Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics☆

Hiroki Mori[a,*], Tomoyuki Satake[a], Makoto Nakamura[b], Hideki Kasuya[a]

[a]*Graduate School of Engineering, Utsunomiya University, 7-1-2, Yoto, Utsunomiya-shi, 321-8585 Japan*
[b]*Faculty of International Studies, Utsunomiya University, 350, Minemachi, Utsunomiya-shi, 321-8505 Japan*

Abstract

The Utsunomiya University (UU) Spoken Dialogue Database for Paralinguistic Information Studies is introduced. The UU Database is especially intended for use in understanding the usage, structure and effect of paralinguistic information in expressive Japanese conversational speech. Paralinguistic information refers to meaningful information, such as emotion or attitude, delivered along with linguistic messages. The UU Database comes with labels of perceived emotional states for all utterances. The emotional states were annotated with six abstract dimensions: pleasant-unpleasant, aroused-sleepy, dominant-submissive, credible-doubtful, interested-indifferent, and positive-negative. To stimulate expressively-rich and vivid conversation, the "4-frame cartoon sorting task" was devised. In this task, four cards each containing one frame extracted from a cartoon are shuffled, and each participant with two cards out of the four then has to estimate the original order. The effectiveness of the method was supported by a broad distribution of subjective emotional state ratings. Preliminary annotation experiments by a large number of annotators confirmed that most annotators could provide fairly consistent ratings for a repeated identical stimulus, and the inter-rater agreement was good ($W \simeq 0.5$) for three of the six dimensions. Based on the results, three annotators were selected for labeling all 4840 utterances. The high degree of agreement was verified using such measures as Kendall's W. The results of correlation analyses showed that not only prosodic parameters such as intensity and f_0 but also a voice quality parameter were related to the dimensions. Multiple correlation of above 0.7 and RMS error of about 0.6 were obtained for the recognition of some dimensions using linear combinations of the speech parameters. Overall, the perceived emotional states of speakers can be accurately estimated from the speech parameters in most cases.

Key words:
Emotional state, Expressive speech, Annotation, Abstract dimensions, Spontaneous speech, Spoken dialogue

1. Introduction

Paralinguistic information generally refers to information that is not linguistic content itself but some meaningful information delivered along with linguistic messages. Sometimes paralinguistic information is even more eloquent than the linguistic message itself. Research on paralinguistic information is attracting growing attention among the speech science community, partly because revealing the nature of paralinguistic information involves not only advanced speech technologies such as human-like agents, but also revealing the hidden nature of speech communication between humans.

Emotion and expressivity in speech is a central topic of paralinguistic information (Cowie et al., 2001; Erickson, 2005). However, paralinguistic information should not be viewed only within the context of biological effects of individuals as in the traditional psychology of emotion, because speech has an important function: interaction. If the aim of a study on paralinguistic information includes not only its emotional aspects but also its socio-linguistic roles, scripted corpus collection is almost useless. Today, there is a trend toward considering natural emotion (Douglas-Cowie et al., 2003; Aubergé et al., 2003) in corpus development for expressive speech studies. Recent works on developing a spontaneous speech corpus for paralinguistic information studies include Greasley et al. (1995); Douglas-Cowie et al. (2000); Campbell (2003); Devillers and Vidrascu (2006); Truong et al. (2008), and Arimoto et al. (2008).

However, there is a serious problem in all attempts to design a speech corpus with natural emotion, i.e. how to label emotions in the corpus? Unlike traditional emotion studies, well-established emotion categories do not apply to most utterances in daily conversations. An alternative scheme is therefore needed for describing the emotional states that are expressed in speech (Cowie and Cornelius, 2003). Possible approaches include lists of key emotions, dimensional representations of underlying emotion, and physiological measures such as heart rate, eye blink, EEG, or facial muscle activities. Among these, effect-oriented emotion labels, which describe the listener's impression, are indispensable if the corpus is to be used for building speech applications such as expressive speech synthesis, or exploring the interactive effects of paralinguistic information in speech communication. Irrespective of the application, estab-

*Corresponding author. Tel:+81 28 689 6120; fax +81 28 689 6119.
Email addresses: hiroki@speech-lab.org (Hiroki Mori)

lishing a common ground for labeling speech expressivity is certainly a major challenge. One of the objectives of our corpus is to provide a reference implementation for future corpus development.

The Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UU Database) is the first public speech corpus specifically designed for studies on paralinguistic information in expressive Japanese dialogue speech. The UU Database is characterized by its unique task design and emotional state labels. Other existing corpora of natural dialogue speech (e.g. Reading/Leeds Emotion in Speech Corpus (Greasley et al., 1995), the Belfast Naturalistic Emotional Database (Douglas-Cowie et al., 2000), and the JST/CREST ESP Corpus (Campbell, 2003)) are not available for public use. Dialogues provided by some Japanese speech corpora (e.g. PASD, the Japanese Map Task Corpus (Horiuchi et al., 1999), and ATR-SLDB (Morimoto et al., 1994)) are too restrictive, or businesslike, for studies of speech expressivity; in fact, there is no public Japanese spoken dialogue corpus that provides speech with natural emotion. The list of relevant corpora is shown in Table 1. Despite the variation in terminology (evaluation = valence = pleasantness, activation = arousal), the table shows that the dimensional description has been adopted as a de facto standard for annotating natural emotion in recent studies. Only the Vera am Mittag Database and the UU Database are public corpora provided with dimension-based emotion annotation. Because most emotions appearing in the Vera am Mittag Database were evaluated as unpleasant (Grimm et al., 2008), the UU Database is distinctive in that it provides speech with a wide variety of emotions, as described in a later section.

The goal of the UU Database project is to provide not only an original collection of expressive dialogue speech but also empirical knowledge about the task design, the annotation scheme and its reliability, and acoustic correlates of expressive speech, as a new resource for emotion and speech scientists.

This paper describes the outline of design, building, and evaluation of the UU Database. Section 2 explains the dialogue task, namely the "4-frame cartoon sorting task," which was designed to collect natural dialogue speech with a wide spectrum of expressivity. Section 3 details the specifications of the recorded speech. Section 4 describes the framework and psychological background of emotional state annotation. Section 5 discusses the consistency and agreement of the emotional state annotation based on this framework. Section 6 shows how "good annotators" were selected, how well the annotations agreed, and how widely the emotional states of all the utterances in the UU Database are distributed. The validity and effectiveness of the emotional state labels with respect to acoustic parameters are evaluated in two ways. In Section 7, various speech parameters are analyzed as correlates of perceived emotional states. In Section 8, the linear regression modeling of emotional states based on these speech parameters is described. Section 9 concludes the paper.

2. Task Design

To collect speech samples of really natural and spontaneous dialogue, extensive recording of daily conversation (the "Pirelli-calendar" approach (Campbell, 2003)) might be a more suitable way than laboratory recording of task-oriented dialogue. Nevertheless, in recording dialogues it is usual to assume a task because of its efficiency. Careful choice of the task for recording dialogues is therefore important to enable the database to be used for investigating rich emotional expression or its pragmatic use in speech communication.

In the UU Database project, the objectives of task design were the following:

- To stimulate expressively-rich and vivid conversation
- To involve opinion exchanges, negotiations and persuasions, and to stimulate the active participation of both speakers
- To enable the subjects to participate with genuine interest, to improve motivation

Conventional tasks for dialogue research, such as the map task (Anderson et al., 1991; Horiuchi et al., 1999), meet some of these criteria. However, neither emotion nor expressivity was a major concern in the design of the map task. Similarly, other existing tasks are considered to be insufficiently interesting to the subjects.

We therefore devised a "4-frame cartoon sorting task." In this task, four cards each containing one frame extracted from a cartoon are shuffled, and each participant has two cards out of the four, and is asked to estimate the original order without looking at the remaining cards. Appendix A shows a typical dialogue that might occur for the 4-frame cartoon sorting task.

The task meets the above objectives because it involves such processes as information elicitation and explanation, as well as discussion to reach an agreement. The variety of expressiveness was also confirmed by a statistical investigation of emotional state annotation described in Section 6.

The usefulness of the 4-frame cartoon sorting task was revealed through the process of building the UU Database. Its main advantages include:

- It is relatively easy to prepare materials for the task.
- The difficulty of the task can be controlled to some extent by the complexity of the original cartoon.
- In the task, the participants need to explain not only what the characters in the cartoon are saying but also describe the situation. This makes the utterances more spontaneous.
- The goal is easy to understand because the story of a 4-frame cartoon should be relatively clear.
- The task strongly motivates the participants because cartoons are familiar and generally popular. (Indeed, all the participants wanted to know the correct order of the original cartoons even after their sessions.)

