Acoust. Sci. & Tech. **33**, 6 (2012)

PAPER

Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment

Yoshiko Arimoto*,†, Hiromi Kawatsu‡, Sumio Ohno and Hitoshi Iida

Tokyo University of Technology, 1404–1 Katakura, Hachioji, 192–0982 Japan

(*Received 22 November 2011, Accepted for publication 21 May 2012*)

Abstract: For the purpose of constructing a naturalistic emotional speech database, a novel paradigm of collecting naturalistic emotional speech during a spontaneous Japanese dialog was proposed. The proposed paradigm was assessed by investigating whether the collected speech contains and conveys rich emotions psychologically and acoustically. To encourage speakers to experience and express their natural and vivid emotions, a Massively Multiplayer Online Role-Playing Game (MMORPG) was adopted as a task for speakers. They were asked to play the MMORPG together while discussing strategies to achieve their tasks through a voice chat system. The recording was performed for one hour per speaker. The total recording time was approximately 14 hours. The results of emotional labeling for the collected speech supported the validity of the paradigm showing higher interlabeler agreement than the chance levels. In addition, it was revealed that the paradigm is superior in the quantity of emotional speech to other paradigm by showing a significantly higher rate of labeling instances for our speech material (73%, $\chi^2(2) = 27659.87$, $p < 0.001$) than other speech materials. Finally, an acoustical analysis supported the validity of the paradigm, showing a significant difference between the nonemotional utterances and the emotional utterances ($p < 0.05$).

Keywords: Emotional speech, Spoken dialog, Online game, Voice chat, Acoustic analysis

PACS number: 43.66.Yw [doi:10.1250/ast.33.359]

1. INTRODUCTION

For automatic speech recognition and speech synthesis to continue to improve, these technologies must be capable of handling emotional information to facilitate human-computer communication and computer-mediated communication as the next step toward advanced functions. In early studies on emotional speech, acted emotional speech was studied to investigate the acoustical correlation with emotion [1–5]. However, computational research can benefit from naturalistic speech material, rather than acted speech, since its applications are designed for the real world, not for laboratory settings. "Naturalistic" emotional speech is the opposing term to idealized emotional speech, which is generated to match someone's conception of what an emotion should be like, according to [6]. One example of the computational research is emotion recognition. Emotion recognition could be applied to human-computer interaction systems in a real world, such as a contact center or an online game. To realize emotion recognition a naturalistic emotional speech database of real-life conversation is dispensable. An application of the naturalistic emotional speech data base to emotion recognition enables one to reveal the characteristics of naturalistic emotional speech and to model naturalistic emotional speech for emotion recognition from real-life conversation. Another example is emotional speech synthesis, which often adopts acted emotional speech. The acted speech includes intense emotional expression and is appropriate for synthesizing prototypical example of one emotion. However, it is too stylized to synthesize delicate nuances of expressions. Naturalistic emotional speech enables one to synthesize more natural and genuine emotional speech that conveys delicate nuances of human emotional expression [7]. Automatic speech recognition (ASR) is also another example that requires a naturalistic emotional speech database. Although current ASR has achieved more than 95% accuracy for reading materials [8], recognizing the verbal content of emotional speech is still a difficult problem [9]. It is expected that a natural emotional speech database would be applied for speech recognition system to address this issue, however, the lack

*e-mail: ar@brain.riken.jp
†Current affiliation: JST, ERATO, Okanoya Emotional Information Project
‡Current affiliation: IBM Japan, Ltd.

of appropriate speech material is one of the main obstacles [9].

Some research groups began studying naturalistic emotional speech in the 2000s [7,10–18]. The main paradigms of collecting naturalistic emotional speech for such research are a speaker's emotion induced by miscommunication during a human-computer dialog or a Wizard-of-Oz scenario. For example, dialogs between the AutoTutor system and students [12], dialogs between a pet robot and child [16], and interviews where the speaker's emotions are controlled by an experimenter [7]. Devillers and Vidrascu dealt with real conversations in a phone call to a call center [11]. Moreover, some attempts have been made to study Japanese emotional speech with naturalistic materials [13,17,18]. Zeng *et al.* [19] and Cowie [6] precisely reviewed the history of an emotional speech corpus and provided suggestions to construct a naturalistic emotional speech corpus. It was pointed out that there was a critical issue concerning naturalistic materials [6]. Very few emotional samples were contained even in a large speech corpus. Campbell recorded phone conversations and labeled each utterance with an emotion perceived from the recorded utterances [17]. They successfully recorded the naturalistic conversations, but very little of the speech displayed any strong emotional content. Ang *et al.* also obtained little emotional speech even though they dealt with approximately 22,000 utterances collected from a pseudodialog [10]. It indicates that the methods of emotion induction in these studies were insufficient to collect naturalistic emotional speech and a novel paradigm is required to record naturalistic expressive speech that contains spontaneous emotions of speakers.

We propose a new paradigm to collect naturalistic emotional speech. To induce speaker's emotion, a Massively Multiplayer Online Role-Playing Game (MMORPG) is adopted with reference to the research on emotion and decision making [20–22] and emotion studies for entertainment computing [23–25], and the subject's emotional response to the game is successfully induced. As our analytical interest is focused on speech, our goal is not just to induce the speaker's emotion but also to record naturalistic expressive utterances containing spontaneous emotion evoked in a usual environment for speakers. Our paradigm also adopts a voice chat system as a communication tool for game players to talk with each other while playing the game. The aim of introducing a voice chat system to our paradigm is to make players talkative and to collect expressive utterances influenced by players' emotions induced by MMORPG. To prove the validity of our paradigm, dialogs between online game players are recorded. The recorded speech material is examined as to 1) whether the collected emotional speech is perceptually distinguishable from each other, 2) whether the recorded material is superior in the quantity of emotional speech compared with other emotional speech databases, and 3) whether the collected emotional speech has distinctive acoustic characteristics in accordance with perceptual emotional categories, by conducting emotional labeling and emotional intensity rating and acoustical analyses for recorded utterances.

The following sections describe the validity of our paradigm for collecting naturalistic emotional speech. In Sect. 2, the proposed collection paradigm and the speech material recorded according to the paradigm are explained. In Sect. 3, we assess whether the collected speech material contains emotional speech that is perceptually distinguishable from each other, and whether the collected material contains a high proportion of emotional speech compared with other emotional speech databases. In Sect. 4, we prove that the recorded utterances are acoustically distinct from nonemotional utterances by conducting acoustic analysis. Section 5 concludes the paper.

