima-u.ac.jp

Abstract
To investigate human brain activities in association with emotional speech, we developed a novel voice analysis system connected to a functional Magnetic Resonance Imaging (fMRI) machine. Participants spoke inside the MR magnet during which BOLD activities of the brain was measured. Speech voice was transmitted through the newly developed mask-microphone inside the magnet to the external computer and was processed by the emotional voice analysis system. Two participants conversed without hindrance and their emotional state was analyzed. Using the system, we were able to detect brain activities during speech and simultaneously evaluate the human emotional voice.

1 Introduction
Human emotion is difficult to investigate. There are several ways of studying emotion by visual stimulation such as pictures, drawings or motion videos, etc (LeDoux, 1996). Another way of studying is to use sound stimulation, namely human voice that contains emotional elements. In a previous study, we developed the voice analysis system that determines human emotional speech. The system allowed us to detect emotional elements of human voice such as fear and joy as well as excitement. Human voice contains prosodic information derived from intonation and fundamental frequency (F0) which is generated at relatively low levels of processing. Since voice production is carried out by vocal cords and shape of voice road (Figure1), vocal cords are connected to the autonomic nervous system, we used human voice, particularly the fundamental frequency (F0) as a parameter of extracting basic human emotion. For a detail description of the system see Voice Emotion Analysis System, (Mitsuyoshi, 2006)(Mitsuyoshi et al., 2006). One problem of emotion studies in general is that emotion is fragile in itself including a large amount of variables. For instance, our previous results showed that only 70% of emotional elements were correctly matched between the analysis system performance and the speaker’s subjective utterance. This may be partly because the emotion process contains several sub processes or stages including simple reaction to physical stimuli, detection of basic emotion, recognizing emotion, and labeling emotion, etc (Schachter, 1964)(Schachter and J.E.Singer, 1962). In order to classify such levels of processing and to investigate the actual process that occurs in humans, it is necessary to extend the research scope from mere voice recognition to the human brain and neuroscience. In this research, we constructed a system of studying neural conditions of human emotion during voice production and conversation, by means of functional Magnetic Resonance Imaging. We connected the emotional voice analysis system and the fMRI machine with which corresponding brain activities can be identified. In the following section, both the emotion voice analysis system and MRI are described in detail.
2 Emotion Voice Analysis System

We used AGI.Japan.Inc. “Sensibility technology ST Ver2.0”(ST2.0), a software for emotion estimation, for determining the emotion parameters in speech recognition. In this study, we investigated the voice generation process, particularly emotional voice. Emotional speech is influenced by the amplitude of voice and by tone. Here we focused on the vibration of vocal cords and their power which may be an expression of natural human emotion. The reason is that the autonomic neural system may be related to the function of vocal cords, which is independent of high-level language processing that occurs in the cortex. Figure 1 shows the relation between vocal chords and emotion.

2.1 Emotion Parameters

As a determinant of analyzing emotional feelings, emotional elements were determined as including anger, joy, sorrow, and calmness. The feeling of excitement was also determined. Because excitement was included in most elements of emotional feelings such as anger or joy (Schachter, 1964) (James, 1890)(Cannon, 1927) Note that excitement is an independent measure because it is classified as low-level emotion. To evaluate the system performance, we asked naive subjects to subjectively classify various recorded voices and we made a database from the subjective classification. To determine emotions, we made a decision tree based on the parameter of F0 and amplitude of the voice, using a data mining tool C5.0. The decision tree was based on the human subject’s judgments of emotions, and an if-then rule was constructed from these judgments. We needed to learn how "Excitement", that is Emotion, influences "Anger", "Joy", "Sorrow" and "Calmness". A common parameter to each feeling was confirmed to gauge the influence of the excitement. Figure 3 shows the feelings voice separated by using the intonation parameter (horizontal axis) and the fundamental frequency parameter (spindle). A yellow point shows the position of the voice file evaluated by the evaluator as joy. A red point shows the position of the voice file evaluated by the evaluator as anger. A blue point shows the position of the voice file evaluated by the evaluator as a sorrow. A green point shows the position of the voice file evaluated by the evaluator as calmness. Joy and anger can be separated in B in Figure 3. Joy and calmness can be separated in A. The aspect distributed in the area where anger and normal are the same is understood from the comparison between A and B.
It is thought that A and B are the comparisons which separate between unpleasant or pleasant. Calmness and anger can be separated in C. Sorrow and anger can be separated in D. Sorrow and calmness can be separated in E. Recording is done simultaneously with real time evaluating, the resulting label from the feelings judgment is added, and the evaluation process is recorded. As a result, the judgment of the system can be evaluated while reproducing after the vocal person and the other party of the conversation end the conversation.