Table 1: Existing speech corpora for studying spontaneous dialogue.

Corpus	Language	Data collection method	Model of emotion annotation	Public availability
HCRC Map Task Corpus (Anderson et al. 1991)	English	Map task	Nothing	Available
Japanese Map Task Corpus (Horiuchi et al. 1999)	Japanese	Map task	Nothing	Freely available
PASD Simulated Spoken Dialogue Corpus	Japanese	Simulated dialogue under various tasks	Nothing	Freely Available
ATR-SLDB (Morimoto et al. 1994)	Japanese	Simulated dialogue under the hotel reservation task	Nothing	Available
CallHome (LDC)	English, Arabic, German, Spanish, Japanese, Chinese	Telephone call to family members or close friends	Nothing	Available
Reading/Leeds Emotion in Speech Corpus (Greasley et al. 1995)	English	Interviews on radio/TV programs	Category (happiness, sadness, anger, fear, disgust), freely choiced label	Not available
Belfast Naturalistic Database (Douglas-Cowie et al. 2000)	English	Inteviews on TV chat shows	Dimension (activation-evaluation, rated using Feeltrace), Category (40 words)	Not available
JST/CREST ESP corpus (Campbell 2003)	Japanese	Daily recording of volunteers' natural spoken interactions	Dimension (activation-evaluation, rated using Feeltrace)	Not available
The Vera am Mittag German Audio-Visual Spontaneous Speech Database (Grimm et al. 2008)	German	Spontaneous and emotional speech recorded from a TV talk show	Dimension (valence, activation, and dominance, using the self assessment manikins)	Freely available
TNO-Gaming corpus (Truong et al. 2008)	Dutch	Talks with friends during playing multiplayer video game with preset events	Dimension (arousal, valence), Category (12 words) Self and Observer ratings	Not available
UU Database (this paper)	Japanese	Four-frame cartoon sorting task	Dimension (pleasantness, arousal, dominance, credibility, interest, and positivity)	Freely available

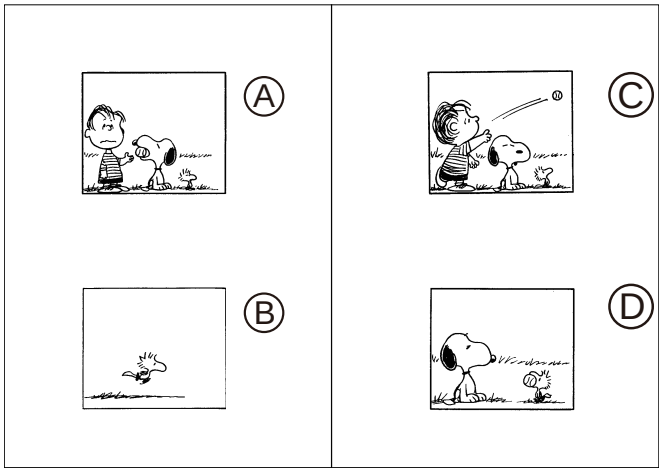

Figure 1: An example of 4-frame cartoon sorting task. (c) 2010 Peanuts Worldwide LLC., dist by UFS, Inc.

Table 2: Attributes to `Chunk` element

`ChunkID`	Constituent ID of this chunk.
`OrthographicTranscription`	Orthographic transcription of this chunk.
`PhoneticTranscription`	Phonetic transcription for this chunk by katakana.
`Disfluency`	True if this chunk is a slip of the tongue (mispronunciation) or a repetition. Some disfluent chunks are followed by a self repair; others are not.
`Filler`	True if this chunk is a filler.
`Backchannel`	True if this chunk is a backchannel (aizuchi).
`Conjunction`	True if this chunk is a conjunction defined in (The Japanese Discourse Research Initiative, 2000).
`DiscourseMarker`	True if this chunk is a discourse marker defined in (The Japanese Discourse Research Initiative, 2000).
`EndOfSlashUnit`	True if the slash unit ends prematurely at this chunk.

3. Dialogue Recording

The current release of the UU Database includes natural and spontaneous dialogues of seven pairs of college students (12 females, 2 males). The participants and pairings were selected carefully to ensure that both people in each pair were of the same grade and able to get along well with each other. All of the participants used "tame-guchi (lit. equal speech)" style in speaking to the partner. Tame-guchi is an informal speaking style commonly used in everyday conversation between close friends or relatives. Because tame-guchi is a colloquial speech where sentence-final elements signaling grammatical moods are often omitted, paralinguistic information was expected to play a larger role in their dialogues.

For the 4-frame cartoon sorting task, cartoons whose story could not be understood merely by reading the script but required some explanation of the scenes were selected. Each pair participated in three to seven independent sessions, using different cartoon materials for each session. A sample material is shown in Fig. 1.

The sessions were performed without eye contact, allowing speech only, and the recording was carried out in a soundproof room. The participants wore headsets (Sennheiser HMD-25), through which the partner's voice could be heard. Although this environment is not ideal in terms of the naturalness of daily emotional life, the participants soon became immersed in the fun exercise and focused on the conversations. The objective of the experiment (to build a spoken dialogue database) was not informed to the participants in advance[1]. The speech data was digitally recorded by DAT. The sampling frequency was 44.1 kHz. In total, dialogue speech was recorded for 27 sessions, lasting about 130 minutes. The recordings were then transcribed with some markups (e.g. backchannel responses, fillers, discourse markers, etc.) Table 2 lists the full inventory of attributes of the `Chunk` element defined in the UUDB XML Document Format. The whole speech signal was segmented into 4840 utterances, where an utterance is defined as a speech continuum bounded by either silence (> 400 ms) or *slash unit* boundaries. A slash unit is something like a sentence in spoken dialogue. See Meteer et al. (1995) and The Japanese Discourse Research Initiative (2000) for the definition of slash units. The details of the slash unit labeling for the UU Database was reported in Mori (2007).

The details of the segmentation procedure were as follows:

1. Extract each channel from the recorded two-channel dialogue speech.
2. Detect endpoints of silences that were 400 ms or longer were detected using the Entropic ESPS tool `find_ep`.
3. Manually adjust incorrect boundaries (mainly due to noise and long geminate stops).
4. If the boundary of a slash unit is in the midstream of the segmented waveform, split the waveform at the boundary.

The average utterance length was 1407 ms, and more than 30% of the utterances were shorter than 400 ms.

4. Annotation of Emotional States

The term "emotion" has multiple meanings. Cowie et al. (2001) carefully distinguished "full-blown emotion" (or "emotion" in the narrow sense) from "underlying emotion," which

[1]The participants were told the real objective of the recordings afterward, and each of them agreed to the distribution of their interactions for academic use with a written consent.

refers to an aspect of mental state that may influence a person's thoughts and actions but the actions are more or less under control. Not surprisingly, it is rather rare to find full-blown emotion in task-oriented dialogues like those in this research. Participants, on the other hand, show various (vocal and facial) expressions that possibly reflect their underlying emotion. Furthermore, one might even simulate emotion that has some communicative role, or to conform to a socially expected or preferred manner of behavior. Similar kinds of speaker's state might be associated with interpersonal relationships and attitudes; although they are not emotion proper, they are emotion-related states.

Speakers' states can be expressed both as linguistic and paralinguistic information. Linguistic information is always discrete, but some paralinguistic information is also discrete. For example, illocutionary acts which are marked by specific prosodic patterns (e.g. affirmation vs. interrogation) may be regarded as discrete information. However, emotion categories (Cowie and Cornelius, 2003) such as anger, happiness and sadness, have the obvious drawback that often there is no appropriate choice for an utterance when evaluating emotional states for spontaneous expressive speech (Campbell, 2003). Therefore, it is natural, or at least safer, to consider emotional states to be continuous because the range of our interest is not limited to full-blown emotions such as anger.

We attempted to annotate any kind of emotion and emotion-like states described above (hereafter *"emotional states"*) in a manner coincident with continuity, rather than category labels. The annotation is expected to cover a large part, but not all, of speakers' states that can be expressed as paralinguistic information.

We used the following six scales for evaluating emotional states:

(1) pleasant-unpleasant
(2) aroused-sleepy
(3) dominant-submissive
(4) credible-doubtful
(5) interested-indifferent
(6) positive-negative

These items were chosen in order to capture several aspects of emotional states, each of which is based on a psychological background as described below. Note that the choice of scales does not mean they are sufficient for evaluating any kinds of emotional state; rather, they are necessary for examining the effectiveness of our database design from the viewpoint of richness in expressivity.