2. EMOTIONAL SPEECH COLLECTION PARADIGM

2.1. Approach to Collecting Naturalistic Emotional Speech

To record naturalistic expressive utterances containing spontaneous emotion in a laboratory setting, the paradigm has the three characteristics below:

- emotion inductivity,
- emotion expressivity,
- environmental ordinariness.

Emotion inductivity is the characteristic that the proposed paradigm can effectively induce the speaker's emotion. The paradigm adopts a MMORPG as the task for speakers to induce their emotion. The effect of a game to induce emotion was proved in earlier studies [20–25]. Emotion expressivity is the characteristic that vocal expression is influenced by induced emotion. To encourage game players to speak with each other and to express their emotions in their voice, a voice chat system was adopted as a communication tool between game players. Players of the MMORPG would usually discuss their strategies to collaboratively achieve their goals in game events, through a text chat function provided by the MMORPG. To reflect the emotional reactions in their speech, the proposed paradigm asks players to communicate through a voice chat system, not the text chat function. The players' emotional reactions to game events and the players' expressive speeches influenced by their internal emotions are expected to be observed. Although wearing headset microphones to talk with each other over an online voice chat system seems an unusual situation, there are an increasing number of online game players who make use of the voice chat system, not a text chat function while

playing games. In fact, voice chat was the normal mode of communication for some online game players who actually participated in our recording. Moreover, it was found, from our questionnaire for the participants, that 82% of them wanted to make use of the voice chat system while playing a game. It suggests that the use of the voice chat system during an online game is not an unnatural experience and is more convenient or comfortable for players than the text chat function. Environmental ordinariness is the characteristic that recording environment should be close to their usual environment of online game playing. To set a usual environment for game players, the players were located at remote sites alone to play the online game. It is quite different from face-to-face conversation, but it is not so unique because some dialogs, such as a conference call or a contact center, operate under the same non-face-to-face online environment as that of our paradigm. The proposed paradigm enables us to record naturalistic expressive utterances containing the spontaneous emotion of speakers who talk about and play an online game in the same environment, even when speakers are located in a laboratory setting.

2.2. Recording

2.2.1. Speakers

The speakers were 13 university students (9 males, 4 females, mean age 22 years ($SD = 1.17$)) with experience playing online games. They participated in our recording as online game players. The players had to participate in the recording as a group with one or two friends of the same gender to make an atmosphere favorable for expressing their emotions easily. A group of the opposite gender was prohibited from to participating in the recording in order to avoid the gender effect on their expression. Six dialogs (five dyadic dialogs and one triad dialog) were recorded. The mean number of months of online game experience per player was 38 months ($SD = 14$) and the mean playing time per month was 33 hours ($SD = 35$). The origin of players were different: six from Tokyo, two from Nagano, and each one from Kanagawa, Shizuoka, Yamanashi, and Aomori. One did not specify his origin.

2.2.2. Recording environment

Figure 1 shows our recording environment. Each player of a group was located in remote laboratories on the campus of Tokyo University of Technology and joined an online game together via the Internet. The players were asked to wear a headset microphone (Audio-Technica Dynamic Headset ATH-30COM) and to talk with each other in a non-face-to-face environment with the voice chat system, Skype [26]. The dialogs among the players were recorded with the voice-recording system, Tapur, for Skype [27]. The dialogs were recorded at all laboratories where each player played the game. Tapur recorded the local player's voice and a remote player's voice completely isolated in individual channels of a stereo sound file. As the local player's voice was recorded directly into the recording PC, there was no distortion or encoding effects caused by transmission via the Internet. Before starting the recording, the players were asked to talk freely with each other to become accustomed to the recording environment. While they were talking freely, the distance between the microphone and the player's mouth was fixed to adjust the input level of the microphone.

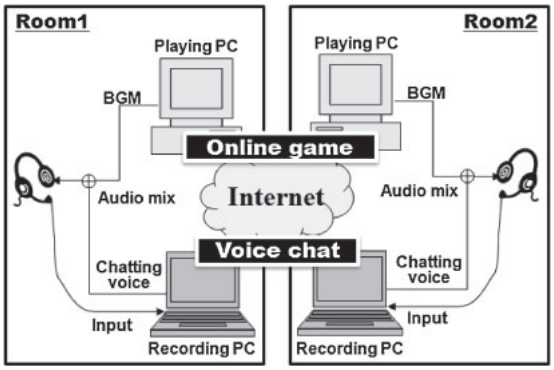

Fig. 1 Recording environment.

The title of the MMORPG used for the recording depended on each group of players. They could choose a game that more than one of the players in each group had actually played and enjoyed in their daily life. The most popular online game was *RagnarokOnline*, which three groups played during the recording. *MonsterHunterFrontier* and *RedStone* were chosen by the other groups. All players were asked to form a party and to participate in quests (tasks in the game) together in the game and to keep talking with each other until the end of the game. They were prohibited from using the text chat function provided by the online game system when they talk with each other. They were allowed to use the text chat function only when they were addressed by a player who was not a participant in the recording.

The recording time was approximately 1 hour for each group, and the total recording time was approximately 14 hours. The sound data were sampled at 48 kHz and digitized to 16 bits.

2.2.3. Segmentation and transcription

The recorded dialogs were segmented into utterances. The inter-pausal unit (IPU) was adopted as the unit of segmentation, which enabled us to segment utterances automatically. Any continuous speech segment between pauses exceeding 400 ms was regarded as a unit of utterance for our dialogs, although some speech corpora, such as the large-scale Corpus of Spontaneous Japanese (CSJ), adopted a 200 ms pause duration for IPU segmentation [28], because the utterance became too short when

Table 1 Number of utterances for each speaker.