3 Experimental procedure

The experiment was done with a following schedule. The procedure of conversation was described next. Whole communication system including the MRI was described in a following section.

3.1 Experimental Schedule

The experiment started with a functional MR scan. We had three conversational sessions for five minutes each with one-minute rest intervals. The rest interval was set to compare MR scans with and without conversation. Generally, subjects tend to move their heads during conversation. By comparing the session with and without conversation, we can evaluate the effect of head motion during conversation. The actual time course was the following: first, subjects waited for a minute while the MR scan began. Then, the conversation session started for five minutes. One minute rest session followed, and the second conversation session (five minutes), the second rest session (one minute), the third conversation session (five minutes), and the third rest session (one minute). The MR scan ceased and the last conversation session was held. This was done to compare the effect on emotion voice analysis performance with and without MR noise. The whole experiment took twenty one minutes.

3.1.1 Making natural conversation

The subject and the experimenter held natural conversation. Subjects were inside the MR magnet and the experimenter spoke to them from the outside through a microphone. Conversation was held naturally without any protocol decided between the experimenter and the subject. It was difficult for the experimenter to deliberately control the subject’s emotion and to make him excited. That is why we chose the third person, referred here as a “di-
rector”, who supervised the whole conversation and regulated it by putting posters in front of the experimenter. For example, the poster said "Make the subject angry" or "Remind the subject of the situation where he was upset before”. When the conversation sounded flat and boring we tried to make the whole conversation more excited by stimulating with the poster. The director was one of the authors (KS) who had more than three years experience of analyzing emotional voice. Conversation was recorded separately for the experimenter and the subject. Emotional voice analysis was done both for the experimenter and the subject. Here, we focused on the feeling of excitement and anger because of their basic nature of emotion. Our goal here is to correlate such elements of emotion to brain activities. We attempted to detect the time of excitement as well as anger using the voice analysis system. MR scans were carried out at the same time for such emotional states. The control scan was carried out to detect the time when no emotional elements were detected. Comparison of these states could create the brain image. In the following section, the MRI system as well as the sound communication system was described.

4 Functional Magnetic Imaging System

We used the functional magnetic imaging (fMRI) system (Trio, Siemens Inc., Germany) to measure brain activities, namely, blood oxygen level dependent (BOLD). The MRI can detect brain activities with an excellent spatial resolution with a magnitude of millimeter range by detecting cerebral blood flow when a cognitive task is performed by the subject. Detailed neural structures are also identified by the structural scan. A trigger pulse (5V pulse wave, 5 msec) was generated from the parallel port on a PC controlled by the experiment control software called ”presentation” (Neurobehavioral Systems Inc., USA). The MRI system included the superconductive coil, RF coil, and the head coil for detecting the magnetic field change of the living body, auditory system (air headphone and microphone for the conversation). There are several drawbacks in making a conversation experiment in the MRI machine. For example, production of strong magnetic fields causes vast amount of sound noise, i.e. about 130 dB of sound noise. This noise was a major obstacle in making a conversation in the MRI experiment. Detection of human voice in such loud noise was very difficult. The MR imaging condition for the functional scan (T2*) was as follows. TR=4 seconds, TE=30ms, slice=32mm, voxel size=3x3x4mm, slice thickness=4 mm, FOV 192 mm delay in TR=500 ms, Bandwidth=1562Hz/Px, Echo space=0.76 ms and flip angle 79 deg. The condition for the anatomical scan (T1) was the following. TR=2 seconds, TE=4.38 ms, TI=990 ms, slice thickness=1 mm, boxel size=0.9x0.9x1 mm, FOV=240 mm, FOV phase=75%, distance factor 50%, phase over sample 8%, slices slab 208, average 1, flip angle 8 deg, bandwidth 130 Hz/pix, and each space 9.6 ms. Data was analyzed by the software SPM2.0 (University College London) on Matlab (MathWorks Inc., USA).

4.1 Solution of fMRI noise

In order to solve the above questions, we made a novel microphone which masked a large amount of the MR noise.