Personal emotional state Items (1) and (2) are for measuring the emotional state of a speaker. The levels of pleasantness and arousal are the most common measures of emotional states and have been repeatedly adopted in emotion studies (e.g., Russell and Bullock (1985); Russell et al. (1989)) since Schlosberg (1952).

Interpersonal relationship Items (3) and (4) concern interpersonal relationships. Dominance (or control) is one of the most studied interpersonal relationships (Burgoon et al., 1989; Palmer, 1989; Stang, 1973), and it was found that we can correctly decode the dominance of a person via nonverbal behavior including paralinguistic cues (Rosenthal et al., 1979). Many studies examined the effect of credibility (or confidence) and deception of the speaker (Keeley and Hart, 1994; Miller and Stiff, 1993), and found that each of these is one of the most important factors affecting the outcomes of a persuasive communication. Because both speakers can be less confident in information gap tasks, doubtful utterances are rather common even though they do not intend to cheat each other.

Attitude Items (5) and (6) are related to the attitude of the speaker to the partner. Attitude is the belief that guides or directs an individual's perception and behavior, which is a critical factor in considering human interaction as well as interpersonal relationships. Positivity includes positive/negative attitudes such as liking, intimacy and affiliation (Argyle and Dean, 1965; Patterson, 1991; Warner et al., 1987). Previous studies of interpersonal behaviors revealed that interest (Duck, 1993) and positive attitude are one of the keys affiliating the participants of a conversation (Argyle et al., 1972; Mehrabian and Ksionzky, 1974).

Regardless of the causes and effects of the aspects mentioned above, all the annotations of emotional states provided by the UU Database are designed to be effect-oriented. In other words, the annotators are to judge "what impression the speaker's way of speaking will have on the listener," rather than "what kind of emotion or message causes the speaker's way of speaking."

5. Statistical Nature of Emotional State Annotation

One of the major difficulties in annotating emotional states for the UU Database was that there was no rigid theoretical foundation that justifies its evaluation and description. Therefore, the authors first tried to explore the nature, at least the statistical nature, of the ratings of emotional states by a large number of annotators.

5.1. Experiment 1: Consistency

The subjects consisted of 13 paid persons in their 20s. They were asked to evaluate the perceived emotional states of the speakers for each utterance on a 7-point scale. In evaluating the pleasant-unpleasant scale, for example, 1 corresponds to extremely unpleasant, 4 to neutral, and 7 to extremely pleasant.

The subjects used the iTunes software to listen to the stimuli. Repeated playback was not allowed. They were instructed as follows:

- Evaluate all the six scales after finishing each playback.
- Evaluate your impression perceived from the way of speaking, not from the content itself.

Prior to the work, the basic theory of emotion dimensions and meanings of each dimension were explained to the subjects. Then, they listened to an example set composed of eight utterances which were considered to accurately express the character of each dimension. In addition, they practiced annotating another four utterances.

The stimuli in Experiment 1 included twelve utterances picked from the corpus, four of which have a high degree of emotional prominence, four with intermediate, and four with low. The entire set of stimuli was composed of $12 \times 8 = 96$ utterances, so each utterance appeared eight times. The order of presentation was randomized.

Before showing the main results, a discussion about the learning effects of annotators is given here. Variances of the emotional state ratings for each individual utterance and for each annotator were calculated separately over the first four repetitions and the second four repetitions, then the variances for the first and the second were compared. If the annotators became more consistent for later presentations, the variance for the second half should be lower than the one for the first half. The number of samples that the variance for the second half was significantly lower (one-tailed F-test, $p < 0.05$) ranged from 28 (17.9%, aroused-sleepy) to 35 (22.4%, pleasant-unpleasant) out of 156 (= 12 utterances $\times$ 13 annotators) samples, which implies that little evidence of the learning effects was observed.

Figure 2 shows the standard deviation (SD) of the emotional state ratings for each individual utterance[2]. The horizontal axis indicates the annotator IDs. The SDs of ratings for each scale and for each annotator were computed over the eight repetitions, each for the twelve utterances. This gives twelve data points per annotator and scale. The lines in boxes show the lower quartile, median, and upper quartile of the SD values. The figure shows that most of the SD values were as low as 0.5. For example, if a subject gave ratings of 4, 4, 4, 4, 4, 5, 5, 5 for an utterance, its SD was 0.518. This implies that most subjects could provide rather consistent ratings for the same stimulus. Some annotators (e.g. #8) were inconsistent for most evaluation items, while others (e.g. #2, #9) were relatively consistent.

Among the six scales, ratings for dominant-submissive showed high deviation as a whole. Especially, annotator #12 rated the dominant-submissive scale in a very inconsistent manner for several utterances. The ratings for the utterances were concentrated at 2 (very submissive) and 6 (very dominant). This could be attributed to the utterances' ambiguity in terms of dominance. For example, one of the utterances expressed strong agreement with her partner's proposal (the last utterance in Figure 4). This could be dominant because the agreement dominated the direction of dialogue; at the same time, it could be submissive because she followed her partner's idea.

5.2. Experiment 2: Inter-Rater Agreement

The stimuli in this experiment were 215 utterances of two sessions by two pairs. The subjects consisted of 14 paid persons including all the subjects for Experiment 1. Two orders of presentation were used. At first, subjects were presented the stimuli in a random order. Following the emotional state evaluation for 215 utterances, they were presented the same utterance set in the correct order of the original sessions. In the random condition, the subjects did not know the context of each utterance; in the in-order condition, they did. Applying this fixed order of experiment condition (first random, then in-order) to all subjects might suffer from the order effect; however, the in-order presentation may allow the subjects to easily understand and memorize the whole story of the sessions. This may also cause another serious side effects on experiments for the random condition when ones for the in-order condition performed first, so the order of experiment was not counterbalanced.

Table 3: Kendall's W in evaluating emotional states by 14 raters. (**: $p < 0.01$)

	random	in-order
pleasant-unpleasant	0.51^{**}	0.48^{**}
aroused-sleepy	0.61^{**}	0.58^{**}
dominant-submissive	0.46^{**}	0.45^{**}
credible-doubtful	0.36^{**}	0.33^{**}
interested-indifferent	0.40^{**}	0.34^{**}
positive-negative	0.31^{**}	0.29^{**}

Table 3 lists Kendall's W coefficient of concordance, which is generally used to assess agreement among more than two raters, ranging from 0 (no agreement) to 1 (complete agreement). For all evaluation items, the coefficient was statistically significant and the null hypothesis "there is no agreement" was rejected. The results also show that the inter-rater agreement in evaluating emotional states was medium to low, but the coefficients were relatively high ($\simeq 0.5$) for the first three evaluation items.

In our previous work, it was shown that the order of presentation had a significant effect on ratings for most subjects (Mori et al., 2005). The results shown in Table 3 imply that the effect of presentation order did exist, namely, random presentation gave more concordant evaluations. However, the difference was subtle.

6. Screening and Actual Annotation

The statistical investigation described in the previous section suggested that we could obtain fairly reliable emotional state annotation if we could employ good annotators. To select annotators for rating emotional states for the UU Database, we first recruited six female adults, and engaged them in a screening test. They were asked to evaluate an utterance set that was composed by unifying the sets described in Sections 5.1 and 5.2.

The screening criteria were as follows.

Criterion 1: Consistency The variance of ratings for the same stimulus should be as low as that of a "good annotator" described in Section 5.1.

Criterion 2: Correlation with the average The tendency of ratings should not differ much from those of the great majority described in Section 5.

[2] Subject #13 did not participate in Experiment 1. The IDs were not renumbered to avoid confusion across different publications.