Speaker	Utterances	Speaker	Utterances
01_MMK	816	04_MNN	934
01_MAD	740	04_MSY	938
02_MTN	884	05_MYH	464
02_MEM	736	05_MKK	539
02_MFM	557	06_FTY	712
03_FMA	561	06_FWA	781
03_FTY	452		
		Total	9114

200 ms pause duration was adopted for our dialogs. The segmented utterances were orthographically transcribed into *kanji* (Chinese logograph) and *kana* (Japanese syllabary). Periods were used when there was a linguistic sentence-final expression in an utterance. The criteria of the sentence-final expression were defined with reference to the definition of the absolute clause boundary of the CSJ [29]. If pauses shorter than 400 ms occurred in an utterance, commas were used to show their position. Jargon and special terms for online games, e.g., BOT or STRAGE, and figures and counters were transcribed in *katakana* (angular Japanese syllabary) as these words were heard. The following three transcription tags were prepared for laughs, coughs, and so on.

- {laughs}, {coughs}
  Laughs, excluding utterances with laughing, and coughs.
- (?), (? (comment))
  An utterance that could not be transcribed due to noise or low sound volume.
- [comment:(comment)]
  Transcriber's comment.

As a result, the total number of utterances in our database was 9114. Table 1 shows the number of utterances for each speaker and Table 2 shows two sample dialogs (translated in English). In Table 1, the speakers are represented by the speaker IDs. For two speakers, 03_FMA and 02_MFM, 1,009 utterances were not adopted for the analysis due to their low sound levels. Moreover, 1,527 utterances with tags were also not used, because these utterances could not be transcribed and the acoustic features could not be calculated. As a result, the number of total utterances for the following analysis was 6578.

Table 2 Examples of dialogs. These sample dialogs were translated into English. The commas, periods and exclamation marks were attached for convenience in reading.

Dialog 1	
A:	Uhmm, the short cut key. OK, '3'.
	Wow, wh- what? What?
	Heeeeeeeeeey, le- let's follow 'em. Let's follow 'em.
B:	Uh?
B:	What? What's going on?
A:	{laughs} Hey, th- that on your right,
	on the right, on the right,
	wow wow wow wow...
A:	{laughs}
B:	Gee.
A:	What a terrible party this is!
B:	Wow. It takes you back, doesn't it!

Dialog 2	
B:	That was close!
B:	It's coming...
B:	Oh no! I'm gonna die {laughs}.
	I'm gonna die {laughs}.
A:	You're gonna die {laughs}.
A:	{laughs}.
B:	Phew! That was close!
B:	OK.
A:	I nearly lost my life there.

3. PSYCHOLOGICAL ASSESSMENT

3.1. Method

The utterances were labeled with emotional states according to their perceived emotional information. After labeling, the labeled utterances were rated for emotional intensity on the basis of how strong the emotion was perceived from the utterance. Both the labelers and the raters were instructed to judge each utterance according to its acoustic characteristics, not the content of the utterance.

3.1.1. Emotional labeling

Twenty-two labelers (14 males and 8 females) participated in emotion labeling. Each utterance was labeled by three labelers, since there were so many utterances to be judged that one labeler could not judge all of them. The labelers had to choose one emotional state to label each utterance from ten alternatives of eight emotional states: fear (FEA), surprise (SUR), sadness (SAD), disgust (DIS), anger (ANG), anticipation (ANT), joy (JOY), and acceptance (ACC), as well as a neutral state (NEU) which does not include any emotion, and states which are impossible to classify into the nine states above, or utterances with noise, etc. (OTH). The eight emotional states were selected with reference to the primary emotions in Plutchik's multidimensional model [30]. The definitions of the ten emotional states were presented to the labelers to give them a common understanding of each emotional state. The definitions were prepared with reference to a dictionary [31]. Each utterance was presented once in random order for each labeler to counterbalance order effects. Table 3 shows the ten emotional states, their abbreviations, and their definitions.

3.1.2. Emotional intensity rating

Each labeled utterance was rated for its emotional intensity by 18 raters (13 males and 5 females). Utterances

Table 3 Abbreviations and definitions of emotional states.

States	Abbr.	Definition
Fear	FEA	Feelings of avoiding people or things that are harmful
Sadness	SAD	Feelings of sorrow for irrevocable consequences such as misfortune or loss
Disgust	DIS	Feelings of avoiding unacceptable states or acts
Anger	ANG	Feelings of irritation or annoyance with an unforgiven subject
Surprise	SUR	Feelings of being disturbed, caught off balance, and confused after experiencing unexpected events
Anticipation	ANT	Feelings of longing for a desirable eventuality or a favorable opportunity
Joy	JOY	Feelings of gladness and thankfulness indicating intense satisfaction with something
Acceptance	ACC	Feelings of active involvement in something fascinating or positive
Neutral	NEU	No feelings at all
Others	OTH	Impossible to classify into the nine states above, or utterances with noise, etc.

Table 4 Results of emotional labeling. Percentages were calculated by dividing the number of utterances for each emotional state by the total number of utterances. The total number of utterances was 6,578.

States	Partial		Complete	
	Utterances	Percent	Utterances	Percent
FEA	142	2.2	33	0.5
SAD	243	3.7	49	0.7
DIS	335	5.1	45	0.7
ANG	237	3.6	60	0.9
SUR	565	8.6	177	2.7
ANT	427	6.5	69	1.0
JOY	595	9.0	174	2.6
ACC	303	4.6	27	0.4
NEU	798	12.1	116	1.8
OTH	200	3.0	30	0.5
Total	3,845	58.5	780	11.0

for witch at least two of the three labelers agreed to one emotional label were adopted for the ratings. The NEU and OTH utterances were excluded from the rating. The utterances were presented once in random order for every rater. The emotional state for each utterance was presented to the raters. The raters were instructed to rate its emotional intensity on a five-point scale from 1 (weak) to 5 (strong).

3.2. Results

Table 4 shows the numbers of utterances for the two types of inter-labeler agreements. The number of utterances was counted for each partially agreed emotional state where two out of the three labelers' choices were consistent (partial), and the number of utterances was counted for each completely agreed emotional state where all three labelers' choices were consistent (complete). The partial and complete agreements were 58.5% (chance level: 28%) and 11.0% (chance level: 1%), respectively. It was found in Table 4 that each positive emotional state, SUR, ANT, JOY, and ACC, shows a comparatively large number of utterances for both the partial agreement and the complete agreement than each negative emotional state, FEA, SAD, DIS, and ANG. These numbers indicate that the labelers could perceive positive emotion more than negative emotion from the utterances. Since the speakers did not tire of the task (MMORPG) and enjoyed the recording with their friends, positive emotions were frequently expressed in their voice.