4.1.1 Mask-microphone

First, we disassembled a ready-made gas mask (Toyo Co. Ltd, Japan, SAFETY, no. 1880-1, TN242, GM small) to prevent MR noise from getting to the microphone. The gas mask had a rubber, sealed cover over mouth and nose. It had a respirator made of plastic and a particle filter. To reduce the sound noise, first silicone was pumped in between the rubber, plastic and mouth. A central hole was made to transfer the voice sound to the microphone. Two sheets of paper towel (kitchen paper, Nepia Co. Ltd, Japan) were embedded in between the rubber and the silicon. The non-magnetic microphone (Hitatchi Medico. Co. Ltd, Japan) was embedded and fixed tightly in front of the mouth. A central hole was made to transfer the voice sound to the microphone. Two sheets of paper towel (kitchen paper, Nepia Co. Ltd, Japan) were embedded in between the rubber and the silicon. The non-magnetic microphone (Hitatchi Medico. Co. Ltd, Japan) was embedded and fixed tightly in front of the mouth. Secondly, we built a sound absorption structure between the microphone and the plastic filter. We made the sandwiched structure composed of two kitchen sponges, and three sheets of paper towel, which made five layers of covers. This was done attempting to reduce the outside vibration noise from affecting the microphone. Figure 11 shows the materials we used for the mask microphone. Since the nose was open, the subject respired through nose.
4.1.2 Head fixation

Another obstacle in the MRI experiment was the body and head movement during conversation. The act of speech moves mouth and chin and produces skull movement. This results in major noise for BOLD detection. To solve this problem, we developed a head fixation system. We made a molded head-shape using expanded polystyrene fitted for each subject. The polyurethane package (Intapac Quick, Sealed Air Japan, Ltd, NJ, USA) was used to make a mold head. The polyurethane was seated on the half spheroid shaped mold with heated air inside that expands for two minutes. Subjects lay their heads down on the polyurethane inside the sphere to mold their own head shape. Note they wore a non-magnetic headphone (Hitatchi Advanced Systems, Japan), to fix their head more firmly to the mold. The outside shape of the sphere was fit to the birdcage attached to the MR system. A nylon belt was used to further tighten the mold to the head, ensuring the subject did not feel uncomfortable. The mold was fixed from outside by the usual fixed head bars that were attached to the birdcage. With this configuration, the subject’s head was fixed to the birdcage as effectively as possible (Figure 8).

4.2 The sound communication system

The sound communication system was described in Figure 9. The subject speech system (Microphone and Mask-microphone) was connected through the filter and the microphone amplifier to the experimenter hearing system which was composed of the speaker and headphone. The voice analysis system for the subject was connected at this level. The experimenter speech system was connected through the microphone amplifier, mixer, and headphone amplifier, and filter to the subject hearing system, which is the headphone in the MR magnet. The voice analysis system for the experimenter was connected to the experimenter speech system.

4.3 Subjects

Two subjects (TO, TN) participated in the experiment. They did not know the purpose of the experiment. The experimenter was YT (author). The subject and experimenter knew each other for at least two years. Experimenter and subjects were all the same sex (male). Both subjects were experienced.
5 Results

The result section is composed of three parts, (I) voice sound acquisition in terms of gain adjustment and fundamental frequency, (II) success of natural conversation and emotional voice analysis, and (III) the evaluation of head motion and the image construction of the brain.

5.1 Gain adjustment

Since the mask-microphone shielded the front part of the mouth, the air from the mouth did not expand. As a consequence, the amplitude of the voice tended to be magnified. In addition, the MR noise masked the voice when the MR machine was operating. To solve these issues, first we operated the MR machine in which the proper gain of the voice amplitude was adjusted through the amplifier. Figure 7 indicates the sound power without the mask-microphone, with the microphone without gain adjustment, and with gain adjusted. Frequency analysis indicates that the mask reduced the noise significantly in high-frequency areas with a magnitude of about 80 dB. The gain adjustment had an effect particularly in the low-frequency range of less than 200Hz. The noise reduction between 200-1000Hz suggests that the silicon structure had some effects on fundamental frequency (F0).

5.2 Fundamental frequency

In this section, fundamental frequency was evaluated inside the MR magnet. Figure 10 is a comparison of the distribution of fundamental frequency (F0) after noise reduction using the Cepstrum method and using the emotional voice analysis system (ST Ver 2.0). We successfully detected the fundamental frequency during the operation of the MR machine. Auditory parameters were detected and they could be analyzed for emotional determination. In contrast, F0 using the cepstrum was scattered. This assures the effective analysis of the voice emotion system.