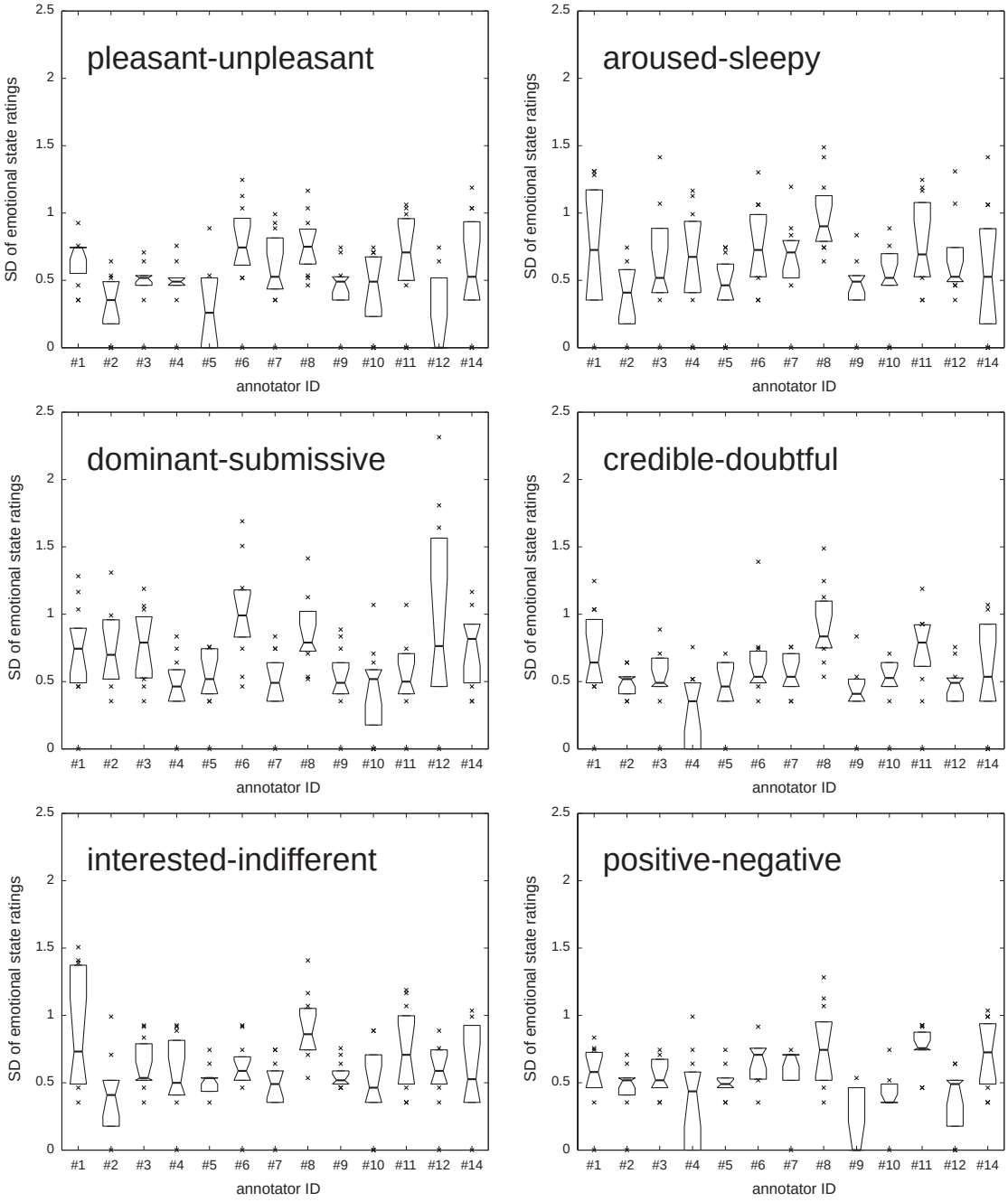

Figure 2: Distribution of the standard deviation of the emotional state ratings for identical utterances.

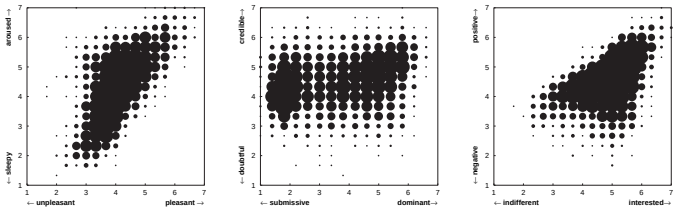

Figure 3: Distribution of rated emotional states for the corpus. The circle area is proportional to the number of occurrences.

Criterion 3: Distinction of evaluation items The annotator should independently interpret each evaluation item and its paralinguistic aspect. This implies that the annotator should not repeatedly assign uniform ratings (e.g. "7" for all).

As a result, two of the six annotators were screened out because they did not meet Criteria 1 and 3. More specifically, the SDs of the two annotators tended to exceed 0.8, and one of them assigned uniform ratings for more than 30% of the evaluations (others less than 6%). Another annotator was unable to continue for personal reasons. Finally, three qualified annotators were selected to evaluate the rest of the corpus. The whole annotation job took approximately four months. Complete (3 of 3) and partial (2 of 3) agreement were 22.09% (chance: 2.04%) and 83.92% (chance: 38.78%), respectively. Kendall's W ranged from 0.56 to 0.81.

Figure 3 shows the distribution of averaged ratings for the six dimensions. The ratings are distributed over a broad range, which means the database covers a wide variety of expressive speech.

Figure 4 shows a short dialogue fragment from the UU Database (originally in Japanese). The actual database is a set of XML structured documents.

	pleasantness	dominance		interest		
		arousal		credibility		positivity
R:{breath}Huh/	2.7	5.3	3.7	3.3	5.3	3.3
R:Um wait/	2.3	3.0	3.7	2.3	5.0	2.7
R:I don't get it, I don't/	2.3	4.7	3.3	3.0	5.3	3.0
L:Um your A what it is/	5.0	5.7	5.3	4.7	5.7	4.3
R:A is, Snoopy, is holding a ball, and the boy is, you know, (D sashite) reaching out his hand/	5.7	6.3	5.7	6.0	6.0	6.0
L:{B uh}{laugh}	5.7	5.0	3.0	5.3	5.3	5.3
L:{B uh}	5.7	4.7	3.0	5.3	5.0	5.3
L:{B uh}	3.3	2.7	2.3	4.3	5.0	4.7
L:After all (D ku), opposite isn't it/	4.7	6.3	5.3	4.3	6.0	4.3
L:The little one first fetched and passed to, Snoopy, then Snoopy fetched/	5.0	6.0	5.7	4.7	5.7	5.0
R:{B uh}	3.7	4.3	3.0	4.3	4.3	4.7
R:Oh that's it maybe/	6.0	6.7	5.3	6.7	6.7	6.3
L:Should be that one/	5.7	6.3	5.3	6.0	6.3	6.3
L:Maybe/	5.7	5.7	4.7	5.7	5.3	6.0
R:That's it that's it that's it that's it/	6.3	6.3	4.7	6.7	6.7	6.7

	pleasantness	dominance		interest		
		arousal		credibility		positivity
R:{D Then}	3.7	4.3	5.0	3.7	4.0	3.7
L:{B uh}	3.0	2.7	2.7	3.3	4.0	3.0
R:that, the last B isn't it/	3.7	4.0	5.0	3.0	5.0	3.0
L:The last B isn't it/	4.3	5.0	4.0	3.7	5.7	3.7
R:Saying it has arrived/	4.0	4.0	4.0	4.7	4.3	5.0
L:I agree/	3.7	3.7	2.7	4.0	3.7	4.0
R:Why the heck mame-mochi/	3.7	4.0	4.0	3.3	5.3	3.3
L:But it looks like	4.7	6.0	5.0	4.0	5.7	4.3
R:{B uh}	3.0	2.7	2.7	3.7	3.3	3.3
L:in D, the girl is saying oh and seems to be noticing something/	3.0	6.0	5.3	4.3	5.7	4.3
R:{B uh}	3.7	3.3	2.7	4.0	4.3	4.0
R:Ah ah ah ah ah/	4.0	4.0	2.7	4.7	4.3	4.7
L:Oh, I got it/	6.0	7.0	6.0	6.0	7.0	6.0
R:Yes/	4.3	4.0	3.0	4.3	4.3	4.7

	pleasantness	dominance		interest		
		arousal		credibility		positivity
R:Huh what a bother isn't it/	3.7	5.0	4.7	4.3	5.7	3.7
L:Why/	3.7	5.0	4.7	3.7	5.3	3.7
R:Um{C because}ha/	4.0	5.3	4.7	4.0	5.7	3.7
L:{B Uh}	3.3	2.3	3.3	3.7	3.7	3.7
R:Ha{C but}{C then}there's no particular reason it's at the base, is there/	3.7	5.0	5.3	3.7	5.3	2.7
R:Isn't that OK if it were at the front-end all along/	3.7	4.3	5.0	4.3	5.0	4.0
L:{sigh}If you say that, you are almost spoiling the world of four-frame cartoon/	4.0	5.3	5.7	3.7	5.0	3.0
R:{laugh}	6.0	5.7	3.7	5.3	5.7	5.0

Figure 4: Some excerpts of dialogues and their corresponding emotional state annotation (averaged values). {B }: backchannel response, (D): disfluent portion, (C): conjunction.