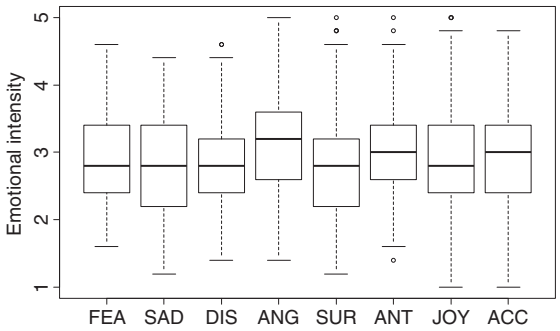

Fig. 2 Distribution of emotional intensities of utterances for each emotional state.

The mean correlation coefficient among the 18 raters was 0.24 (range = −0.01–0.52). The range of correlation coefficients among the 18 raters was widely spread, so that the criteria to rate emotional intensity were different among the raters. To obtain a more consistent emotional intensity rating, the mean rating score for each utterance was calculated with the ratings of the five best raters who showed the highest mean correlation coefficients with the other raters. The mean correlation coefficient among the five best raters was 0.42 (range = 0.36–0.52). All pairs of the five raters showed moderate correlations at a significant level of 0.05, as a result of correlation tests.

Figure 2 shows distributions of emotional intensities of utterances for each emotional state according to the result of emotional intensity rating. It indicates that the emotional intensity of each utterance spread from 1 to 5 for all eight emotional states, although, FEA and ANT have fewer utterances of weak emotion and SAD and DIS have fewer utterances of strong emotion. Our speech material contains

Table 5 Confusion matrix between partially agreed emotional states vs the remaining labeler's choice.

		Remaining labeler's choice									
		FEA	SAD	DIS	ANG	SUR	ANT	JOY	ACC	NEU	OTH
Partially agreed emotion	FEA	**23%**	15%	11%	2%	17%	6%	4%	1%	11%	10%
	SAD	13%	**20%**	14%	4%	5%	4%	5%	9%	15%	11%
	DIS	7%	12%	**13%**	19%	4%	3%	4%	9%	22%	6%
	ANG	3%	3%	29%	**25%**	9%	7%	5%	6%	11%	3%
	SUR	6%	3%	6%	7%	**31%**	11%	12%	6%	12%	7%
	ANT	3%	3%	6%	4%	5%	**16%**	22%	9%	23%	11%
	JOY	2%	5%	4%	3%	9%	18%	**29%**	13%	12%	5%
	ACC	3%	7%	11%	2%	3%	11%	12%	**9%**	35%	7%
	NEU	5%	7%	10%	5%	6%	14%	7%	22%	**15%**	9%
	OTH	6%	13%	6%	4%	15%	6%	6%	7%	24%	**15%**

various strengths of emotional intensity. The paradigm could enable us to collect utterances of not one level of emotional intensity, but wide-spread emotional intensity from weak to strong.

3.3. Discussion

3.3.1. Inter-labeler agreement

We obtained very good results for the partial and complete agreements. The partial and complete agreements were 58.5% and 11.0%, respectively, which are much higher than the chance levels of the partial and complete agreements (28% and 1%, respectively). The result suggests that the labelers could judge each utterance by using common cues of emotional speech.

However, since around 3,000 utterances did not agree with one emotional state, our speech material contains complex expressions displaying multiple emotional states. To investigate which emotional state is confused with another, the rates of the remaining labeler's choice against the partially agreed emotional states are calculated (Table 5). In Table 5, the rates of consistent emotional states with the partially agreed emotional states in the diagonal components are in bold face. The cells that show at least 10% are shaded in gray. It was found that the nonemotional state, NEU, is often chosen as inconsistent emotional states by the remaining labelers. This result suggests that these utterances were considered to contain a weak emotion. This topic is discussed in Sect. 3.3.2.

It was also found that the utterances labeled with one of the negative emotional states, FEA, SAD, DIS, or ANG, were often labeled with one of the negative emotional states by the remaining labelers. For example, the SAD utterances were often labeled with FEA or DIS by the remaining labelers. Similarly, the utterances labeled with one of the positive emotional states, JOY, ACC, or ANT, were also often labeled with one of the positive emotional states. SUR, which is often experienced with other positive and negative emotions, was often labeled with a positive emotional state. This suggests that the judgment of the higher taxon (positive or negative) for each utterance was consistent among labelers. To calculate the inter-labeler agreement for the higher taxon, the emotional states were grouped into three emotional categories of positive (JOY+ACC+ANT+SUR), negative (FEA+SAD+DIS+ANG), and other (NEU+OTH). The result shows high agreement of 89.2%. This implies that the emotional speech collected by the proposed paradigm was perceptually distinguishable for the listeners.

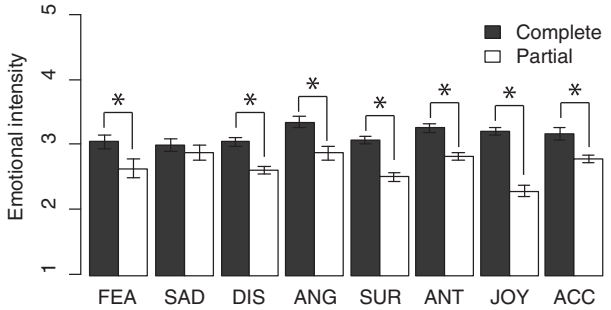

Fig. 3 Mean emotional intensity and standard error of each partially agreed and completely agreed emotional state. The partially agreed bars with asterisks represent significantly weaker mean intensity than the completely agreed bars ($p < 0.05$).

3.3.2. Utterances with weak emotion [32]*

Table 5 indicates that NEU was often chosen by the remaining labelers, showing that its rates against every partially agreed emotional states were more than 10%. These utterances were considered to contain weak emotion because NEU is the state containing no emotion. Figure 3 shows the mean emotional intensities of the utterances with completely agreed emotional states and with partially agreed emotional states. The utterances with partially agreed emotional states only include the utterances labeled with consistent emotional states or NEU by the remaining labeler. The mean emotional intensity of all partially

*The analysis in this section is a reproduction of that in [32]

agreed emotional states, except SAD, was significantly weaker than the mean emotional intensity of the completely agreed emotional states ($p < 0.05$). It was proved that utterances labeled with NEU by the remaining labeler contain weak emotion.