5.3 Natural conversation

We employed three sessions of five-minute conversation (total: 900 seconds). After a one minute break each conversation was implemented naturally without any delay. The emotional conversation was controlled by the director, who asked the experimenter to elicit excitement three times with TO, and four times with TN. Figure 11 demonstrates the frequency of
emotional events, namely excitement, laughter, and anger. The voice analysis system indicated significant frequency of excitement for subject TO. The frequency of calmness and laughter was also significant. The contents were student life and friends in daily life. In contrast, subject TN showed little excitement while significant frequency of anger and calmness. No laughter was detected. After the experiment, an interview was held with each subject. TO expressed a fun feeling for the experiment. TN suggested the relative loss of oxygen during the experiment made him feel negative during the conversation.

5.4 Evaluation of head motion

The head motion was evaluated for each subject. The translation of the head position was within +/-0.2mm except around 230 scans for TN. There was little rotation for roll and yaw directions, yet around 0.6 degree rotation was found for pitch direction (Figure 13). This may be due to laughter during the free conversation. The magnitude of head motion is less than 10% (3 to 4mm) of each voxel. A similar tendency of head motion was found for the subject TO, with the magnitude of +/-0.18mm for direction of x, y, z axis.

5.5 Constructing the brain image

Based on above evaluation of head motion, the brain image was constructed after adjustment of the realigned components of spatial normalization. As is seen in Figure (Figure 14), the head motion was absorbed in sagittal, coronal, and axial directions each with the constructed filters (left panels), and consequently the normalized brain image was successfully constructed. This image is a basis to map functional activities of the brain that corresponds to emotional elements such as excitement and anger.

6 Discussion

We set up a novel system with which people are able to make emotional conversation inside the MR magnet. The MR voice and sound system including the newly developed mask-microphone was able to transmit the human voice between the experimenter and the subject with significant (80dB) noise reduction. Natural conversation was successfully conducted between persons inside and outside the magnet. The large MR noise and the noise from the head motion was carefully controlled to achieve the MR image of the brain during the conversation. As a consequence, subject’s speech voice was success-
fully detected during the MR detection of brain activities and the analysis of the voice was done to classify emotional elements. The head motion was quantitatively evaluated. The motion artifact was small enough so that the basic brain image was successfully constructed after the spatial normalization (procedure of the Statistical Parametrical Mapping).

Yet, to construct functional images corresponding to emotional elements is still under development. Importantly, conversation contains different types of cognitive processes such as hearing, speech, and imagination of speech (inner speech). In this study, we focused on speech activities picked up by the voice analysis system. However, it is possible that brain activities of hearing, inner speech, and speaking overlap each other while subjects speak. It may be necessary, in future research, to dissociate these hemodynamic factors in some way. Furthermore, the act of speaking itself could produce various kinds of noise in the image which should be masked out. This time we tested only two subjects, yet, to obtain plausible functional brain activities as well as evaluate statistical differences between subjects, we need to test at least double digit numbers and average data from them.

In this study, we focused on the detection of the fundamental frequency. However, it is important to find other formant information such as F1 or F2, for consonant detection. These factors are critical to analyze other emotional elements such as joy or sorrow. For this, it is necessary to modify the mask-microphone to improve the detection of such information. In our previous study, we constructed a similar head fixation system (Mitsuyoshi, 2006)(Mitsuyoshi et al., 2006), in which the subject’s head was also fixed tightly by a nylon belt to the bird cage. However, tightening the head turned out to make subjects uncomfortable and as a consequence they felt pretty stressed, therefore the tightening method could not be used for naive subjects. In the current study we were successful in conducting the experiment without that level of belt tightening with a decently small magnitude of head motion.

We used the voice analysis system to detect emotional elements and their timing. The timing was used as a time stamp. A similar time stamp could be obtained by means of physiological measures such as heart-rate or respiratory rate. Further research is needed.

6.1 Importance of the hearing and expressing voice in communication

Human voice is important for transmitting lexical information as well as emotion. In this study, we focused on the condition where human subjects speak and at the same time hear what they speak. The brain and voice were analyzed simultaneously. However, it is not clear how we dissociate effects of the subject hearing and speaking at the same time. There is a theory of feedback loop during the conversation.
which is connected to motor theory of conversation. Further research is necessary for this issue.

7 Conclusion

We made the emotion voice analysis system connected to an MRI machine. This system was able to detect the human voice of speakers inside the MRI by reducing the MR noise significantly, and construct the brain images during the conversation that contained emotional voice.

8 Acknowledgement

We appriciate Seishi Itoi, Takeshi Nogai in NICT, and Fumiaki Monma in AGI.Japan. Inc for their technical assistance to construct the MRI sound system. We also thank Guoxiang Liu in NICT for help of constructing a novel MR sequence.

References