7. Acoustic Correlates of Emotional States

Studies on speech and emotion have revealed that the acoustic cues of expressive speech involve information about f_0, loudness, duration, and voice quality (Murray and Arnott, 1993; Erickson, 2005). In terms of emotion dimensions, Schröder (2004) reviewed the literature and concluded that the correlation of activation (equivalent to the aroused-sleepy dimension) with mean f_0, mean intensity and speech rate is assured, but there is less evidence regarding the acoustic correlates of evaluation (equivalent to the dimension of pleasant-unpleasant) and power (equivalent to the dimension of dominant-submissive). So he conducted his own analysis of the Belfast Naturalistic Emotion Database and found a strong correlation between the activation dimension and f_0 parameters as in the literature, but the correlation with the evaluation dimension was reportedly much less stable.

In this section, speech parameters reflecting prosody and voice quality are analyzed as correlates of perceived emotional states of speakers using the UU Database. Because of its rich

content including expressive utterances and well-controlled recording environment, this analysis is expected to open up new possibilities in the field of speech and emotion. The inter-speaker variability is also considered in this section.

7.1. Speech Parameters

Fundamental frequency (f_0) is a speech parameter that has been widely reported to be linked to emotional expressions, especially to the activation dimension (Scherer, 1989). f_0 analyses were performed using Praat, and erroneous values (octave jumps) were corrected using Praat's PitchEditor. Then the **f_0 range** and **f_0 slope** were calculated for each utterance in the semitone scale. The f_0 range is defined as the difference between the highest and lowest f_0 values. The f_0 slope, defined as the first-order regression coefficient of the f_0 contour, can be meaningful only for short utterances like the interjection "e" (Campbell and Erickson, 2004). Both parameters are considered to be less sensitive to the f_0 baseline intrinsic to speakers.

Intensity is another parameter that is believed to be a strong vocal correlate to emotional expressions. Again, Praat was used to obtain intensity curves, then **peak intensity** was obtained for each utterance. Because the recording level was adjusted individually for each speaker, the intensity values are directly comparable only within speaker from the viewpoint of speech production. However, because the playback level was kept constant over all annotators, the intensity values are comparable throughout the speakers from the viewpoint of speech perception.

Utterance length is an utterance's number of morae (not a phonetic but a phonological feature).

Speech rate is defined as an utterance's duration (not including pauses inside) divided by utterance length.

Among the various kinds of voice quality parameters that can be automatically obtained from the speech signal, we focused on the one related to the breathiness of voicing source, because glottal noise is associated with utterances expressing suspicion and disappointment from the viewpoints of both speech production and perception (Kasuya et al., 1999, 2000), which are likely to be related to unpleasantness. As the parameter reflecting glottal noise, we used $f_{\text{aperiodic}}$ (Ohtsuka and Kasuya, 2001; Mori and Kasuya, 2007); this is defined as the boundary frequency that partitions the whole band of the estimated source signal into a lower band in which harmonic components dominate, and a higher band in which aperiodic components dominate. A lower value of $f_{\text{aperiodic}}$ indicates a relative increase of the aperiodic component to the periodic one. For each utterance, **mean $f_{\text{aperiodic}}$** is defined as the average of $f_{\text{aperiodic}}$ for voiced frames.

7.2. Relationship between speech parameters and emotional states

Correlation coefficients between the six speech parameters described above and averaged emotional state ratings for the 14 speakers in the UU Database were calculated. The dataset included all utterances in the UU Database except those composed entirely of nonlinguistic sounds such as laughter. The number of utterances for each speaker ranged from 153 to 741.

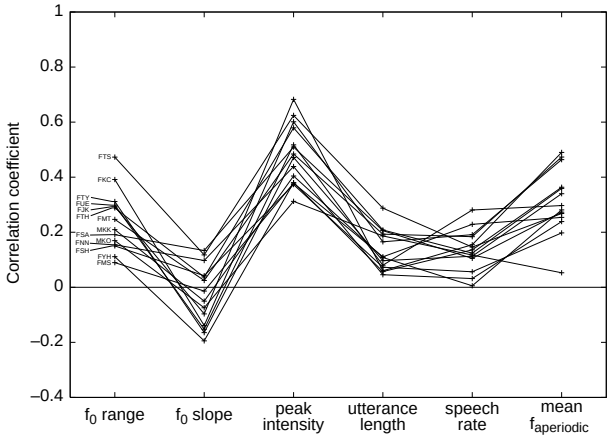

Figure 5: Correlation coefficients of f_0 range, f_0 slope, peak intensity, utterance length, speech rate, and mean $f_{\text{aperiodic}}$ for the dimension of pleasant-unpleasant. Each line corresponds to a speaker.

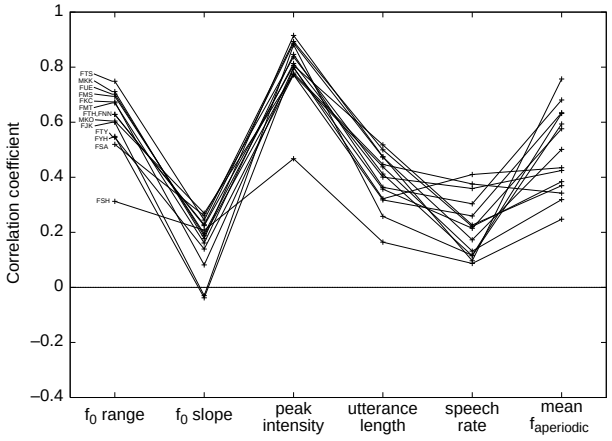

Figure 6: Correlation coefficients of f_0 range, f_0 slope, peak intensity, utterance length, speech rate, and mean $f_{\text{aperiodic}}$ for the dimension of aroused-sleepy. Each line corresponds to a speaker.

Figures 5 to 10 show the distribution of correlation coefficients between each speech parameter and emotional state ratings. The following subsections discuss the details for the dimensions of pleasant-unpleasant, aroused-sleepy, and dominant-submissive.

7.2.1. Pleasant-unpleasant

Figure 5 shows the distribution of correlation coefficients between each speech parameter and averaged "pleasantness" ratings for the 14 speakers. From the figure, it can be observed that the correlation of the dimension of pleasant-unpleasant is:

- high with peak intensity (mean: 0.48),
- relatively high for some speakers with mean $f_{\text{aperiodic}}$,
- moderately high with f_0 range (mean: 0.24), and
- low with utterance length and speech rate; inconsistent with f_0 slope.

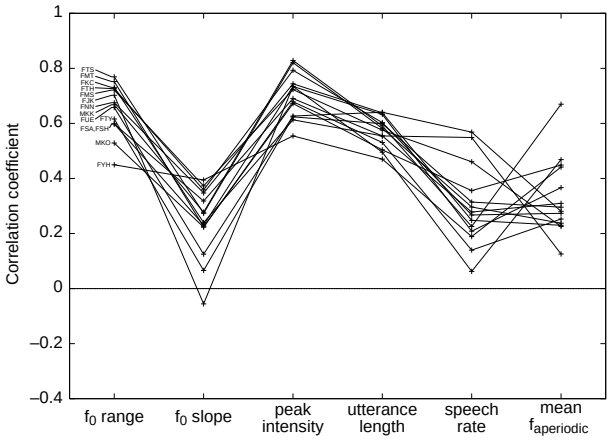

Figure 7: Correlation coefficients of f_0 range, f_0 slope, peak intensity, utterance length, speech rate, and mean $f_{\text{aperiodic}}$ for the dimension of dominant-submissive. Each line corresponds to a speaker.

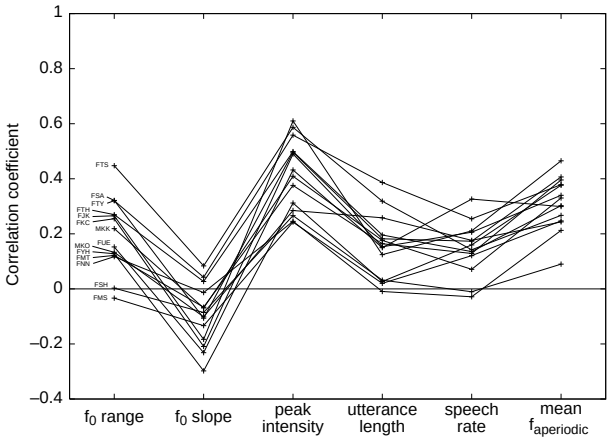

Figure 8: Correlation coefficients of f_0 range, f_0 slope, peak intensity, utterance length, speech rate, and mean $f_{\text{aperiodic}}$ for the dimension of credible-doubtful. Each line corresponds to a speaker.