3.3.3. Expressiveness

To assess the efficiency of our paradigm to collect naturalistic emotional speech, the number of labeling instances of our speech material was compared with those of the two other speech materials adopted in earlier studies. One of these is spontaneous pseudodialogs for anger speech classification [10], and the other is a speech database for paralinguistic information studies, UUDB [18]. The rate of instances was calculated by dividing the number of emotion labels by the total number of labels, in accordance with [10]. Note that the labeling scheme for each of the three materials is not completely the same and the calculation was done for the sake of comparison among them. Each utterance of the anger speech material [10] was labeled with one of 7 emotional states: neutral, annoyed, frustrated, tired, amused, other, or notapplicable (contained no speech data from the user). The annoyed, frustrated, tired, amused, and other utterances were regarded as emotional utterances for comparison. No utterance of the UUDB was labeled with a single emotional state. It was rated on a seven-point scale of six paralinguistic information values: pleasant–unpleasant, aroused–sleepy, dominant–submissive, credible–doubtful, interested–indifferent, and positive–negative. The utterances had all six paralinguistic information values; hence, the nonemotional state was never judged. To compare the rates of labeling instances between our speech material and the UUDB, the utterances rated with scores from 3 to 5 (weak or none) for every 6 values were regarded as nonemotional utterances, and the rest were regarded as emotional utterances.

Table 6 shows the number and rate of labeling instances of each emotional state for our speech material. The total number of labeling instances for our speech material was 19,734 labels (6,578 utterances × three labelers). Among them, 14,414 labels were the emotional labels of eight types of emotional state, so that a very high rate, 73.0%, of the total labeling instances were emotional labels. The total number of labeling instances for the speech material of Ang *et al.* [10] were 49,553, which were judged by 2.62 mean labelers per utterance, and included 4,904 emotional labels. The rate of the emotional labels was quite low, 9.9%. The UUDB had 14,520 labels judged by three labelers. 58.2% of the total labels (8,446 labels) were emotional labels. Figure 4 shows the frequency of the emotional labels of each speech material. A proportional test was conducted with the number of labels for each speech material. A significant difference existed among the three speech materials ($\chi^2(2) = 27659.87$, $p < 0.001$), and our speech materials had a significantly higher rate than did the other two speech materials ($p < 0.01$, indicated with asterisks in Fig. 4). This result implies that the proposed paradigm could yield a lot of emotional speeches that were perceptually distinguishable by the listeners.

Table 6 Number of labeling instances of each emotional state for our speech material. Percentages were calculated by dividing the number of labelings for each emotional state by the total number of labeling instances.

States	Sum. of labels	
	Instances	Percent
FEA	967	4.9
SAD	1,330	6.7
DIS	1,912	9.7
ANG	1,209	6.1
SUR	2,194	11.1
ANT	2,215	11.2
JOY	2,469	12.5
ACC	2,118	10.7
NEU	3,797	19.2
OTH	1,523	7.7
Total	19,734	100

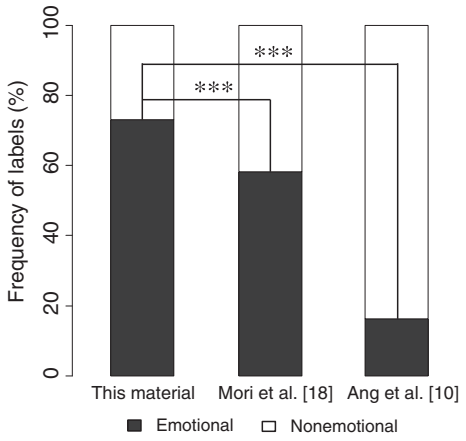

Fig. 4 Frequency of emotional labels.

4. ACOUSTICAL ASSESSMENT

4.1. Method

For the acoustical assessment of the proposed collection paradigm, the recorded speech material was tested to determine whether they showed an acoustical difference between emotional utterances and nonemotional utterances. The utterances with completely agreed emotional states were adopted for the acoustic analysis. Nine acoustic parameters were extracted from these utterances. Our parameters included prosodic and segmental parameters, pitch, duration and speaking rate, power, and voice quality,

which were traditionally adopted in emotional speech studies. A one-way analysis of variance (ANOVA) was conducted with the factor of the nine emotional states (JOY, ANT, FEA, SUR, DIS, ANG, ACC, SAD, and NEU) for each acoustic parameter. The OTH utterances were not adopted for the analysis because these contain noise or have a low sound volume, causing difficulty in extracting the acoustic parameters. To compare the difference in the mean value of each acoustic parameter between the emotional utterances and the nonemotional utterances, Dunnett-type multiple comparison tests were also conducted for each acoustic parameter. The control state was the NEU utterances.

4.2. Parameter Extraction

The nine acoustic parameters were prepared for our analysis with reference to [13]. The sampling frequency of each utterance was down-sampled at 16 kHz from 48 kHz.

4.2.1. Fundamental frequency

The three pitch parameters, F_{mini}, F_{maxi}, and F_{stdv}, were calculated with fundamental frequency values within an utterance. Every parameter was represented in descriptive statistics within an utterance: minimum value (F_{mini}), maximum value (F_{maxi}), and the standard deviation (F_{stdv}) of the logarithmic F_0 value. F_0 values were automatically extracted at intervals of 10 ms with the speech manipulation system, STRAIGHT [33]. If a smooth F_0 contour was not produced within an utterance, F_0 values of the utterances were extracted again after the extraction settings of STRAIGHT were adjusted. The upper and lower 10% points of the F_0 values within an utterance were regarded as our F_{mini} and F_{maxi} parameters, because F_0 extraction seriously affects the minimum and maximum values on our parameters. F_{mini} and F_{maxi} were the speaker-normalized values of these statistics. Our normalization technique was a simple calculation in which the mean F_0 value of all utterances of a speaker was deducted from each F_0 value of the speaker to set the mean F_0 values for each speaker to 0.

4.2.2. Power

The two parameters, P_{maxi} and P_{magn}, were calculated as the values of short-term power within an utterance. The value of short-term power was calculated for a 20 ms window length at 5 ms intervals according to

$$P(m) = 10\log\left(\frac{1}{N}\sum_{n=0}^{N-1} S^2(Nm+n)\right), \qquad (1)$$

where $S(i)$ is the amplitude of the i-th sample, and m is the frame number for the analysis ($m = 0, 1, 2, \ldots$).

P_{maxi} was the maximum short-term power within an utterance. P_{magn} was the magnitude of the short-term power shifts within an utterance. P_{magn} quantified the speed of the short-term power shifts by calculating the root-mean-square slopes of linear regression lines over eleven frames at every frame within an utterance. The eleven frames consist of the five frames before and after the center frame.

4.2.3. Voice quality

The voice-quality parameters, C_{mean} and C_{stdv}, were calculated with the first mel-cepstral coefficients of voiced frames within an utterance. Cepstrum is commonly used in speech technology applications because of its ability to capture the formant structure and spectral tilt, and because the low-order cepstral coefficients are sensitive to overall spectral slope [34]. Moreover, the first cepstral coefficient corresponds to the overall spectral tilt [35–37]. Spectral tilt is often analyzed for vocal expression of emotion as an index of harshness or softness [38], or an index of breathiness [39], e.g. a flatter slope with positive emotion that sounds less breathy and less coarse. For each parameter, the mean value (C_{mean}) and the standard deviation (C_{stdv}) were represented in descriptive statistics within an utterance. The first mel-cepstral coefficient of voiced frames was extracted with the Speech Processing Toolbox, VoiceBox, for MATLAB [40]. The voiced frames were defined with reference to the voiced/unvoiced information extracted by STRAIGHT.

4.2.4. Duration/speaking rate

The two duration/speaking rate parameters, D_{mora} and D_{rate}, were calculated from the mora numbers and duration within an utterance. D_{mora} was the mean mora number within a breath group (unit; mora). The breath group was defined as one or more segments within an utterance separated by commas in the transcription. A comma represents a pause shorter than 400 ms within an utterance. The mora numbers within an utterance were counted in the transcription, after *kanji* (Chinese logograph) and *kana* (Japanese syllabary) were converted to *katakana* (angular Japanese syllabary) with the morphological analyzer MeCab [41] which was trained with the electronic Japanese dictionary, UniDic [42]. The conversion errors to Katakana were manually corrected. D_{rate} was the mean speaking rate of an utterance (unit; mora/s).

4.3. Results

The one-way ANOVA revealed significant main effects of the nine emotional states for each of the acoustic parameters ($p < 0.05$). The more precise results are shown in Table 7. The numbers in Table 7 show the mean values of each acoustic parameter for each emotional state. The numbers in bold face indicate the mean values of the emotional states that show a significant difference from the control state of NEU, according to the results of the Dunnet-type multiple comparison test ($p < 0.05$). According to Table 7, all emotional states except ACC show a significant difference for any of the nine acoustic parameters from the nonemotional state of NEU.

Table 7 Mean values of each acoustic feature of the emotional states.

Acoustic feature	NEU	FEA	SAD	DIS	ANG	SUR	JOY	ANT	ACC
F_{mini}	−0.33	**−0.06**	**−0.22**	−0.32	−0.29	**−0.06**	**−0.19**	−0.29	−0.34
F_{maxi}	0.21	0.25	0.17	0.11	**0.37**	**0.59**	0.29	0.23	0.2
F_{stdv}	0.16	**0.09**	**0.1**	**0.11**	0.19	**0.2**	0.13	0.14	0.15
P_{maxi}	−17.37	−18.02	**−23.37**	−18.09	**−11.16**	**−14.1**	**−13.52**	−16.57	−16.79
P_{magn}	291.53	259.86	**235.8**	**246.02**	293.88	**370**	**310.99**	295.92	271.45
C_{mean}	4.54	**3.49**	**3.71**	4.68	3.9	**3.26**	**3.13**	**3.54**	4.25
C_{stdv}	1.98	1.78	1.82	**1.58**	2.16	1.83	**2.21**	2.01	1.71
D_{mora}	7.59	6.82	6.17	**8.25**	9.74	3.35	8.5	**9.42**	6.2
D_{rate}	8.35	8.84	7.24	6.96	8.74	**7.7**	8.38	9.83	9.13

4.4. Discussion

The results of the one-way ANOVA suggest that the utterances of all emotional states except ACC could be distinguished from the utterances of the nonemotional state NEU by the acoustic characteristics. This implies that the recorded emotional speech contains rich acoustic expressions. The seven parameters show significant differences between the SUR utterances and the NEU utterances. The SUR utterances were recognized by listeners on the basis of multiple cues of pitch, power, voice quality, and speaking rate. The ANT utterances were distinguished from the NEU utterances by the two cues of voice quality and duration.

However, the ACC utterances could not be distinguished from the NEU utterances on the basis of any of the nine acoustic parameters. This result indicates that the acoustic characteristics for ACC and NEU are quite similar. Since the labelers were able to distinguish the ACC utterances from the NEU utterances, nonacoustical cues, such as linguistic information, might help them to judge ACC utterances. To investigate the influence of linguistic information on emotion judgment, the number of words that form emotional utterances and its perplexity were calculated for each emotion. The result shows that the number of words for 27 ACC utterances is 61 and its perplexity is 46.29. This suggests that the ACC utterances consisted of a variety of words so that the labelers did not judge ACC utterances on the basis of specific words. As for the utterances of the other emotional states, there was not a strong influence of linguistic information on emotion judgment because its perplexity is not small compared with the number of words for each emotional state (see Appendix for further discussion on the number of words and the perplexity for each emotional state and the calculation of perplexity). Upon a comparison of the transcription of the ACC utterances and the NEU utterances, a difference is found between them. The ACC utterances were often associated with repetition, such as "yes, yes, yes, yes" ("Hai, hai, hai, hai" in Japanese). The number of ACC utterances with repetition was 8 out of 27 utterances, or approximately 30% of the ACC utterances. However, the number of NEU utterances with repetition was only 5 out of 116, or 4% of the NEU utterances. ACC utterances could be distinguished from NEU utterances by their linguistic characteristic of repetition.