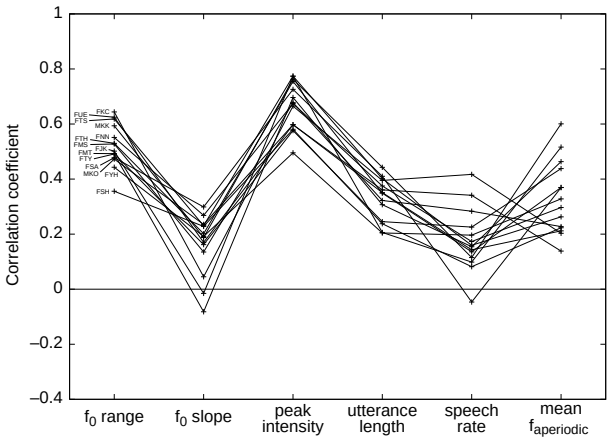

Figure 9: Correlation coefficients of f_0 range, f_0 slope, peak intensity, utterance length, speech rate, and mean $f_{\text{aperiodic}}$ for the dimension of interested-indifferent. Each line corresponds to a speaker.

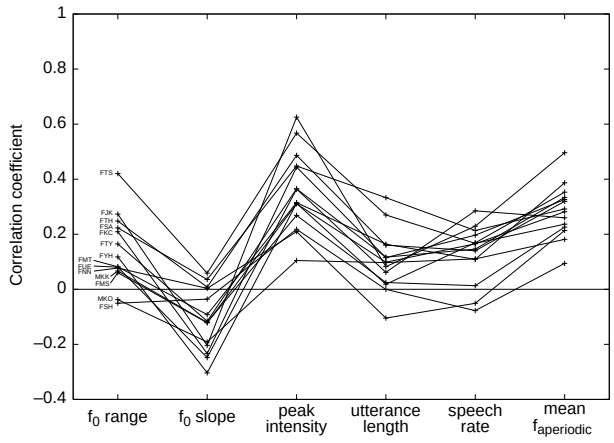

Figure 10: Correlation coefficients of f_0 range, f_0 slope, peak intensity, utterance length, speech rate, and mean $f_{\text{aperiodic}}$ for the dimension of positive-negative. Each line corresponds to a speaker.

Similar tendencies were observed for the dimensions of credible-doubtful and positive-negative.

The positive correlation with f_0 range and speech rate is compatible with the previous perception study, where tone sequences were used as stimuli (Scherer and Oshinsky, 1977). The positive correlation with $f_{\text{aperiodic}}$ can be interpreted as meaning that some speakers used breathy phonation in speaking in unpleasant states such as suspicion, uneasiness or uncertainty. Finally, a positive correlation with intensity was observed, which may partly reflect the positive correlation between pleasant-unpleasant and aroused-sleepy dimensions in the UU Database (see Fig. 3); if it contained a larger number of unpleasant and aroused utterances such as hot anger, the correlation coefficient would be smaller.

7.2.2. Aroused-sleepy

Figure 6 shows the distribution of correlation coefficients between each speech parameter and averaged "arousal" ratings for the 14 speakers. From the figure, it can be observed that the correlation of the dimension of aroused-sleepy is:

- very high with peak intensity (mean: 0.80), and
- high with f_0 range (mean: 0.61)

except for one speaker. In addition,

- speaker-sensitive with mean $f_{\text{aperiodic}}$, and
- relatively low with f_0 slope, utterance length, and speech rate.

Similar tendencies were observed for the dimension of interest-indifferent.

Arousal's positive correlation with f_0 range and intensity is perfectly compatible with previous studies (Huttar, 1968; Scherer, 1989). Some speakers tend to produce short utterances with very weak and breathy phonation, which seems to be a cause of the positive correlation between arousal and $f_{\text{aperiodic}}$.

7.2.3. Dominant-submissive

Figure 7 shows the distribution of correlation coefficients between each speech parameter and averaged "dominance" ratings for the 14 speakers. From the figure, it can be observed that the correlation of the dimension of dominant-submissive is:

- high with f_0 range, peak intensity and utterance length,
- moderately high for most speakers with f_0 slope, and
- speaker-sensitive with speech rate and mean $f_{\text{aperiodic}}$.

This tendency was similar to that of the aroused-sleepy dimension for f_0 range, intensity and $f_{\text{aperiodic}}$. It is also reasonable that utterances with more content (=longer utterances) are perceived as dominant. Figure 7 also suggests that steeper f_0 slope (=larger negative slope) is associated with submission. A possible interpretation is that the intonation pattern of dominant utterances tends to be raised toward the end of an utterance, which can reduce f_0 declination.

8. Emotional State Recognition

Based on the acoustic parameters described in the previous section, we examined how accurately machines can estimate a given speaker's emotional states as perceived by human raters.

Although nonlinear machine learning methods including the regression tree and model tree were tested as well, we did not see any significant improvement over the linear model. We therefore describe only the results of the linear method here.

The predicted emotional state of the utterance i for the dimension k is modeled by the linear combination of standardized acoustic parameters of the utterance i ($x_{i1k} \ldots x_{i6k}$) as follows:

$$y_{ik} = \sum_{j \in \mathrm{FS}} \beta_{jk} x_{ijk} + \varepsilon_{ik}, \tag{1}$$

where x_{ijk} is one of the six acoustic features (e.g. f_0 range, f_0 slope, ...) that is normalized by subtracting its mean and dividing by its standard deviation. FS $\subseteq \{1, \ldots, 6\}$ is a subset of features selected so as to maximize the prediction power for unseen data.

Linear regression models were set up in two conditions: SpD (speaker-dependent) and SpI (speaker-independent). In the SpD condition, the speaker was assumed to be known, and the model for each speaker was trained with her/his utterances. In the SpI condition, the speaker was assumed to be unknown, and a single model was trained with all speakers' utterances. In the training procedure, feature selection was performed using the AIC (Witten and Frank, 2005).

Table 4 shows the standardized coefficients for the speech parameters, multiple correlation, and RMS error of the SpI linear regression model. The distributions of the RMS errors of the SpD and SpI models are illustrated in Fig. 11, where each hairline represents the RMS error for a speaker, and the bold line represents that of the SpI model for all speakers. The RMS errors were evaluated with 10-fold cross validation.

Table 4 indicates that the peak intensity has the greatest effect on all the emotion dimensions. Other major results are: (i) the f_0 range has a greater effect on the dimensions of arousal, dominance, and interest, (ii) the mean $f_{\text{aperiodic}}$ has a greater effect on the dimensions of pleasantness, credibility, and positivity. From the viewpoint of perception, this can be interpreted as follows.

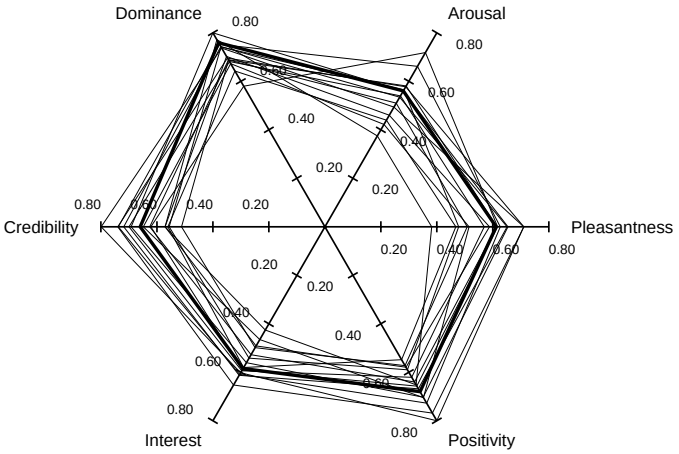

Figure 11: The RMS error of the SpD (thin) and SpI (thick) linear regression models.

- Loudness and pitch range principally account for arousal, dominance, and interest perceived from utterances.
- Loudness and breathiness principally account for pleasantness, credibility, and positivity perceived from utterances.

Other minor influences include: f_0 slope contributed to submission, credibility, and positiveness; utterance length contributed to dominance. The speech rate had only a small effect.

From the fairly high (> 0.7) multiple correlation in Table 4, it is understood that the overall tendency in estimating emotional states with the SpI model is very close to that of human raters for the dimensions of arousal, dominance, and interest. Regarding similarity to human perception, we could precisely estimate these emotional states from speech signals. The precision for the other dimensions was somewhat worse.

From Fig. 11, we can see that the RMS error of the SpD models was distributed in the vicinity of 0.6 for the dimensions of pleasantness, arousal and interest. The value is comparable to the typical intra-rater SD value of 0.5 described in Section 5.1. In other words, speakers' pleasantness, arousal and interest can be estimated from the speech parameters nearly as accurately as by human raters.