5. CONCLUSION

For the purpose of constructing a naturalistic emotional speech database, we proposed the novel paradigm of collecting naturalistic emotional speech during spontaneous Japanese dialog in a laboratory setting. To prove the validity of the paradigm, the collected speech material was assessed concerning whether it conveys rich emotions psychologically and acoustically. The results of emotional labeling for the collected speech supported the validity of the paradigm, showing higher inter-labeler agreement than chance levels. In addition, it was revealed that the paradigm is superior in the quantity of emotional speech compared with other paradigm, as it shows a significantly higher rate of labeling instances for our speech material (73%, $\chi^2(2) = 27659.87$, $p < 0.001$) than did the two other emotional speech materials adopted in earlier studies. The acoustical analysis also supported the validity of our paradigm by showing that utterances of seven out of eight emotions show a significant difference from the nonemotional utterances. The results of this study proved that the proposed paradigm is effective for collecting naturalistic, vivid vocal emotional expressions. The authors believe that speech technology can be advanced using naturalistic emotional speech material collected using the proposed paradigm.

REFERENCES

[1] C. E. Williams, "Emotions and speech: Some acoustical correlates," *J. Acoust. Soc. Am.*, **52**, 1238–1250 (1972).

[2] K. Itoh, "A basic study on voice sound involving emotion. III. Non-stationary analysis of single vowel [e]," *Jpn. J. Ergon.*, **22**, 211–217 (1986) (in Japanese).

[3] Y. Kitahara, "Prosodic components of speech in the expression of emotions," *J. Acoust. Soc. Am.*, **84**, S98–S99 (1988).

[4] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *J. Pers. Soc. Psychol.*, **70**, 614–636 (1996).

[5] I. S. Engberg, A. V. Hansen, O. Andersen and P. Dalsgaard,

"Design, recording and verification of a Danish emotional speech database," *Proc. EUROSPEECH '97*, pp. 1695–1698 (1997).
[6] R. Cowie, "Perceiving emotion: Towards a realistic understanding of the task," *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **364**, 3515–3525 (2009).
[7] E. Douglas-Cowie, "Emotional speech: Towards a new generation of databases," *Speech Commun.*, **40**, 33–60 (2003).
[8] A. Lee, T. Kawahara and K. Shikano, "Julius — An open source real-time large vocabulary recognition engine," *Proc. EUROSPEECH 2001*, pp. 1691–1694 (2001).
[9] T. Athanaselis, S. Bakamidis, I. Dologlou, R. Cowie, E. Douglas-Cowie and C. Cox, "ASR for emotional speech: Clarifying the issues and enhancing performance," *Neural Networks*, **18**, 437–44 (2005).
[10] J. Ang, R. Dhillon, A. Krupski, E. Shriberg and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," *Proc. ICSLP 2002*, pp. 2037–2040 (2002).
[11] L. Devillers and L. Vidrascu, "Real-life emotion detection with lexical and paralinguistic cues on human-human call center dialogs," *Proc. Interspeech 2006*, pp. 801–804 (2006).
[12] D. Litman and K. Forbesriley, "Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogs with both human and computer tutors," *Speech Commun.*, **48**, 559–590 (2006).
[13] Y. Arimoto, S. Ohno and H. Iida, "An estimation method of degree of speaker's anger emotion with acoustic and linguistic features," *J. Nat. Lang. Process.*, **14**, 147–163 (2007) (in Japanese).
[14] Y. Arimoto, H. Kawatsu, S. Ohno and H. Iida, "Emotion recognition in spontaneous emotional speech for anonymity-protected voice chat systems," *Proc. Interspeech 2008*, pp. 322–325 (2008).
[15] H. Kawatsu, Y. Arimoto and S. Ohno, "Development of anonymous voice chat system conveying emotional information," *Proc. IEICE Gen. Conf. '08*, D-14-3, p. 176 (2008) (in Japanese).
[16] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson and L. Kessous, "Whodunnit — Searching for the most important feature types signaling emotion-related user states in speech," *Comput. Speech Lang.*, **25**, 4–28 (2011).
[17] N. Campbell, "Speech & expression; The value of a longitudinal corpus the JST ESP Corpus," *Proc. LREC 2004* (2004).
[18] H. Mori, T. Satake, M. Nakamura and H. Kasuya, "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," *Speech Commun.*, **53**, 36–50 (2011).
[19] Z. Zeng, M. Pantic, G. I. Roisman and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, **31**, 39–58 (2009).
[20] C. A. Anderson and B. J. Bushman, "Effects of violent video games on aggressive behavior, aggressive cognition, aggressive affect, physiological arousal, and prosocial behavior: A meta-analytic review of the scientific literature," *Psychol. Sci.*, **12**, 353–359 (2001).
[21] M. van't Wout, R. S. Kahn, A. G. Sanfey and A. Aleman, "Affective state and decision-making in the ultimatum game," *Exp. Brain Res.*, **169**, 564–568 (2006).
[22] N. Ravaja, M. Turpeinen, T. Saari, S. Puttonen and L. Keltikangas-Järvinen, "The psychophysiology of James Bond: Phasic emotional responses to violent video game events," *Emotion*, **8**, 114–120 (2008).
[23] R. L. Hazlett, "Measuring emotional valence during interactive experiences," *Proc. SIGCHI Conf. Human Factors in Computing Systems — CHI '06*, pp. 1023–1026 (2006).
[24] R. Hazlett and J. Benedek, "Measuring emotional valence to understand the user's experience of software," *Int. J. Hum. Comput. Stud.*, **65**, 306–314 (2007).
[25] T. Tijs, D. Brokken and W. Ijsselsteijn, "Creating an emotionally adaptive game," *Proc. ICEC 2008*, Vol. 5309 of LNCS 5309, S. M. Stevens and S. Saldamarco, Eds. (Springer-Verlag, Berlin, 2008), pp. 122–133.
[26] *Skype*, http://www.skype.com/.
[27] *Tapur*, http://www.tapur.com/.
[28] H. Koiso, K. Nishikawa and Y. Mabuchi, "Chap. 2 Transcription," in *Construction of the Corpus of Spontaneous Japanese* (National Institute of Japanese Language and Linguistics, 2006), pp. 23–132 (in Japanese).
[29] T. Maruyama, K. Takanashi and K. Uchimoto, "Chap. 5 Clause Unit," in *Construction of the Corpus of Spontaneous Japanese* (National Institute of Japanese Language and Linguistics, 2006), pp. 255–322 (in Japanese).
[30] R. Plutchik, *Emotions: A Psychoevolutionary Synthesis* (Harper & Row, New York, 1980).
[31] T. Yamada, T. Shibata, Y. Kuramochi and A. Yamada, *Shin-Meikai Japanese Dictionary*, 6th ed. (Sanseido, Tokyo, 2005) (in Japanese).
[32] Y. Arimoto, S. Ohno and H. Iida, "Assessment of spontaneous emotional speech database toward emotion recognition: Intensity and similarity of perceived emotion from spontaneously expressed emotional speech," *Acoust. Sci. & Tech.*, **32**, 26–29 (2011).
[33] H. Kawahara, A. de Cheveigne, H. Banno, T. Takahashi and T. Irino, "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT," *Proc. Interspeech 2005*, pp. 537–540 (2005).
[34] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition* (Prentice Hall, Englewood Cliffs, NJ, 1993).
[35] R. Nakatsu, H. Nagashima, J. Kojima and N. Ishii, "A speech recognition method for telephone voice," *IEICE Trans.*, **J66-D**, 377–384 (1983) (in Japanese).
[36] Y. Tohkura, "A weighted cepstral distance measure for speech recognition," *IEEE Trans. Acoust. Speech Signal Process.*, **35**, 1414–1422 (1987).
[37] J. Busset, Y. Laprie, L. C. Umr and J. Botanique, "Adaptation of cepstral coefficients for acoustic-to-articulatory inversion," *Proc. ISSP '11* (2011).
[38] M. Schröder, "Speech and emotion research: An overview of research frameworks and a dimensional approach to emotional speech synthesis," Doctoral thesis, *Phonus 7, Res. Rep. Inst. Phonetics, Saarland Univ.* (2003).
[39] C. T. Ishi, H. Ishiguro and N. Hagita, "Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality," *Speech Commun.*, **50**, 531–543 (2008).
[40] M. Brookes, *VOICEBOX: Speech Processing Toolbox for MATLAB*, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voice/, (1999).
[41] T. Kudo, K. Yamamoto and Y. Matsumoto, "Applying conditional random fields to Japanese morphological analysis," *Proc. EMNLP 2011*, pp. 230–237 (2004).
[42] Y. Den, T. Ogiso, H. Ogura, A. Yamada, N. Minematsu, K. Uchimoto and H. Koiso, "The development of an electronic dictionary for morphological analysis and its application to Japanese corpus linguistics," *Jpn. Linguist.*, **22**, 101–123 (2007) (in Japanese).

APPENDIX

To assess the influence of linguistic information on emotional judgment, the number of words and perplexities for each emotional state were calculated. Each utterance was split into separate morphemes by MeCab [41] and Unidic [42] to count morphemes as the number of words. The perplexity, PP, for each emotional state was calculated according to

$$H(p) = -\sum_i p_i \log_2 p_i,$$

$$\mathrm{PP} = 2^{H(p)},$$

where $H(p)$ is the entropy of the words, and p_i is the probability of $word_i$ for each emotional state. Table A.1 shows the number of words (Word) and perplexity (PP) for each of the partially agreed (Partial) and completely agreed (Complete) emotional states.

For the assessment, large perplexity for each emotional state would be expected to show high complexity of the database since the purpose of the assessment is to reveal that there is no influence of the use of a specific word on emotional judgment. All emotional states show fairy high perplexities for both the partially agreement and completely agreement utterances. In particular, the completely agreed emotional states show greater perplexities than approximately half their number of words. This implies that there is not a strong influence of linguistic information on emotion judgment and that our collected emotional speech consisted of a variety of words, not a specific word.

Table A.1 Number of words and perplexity for each emotional state.

States	Partial		Complete	
	Word	PP	Word	PP
FEA	248	144.71	72	51.02
SAD	362	189.98	110	87.66
DIS	547	251.34	140	114.29
ANG	470	228.40	180	125.23
SUR	414	182.62	172	104.86
ANT	781	314.21	220	152.62
JOY	984	348.46	401	196.94
ACC	388	157.74	61	46.29
NEU	1,028	355.57	297	200.59
OTH	246	169.14	50	49.00

Yoshiko Arimoto received the B.A. degree in literature from Aoyama Gakuin University in 1996, and the B.S. degree in media science and the M.S. degree in engineering from Tokyo University of Technology in 2004 and 2007, respectively. Since 2007, she has been a Ph.D. candidate in Tokyo University of Technology. She is currently a researcher in JST, ERATO, Okanoya Emotional Information Project. Her research interests include automatic emotion recognition, emotion perception, and speech processing. She is a member of ASJ, NLP and ISCA.

Hiromi Kawatsu received the B.E. and Ph.D. degrees in engineering from Tokyo University of Technology, Japan, in 2002 and 2008, respectively. She was a JSPS Research Fellow (DC2) in 2007–2008, and a post doctoral researcher at National Institute for Japanese Language and Linguistics in 2008. In 2009, she joined IBM Japan, Ltd. and is currently a development engineer. Her research interests include emotional speech synthesis and automatic emotion recognition from speech.

Sumio Ohno received the B.E. degree in electrical engineering in 1988, and the M.E. and Dr.Eng. degrees in electronic engineering respectively in 1990 and 1993 from the University of Tokyo. From April 1993 to March 1999 he was a faculty member of the Department of Applied Electronics, Science University of Tokyo. Since 1999 he has been with the Department of Information Networks, Tokyo University of Technology, where he is presently a Professor at the School of Computer Science. He has been engaged in studies of speech perception, automatic speech recognition, and prosody. He is a member of the Institute of Electronics, Information and Communication Engineers, and the Acoustical Society of Japan.

Hitoshi Iida received the Ph.D. degree in engineering from Tokyo Institute of Technology. He was with the Basic Research Department in Nippon Telegraph and Telephone Public Corporation, the Interpreting Telecommunications Research Laboratories and Spoken Language Translation Research Laboratories in Advanced Telecommunications Research Institute International (ATR), and Sony CSL. He has been a Professor at the School of Media Science, Tokyo University of Technology since 2003, and the Dean of the School since 2009. He received the Director-General's award for Academic Merit from the Science and Technology Agency, International Association for Machine Translation (IAMT) Award of Honor, and others. He has been engaged in studies of machine translation, natural language processing, spoken language processing, artificial intelligent, multimodal interaction, and Kansei engineering.