For the dominance dimension, on the other hand, the RMS error was relatively high. The result corresponds with previous review of Schröder (2004), that there is less evidence regarding the acoustic correlates of power. This may be also related to the ambiguity in evaluating dominance, described in Section 5.1.

The RMS error for the dimensions of credibility and positivity was also high; however, the intra-rater SD values for these dimensions were not so high, meaning that the speech parameters in the current study are not sufficient to estimate a

Table 4: Speech parameters and their corresponding standardized coefficients (with 95% confidence intervals), multiple correlation, and RMS error of the SpI linear regression model.

	Standardized coefficients β						Mult. correl.	RMS error
	f_0 range	f_0 slope	peak intensity	utt. length	speech rate	mean $f_{\text{aperiodic}}$	R	
pleasantness		–0.06±0.02	0.39±0.02	–0.04±0.02	0.07±0.02	0.21±0.02	0.52	0.61
arousal	0.22±0.02	0.05±0.02	0.60±0.02	–0.04±0.02	0.07±0.02	0.15±0.02	0.84	0.56
dominance	0.27±0.03	0.12±0.02	0.46±0.03	0.17±0.03	0.09±0.02	–0.09±0.03	0.79	0.76
credibility	–0.05±0.03	–0.11±0.02	0.37±0.03		0.10±0.02	0.19±0.02	0.49	0.66
interest	0.22±0.02	0.07±0.02	0.54±0.02	–0.04±0.02	0.03±0.02		0.70	0.58
positivity	–0.08±0.03	–0.13±0.02	0.31±0.03		0.10±0.02	0.22±0.02	0.44	0.68

speaker's credibility and positivity. A further inspection of utterances with large error suggested that ratings for most of them were influenced by their linguistic contents and contexts.

9. Conclusion

In this paper, the Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies was introduced. It is the first public speech corpus specifically designed for studies on paralinguistic information in expressive Japanese dialogue speech.

We devised a useful method for stimulating expressively-rich and vivid conversation. With the "4-frame cartoon sorting task," the participants were successfully interested and motivated. The effectiveness of the method was also supported by the fact that the subjective ratings for emotion dimensions were distributed over a broad range, which means the database covers a wide variety of expressive speech.

Although emotion and expressivity as paralinguistic information delivered by speech is one of the most important aspects of speech communication, there is no common ground for labeling such information. We attempted to annotate perceived emotional states of speakers with the six abstract dimensions: pleasant-unpleasant, aroused-sleepy, dominant-submissive, credible-doubtful, interested-indifferent, and positive-negative. The dimensions were defined based on various psychological backgrounds, and were intended to capture the aspects of personal emotional state, interpersonal relationship, and attitude. The appropriateness of these dimensions is still an open question. On the one hand, there certainly exists a redundancy in the six scales; for example, the correlation coefficient between the dimensions of credible-doubtful and positive-negative is as high as 0.88. On the other hand, the paralinguistic phenomena that we can observe from this corpus are merely the reflections of some (not all) aspects of our emotional/social lives. For example, the UU Database contains very few calm, relaxed utterances because it is task-oriented. Further assessment using multiple speech corpora will help clarify the question.

Prior to the actual annotation work, we examined the statistical nature of emotional state ratings with the abstract dimensions by a large number of annotators. Consequently, it was found that most annotators could provide fairly consistent ratings for a repeated identical stimulus. The standard deviation values for most stimuli were as low as 0.5 for the scale of 1 to 7. It was also revealed that the inter-rater agreement in evaluating emotional states was reasonably good, at least for the dimensions of pleasant-unpleasant, aroused-sleepy, and dominant-submissive.

Three annotators were selected based on the three criteria: consistency, correlation with the average, and distinction of evaluation items. The qualified annotators assigned emotional state labels with abstract dimensions for all 4840 utterances. The high degree of agreement among the three annotators was verified by the complete/partial agreement rate and Kendall's W.

The validity of the emotional state annotation was also examined by its relevance to acoustic parameters. Among them, peak intensity and f_0 range were highly correlated with most dimensions. In addition, a voice quality parameter, $f_{\text{aperiodic}}$, showed a considerable effect for some dimensions depending on the speaker.

Finally, a linear regression-based estimator of perceived emotional states of speakers was tested. Using speaker-independent models, we obtained fairly high (> 0.7) multiple correlation for the dimensions of arousal, dominance and interest. We also showed that the RMS error of the speaker-dependent models was as low as 0.6 for the dimensions of pleasantness, arousal and interest. Overall, it is concluded that the perceived emotional states of speakers can be accurately estimated from the speech parameters in most cases.

It is desirable to distinguish paralinguistic causes from linguistic ones in analyzing effects on speech parameters. But as is the case for other corpora of spontaneous (uncontrolled) speech, separating the two causes is not straightforward. One possibility is to limit the modeling to a world of closed vocabulary, such as interjections. Mori and Kasuya (2007) showed the effect of phonetic variations on perceived emotional states for the single-vowel interjection /a/ in the UU Database. Another possibility is to use linguistic information together with paralinguistic information such as emotional states to model acoustic variations of speech. The authors are now working to append linguistic information such as dialogue act tags to the UU Database in order to analyze the interaction of the two

causes (Satake and Mori, 2009).

There should be considerable variations in manifestations of emotion depending on gender, social background, interpersonal relationship, and culture. Because the dialogues were recorded by Japanese college students, most of whom were female, the findings described in this paper cannot directly be generalized. Related to this, the distinctness of the "tame-guchi" speaking style was not covered in this paper. Although tame-guchi is a very interesting, commonly used speaking style, few speech corpora have focused on it. The pragmatics of tame-guchi would also be an interesting topic in sociolinguistics, so the authors hope the database will serve as an example of the style in such studies.

The current release of the UU Database is compact, and is too small to meet all demands in speech and language science as well as applications such as speech recognition and synthesis. Nevertheless, using the UU Database we have obtained many important findings in various fields such as the analysis of expressive speech (Mori and Kasuya, 2007), facial expression generation of virtual characters (Mori and Ohshima, 2008), prosody in discourse (Mori, 2009), and so on. The authors believe that these achievements reflect the correct choice of corpus design, in which cost effectiveness was given priority over other demands under the limitation of manpower, budget and time.

The UU Database is distributed by the NII Speech Resources Consortium without charge for academic use. It is also downloadable from the UU Database website at http://uudb.speech-lab.org/ .

A. A typical example of the cartoon sorting task dialogue

The section gives a typical dialogue, which is taken from session C032 of the UU Database, where two female speakers, FKC and FUE, participated. The cartoon material was taken from "Peanuts" by C. M. Schulz, and shuffled as explained in Section 2. In this session, the speaker FUE had cards A and B, and the speaker FKC had cards C and D, as shown in Fig. 1.

If one can see all the four cards together, it is quite easy to guess the correct order—C, B, D, A. In this study, however, both participants had only incomplete, but complementary, information for understanding the story. So, in most cases participants first tried to describe the information drawn on their own cards to the partner. The progress of the dialogue is summarized below.

(1) Both the participants became embarrassed because there were no words at all.
(2) FKC requested FUE to explain the cards A and B.
(3) FUE tried to explain the contents of card A but she did not know the name of the character. FKC called him "a tiny, yellow one" and explained what he and Snoopy were doing. Then she explained the action of "a boy" by arbitrarily making up his speech.
(4) FUE explained that on card B "the yellow one" was running.
(5) FKC explained card C by imagining the thoughts of Snoopy and "the yellow one."
(6) FKC explained card D. On illustrating the action of "the yellow one," she said something like "Here it is, sir."
(7) FKC presented her idea that first "the boy" threw the ball, then Snoopy fetched and threw it, and finally "the yellow one" fetched it.
(8) FUC was suspicious about Snoopy's throwing. FKC's opinion was indefinite.
(9) FUE tried to make a partial agreement that the throwing should come first.
(10) FUE tried to explain her own cards again.
(11) FKC proposed another story but FUE ignored it.
(12) FUE asked FKC to explain cards C and D again.
(13) FUE was confused because the ball fetching was illustrated on multiple cards, and she groaned with impatience.
(14) FKC asked FUE to reexplain card A, and FUE did so.
(15) FKC presented an alternative story. FUE excitedly agreed with her.
(16) Both participants cooperatively organized the discussion and concluded that the answer was B, D, A, and C.

After all, the participants reached the wrong answer, but correctness does not matter in corpus collection.

References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E. H., Doherty, G. M., Garrod, S. C., Isard, S. D., Kowtko, J. C., McAllister, J. M., Miller, J., Sotillo, C. F., Thompson, H. S., Weinert, R., 1991. The HCRC map task corpus. Lang. Speech 34 (4), 351–366.

Argyle, M., Alkema, F., Gilmour, R., 1972. The communication of friendly and hostile attitudes by verbal and non-verbal signals. Eur. J. Soc. Psychol. 1 (3), 385–402.

Argyle, M., Dean, J., 1965. Eye contact, distance and affiliation. Sociometry 28 (3), 289–304.

Arimoto, Y., Kawatsu, H., Ohno, S., Iida, H., 2008. Emotion recognition in spontaneous emotional speech for anonymity-protected voice chat systems. In: Proc. Interspeech 2008. pp. 322–325.

Aubergé, V., Audibert, N., Rilliard, A., 2003. Why and how to control the authentic emotional speech corpora. In: Proc. Eurospeech 2003. pp. 185–188.

Burgoon, J. K., Buller, D. B., Woodall, W. G., 1989. Nonverbal Communication: The Unspoken Dialogue. Harper and Row, New York.

Campbell, N., 2003. The JST/CREST ESP project — a mid-term progress report. In: 1st JST/CREST Intl. Wkshp. Expressive Speech Processing. pp. 61–70.

Campbell, N., Erickson, D., 2004. What do people hear? A study of the perception of non-verbal affective information in conversational speech. J. Phonet. Soc. Jpn. 8 (1), 9–28.

Cowie, R., Cornelius, R. R., 2003. Describing the emotional states that are expressed in speech. Speech Commun. 40 (1–2), 5–32.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J. G., 2001. Emotion recognition in human-computer interaction. Signal Processing Magazine, IEEE 18 (1), 32–80.

Devillers, L., Vidrascu, L., 2006. Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogues. In: Proc. Interspeech 2006. pp. 801–804.

Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P., 2003. Emotional speech: Towards a new generation of databases. Speech Commun. 40 (1–2), 33–60.

Douglas-Cowie, E., Cowie, R., Schröder, M., 2000. A new emotion database: considerations, sources and scope. In: Proc. ISCA Wkshp. Speech & Emotion. pp. 39–44.

Duck, S. W. (Ed.), 1993. Social Context and Relationships. Vol. 3 of Understanding Relationship Processes. Sage, Newbury Park, CA.

Erickson, D., 2005. Expressive speech: Production, perception and application to speech synthesis. Acoust. Sci. & Tech. 26 (4), 317–325.
Greasley, P., Setter, J., Waterman, M., Sherrard, C., Roach, P., Arnfield, S., Horton, D., 1995. Representation of prosodic and emotional features in a spoken language database. In: Proc. 13th ICPhS. pp. 242–245.
Grimm, M., Kroschel, K., Narayanan, S., 2008. The Vera am Mittag German audio-visual emotional speech database. In: Proc. ICME 2008. pp. 865–868.
Horiuchi, Y., Nakano, Y., Koiso, H., Ishizaki, M., Suzuki, H., Okada, M., Naka, M., Tutiya, S., Ichikawa, A., 1999. The design and statistical characterization of the Japanese map task dialogue corpus. J. JSAI 14 (2), 261–272.
Huttar, G. L., 1968. Prosodic variables and emotions in normal American English utterances. J. Speech Hear. Res. 11 (3), 481–487.
Kasuya, H., Maekawa, K., Kiritani, S., 1999. Joint estimation of voice source and vocal tract parameters as applied to the study of voice source dynamics. In: Proc. 14th ICPhS. pp. 2505–2512.
Kasuya, H., Yoshizawa, M., Maekawa, K., 2000. Roles of voice source dynamics as a conveyer of paralinguistic features. In: Proc. ICSLP 2000. pp. 345–348.
Keeley, M., Hart, A., 1994. Nonverbal behaviors in dyadic interaction. In: Duck, S. W. (Ed.), Dynamics of Relationships. Vol. 4 of Understanding Relationship Processes. Sage, Newbury Park, CA, pp. 135–162.
Mehrabian, A., Ksionzky, S., 1974. A theory of affiliation. Lexington Books, Lexington, MA.
Meteer, M., Taylor, A., MacIntyre, R., Iver, R., 1995. Dysfluency annotation stylebook for the switchboard corpus. The Linguistic Data Consortium.
Miller, G. R., Stiff, J. B., 1993. Deceptive Communication. Vol. 14 of Interpersonal Communication. Sage, Newbury Park, CA.
Mori, H., 2007. Basic unit in annotating paralinguistic information for conversational speech. Tech. Rep. SIG-SLUD-A603-03, JSAI.
Mori, H., 2009. An analysis of switching pause duration in expressive dialogues as a paralinguistic feature. Acoust. Sci. & Tech. 30 (5), 376–378.
Mori, H., Aizawa, H., Kasuya, H., 2005. Consistency and agreement of paralinguistic information annotation for conversational speech. Journal of Acoustical Society of Japan 61, 690–697.
Mori, H., Kasuya, H., 2007. Voice source and vocal tract variations as cues to emotional states perceived from expressive conversational speech. In: Proc. Interspeech 2007. pp. 102–105.
Mori, H., Ohshima, K., 2008. Facial expression generation from speaker's emotional states in daily conversation. IEICE Trans. Inf. & Syst. E91-D (6), 1628–1633.
Morimoto, T., Uratani, N., Takezawa, T., Furuse, O., Sobashima, Y., Iida, H., Nakamura, A., Sagisaka, Y., Higuchi, N., Yamazaki, Y., 1994. A speech and language database for speech translation research. In: Proc. ICSLP 1994. pp. 1791–1794.
Murray, I. R., Arnott, J. L., 1993. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. J. Acoust. Soc. Am. 93 (2), 1097–1108.
Ohtsuka, T., Kasuya, H., 2001. Aperiodicity control in ARX-based speech analysis-synthesis method. In: Proc. Eurospeech 2001. Vol. 3. pp. 2267–2270.
Palmer, M. T., 1989. Controlling conversations: Turns, topics and interpersonal control. Commun. Monogr. 56 (1), 1–18.
Patterson, M. L., 1991. A functional approach to nonverbal exchange. In: Feldman, R. S., Rime, B. (Eds.), Fundamentals of Nonverbal Behavior. Cambridge University Press, New York, pp. 458–495.
Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L., Archer, D., 1979. Sensitivity to Nonverbal Communication: The PONS Test. Johns Hopkins University Press, Baltimore, MD.
Russell, J. A., Bullock, M., 1985. Multidimensional scaling of emotional facial expressions: Similarity from preschoolers to adults. J. Pers. Soc. Psychol. 48, 1290–1298.
Russell, J. A., Weiss, A., Mendelsohn, G. A., September 1989. Affect grid: A single-item scale of pleasure and arousal. J. Pers. Soc. Psychol. 57 (3), 493–502.
Satake, T., Mori, H., 2009. Estimation of speaker's emotional states based on discourse analysis of conversational dialogue. Tech. Rep. SIG-SLUD-A803-13, JSAI.
Scherer, K. R., 1989. Vocal correlates of emotional arousal and affective disturbance. In: Wagner, H. L., Manstead, A. S. R. (Eds.), Handbook of Social Psychophysiology. Wiley, pp. 165–197.
Scherer, K. R., Oshinsky, J. S., 1977. Cue utilization in emotion attribution from auditory stimuli. Motivation and Emotion 1, 331–346.
Schlosberg, H., 1952. The description of facial expressions in terms of two dimensions. J. Exp. Psychol. 44 (4), 229–237.
Schröder, M., 2004. Speech and emotion research: An overview of research frameworks and a dimensional approach to emotional speech synthesis. Ph.D. thesis, Saarland Univ.
Stang, D. J., 1973. Effect of interaction rate on ratings of leadership and liking. J. Pers. Soc. Psychol. 27 (3), 405–408.
The Japanese Discourse Research Initiative, 2000. Japanese dialogue corpus of multi-level annotation. In: Proc. 1st SIGdial. pp. 1–8.
Truong, K. P., Neerincx, M. A., van Leeuwen, D. A., 2008. Assessing agreement of observer- and self-annotations in spontaneous multimodal emotion data. In: Proc. Interspeech 2008. pp. 318–321.
Warner, R. M., Malloy, D., Schneider, K., Knoth, R., Wilder, B., 1987. Rhythmic organization of social interaction and observer ratings of positive affect and involvement. J. Nonverbal Behav. 11 (2), 57–74.
Witten, I. H., Frank, E., 2005. Data Mining, 2nd Edition. Morgan Kaufmann, San Francisco.