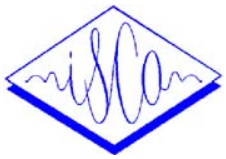

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

PARALINGUISTIC FEATURES AS SUPRASEGMENTAL ACOUSTICS OBSERVED IN NATURAL JAPANESE DIALOGUE

S. Kitazawa & S. Kobayashi
Department of Computer Science, Faculty of Information, Shizuoka University,
5-1, 3-Chome, Jouhoku, Hamamatsu, 432, JAPAN
kitazawa@cs.inf.shizuoka.ac.jp

ABSTRACT

The paralinguistic features, however this conference classifies the "Paralinguistic analysis" as Speaker identification, Keyword/topic spotting, and Language identification, include emotional aspects of voice, which is focused recently on interpersonal communications. Study starts from description and statistics of those features. Among many acoustic characteristics, we investigated features represented in pitch raising and lowering, loudness of speech, and rate of talking; these are not represented in textual meaning of speech that is the content of conventional speech recognition. The paralinguistic features are distinguished from prosodic features, however these two are often confused.

Keywords: paralanguage, prosody, spontaneous speech, transcription

1. INTRODUCTION

Non-verbal communication as a hyper concept of paralanguage is a communication behavior with non-linguistic gesture or physical feature, property, style of handwriting, and communications with non-linguistic speech sounds and features. Paralanguage is a suprasegmental feature as well as prosodic feature.
Trager listed components of paralanguage as Voice-Qualifiers, Vocal Characterizers, Vocal Qualifiers, and Vocal Segregates [1]. Fujisaki listed paralanguage according to the contents of the expression [2]. He classifies non-verbal speech as paralanguage and non-language. Paralanguage is information that is not language transmitted by speech sound, such as intention, manner and style. Non-language means information that is not related to the content of speech and that is not intentionally controllable, such as personal characteristics, physiological state and psychological state. Usual studies in nonverbal communication, Fujisaki's non-language is included in paralanguage. Here we follow this definition and we include non-language into paralanguage.

2. PARALANGUAGE

Paralanguage is not a language that has syntax and lexicon. Therefore measurement of paralanguage is only psychological.

2.1 Paralanguage and Prosodic Feature

2.1.1 Prosody
Phonology studies phonetic features related to distinction of meaning of a language. Prosody is also a concept interpreted within the phonology. Prosodic feature includes all of phonetic features except segmental phoneme: tone, intonation, stress and duration. These are presented as variation within segment or contrast between segments to mean semantic contrast.

2.1.2 Paralanguage
We define paralanguage as expressions related to speaker's corporal and mental states such as attitude, intention, and emotion. Even in read speech, fundamental frequency of speech is raised or changes pattern in emotionally emphasized speech.

2.2 Comparison of Prosody and Paralanguage

Paralanguage and prosody transmits different content. We consider prosodic and paralinguistic features as follows:
1. Prosodic features represent distinction of meanings, while paralinguistic features represent intentions, attitudes, feelings and others.
2. Both of them use pitch, intensity (loudness) and tempo (duration of segments) as physical and psychological measures of these features. Paralanguage uses various characteristics of voices as well.
3. When a speaker represents something in paralanguage, overlapped prosodic features seems to change simultaneously.
4. Paralinguistic voice characteristics continue for longer interval than prosodic (segmental) changes. Prosody is expressed in contrast within segments or between segments in a relatively short period.

3. METHOD

In order to study paralanguage as a subject, there are several problems: how to record natural spontaneous dialogue or utterance, how to instruct or not to instruct expression of paralanguage to participating informants,

and how to evaluate or judge paralanguages.

3.1 Recording of Spontaneous Speech

3.1.1 Set Up

Some researchers prepare test speech as performance of actors and actresses, which is a played speech. However we do not believe played speech contain real paralanguage. We tried to collect speech samples from natural situation. Speech samples are taken from spontaneous unprepared Japanese free dialogs recorded three times, each dialog lasts about an hour. They are informed that this recording is used for academic research purpose of paralanguage and dialogue. It takes several tens minutes to relax and to activate free discussion. There was no special instruction how to dialog or how to express paralanguages. We just wait for natural spontaneous speech to start to appear.

3.1.2 Informants

Each dialog participants are three young Japanese male native speakers from the same dialect, and they know each other well. We intended three participants are smoother to continue dialogue than two participants are.

3.1.3 Environments

Recordings were practiced in a sound proofed non-echoic room. For each speaker a directional microphone is assigned, and one non-directional microphone is placed at the center of three speakers. Recording equipment was 2 DAT recorders with 4 microphone channels. Speech sound was digitized through a DATLINK at 16kHz sampling rate.

3.1.4 Post Processing

Recorded sound includes overlapped utterances, noises from body movements such as kinesics, non-lexical speech such as coughs. Some utterances are too small to hear, while some are too loud. These difficult parts are inappropriate to use as test material. We prefer long monologue than short replies because a long monologue consists of several sentences accompanied with different kinds of paralinguistic information. In order to keep the amount of work for labelers within tractable range, 5 utterances of duration about 20 seconds from each of three subjects that is 15 utterances are extracted for test materials.

3.2 Labeling Task

3.2.1 Why Labeling

Paralanguage, like ordinary language, completes when a listener recognizes its contents. Only human observers can transcribe paralinguistic features. Here we study paralanguage through perceptual subjective identification in test materials.

3.2.2 What to Label

To ask to labeler to pick up paralanguage is too vague, because paralanguage includes wide ranges of features. We restricted to the following aspects of paralanguage in order to be clear to understand the feature to focus: voice pitch level, voice loudness, and speech rate. These features are seen in prosodic level at the same time, however, we instruct the subjects to label at those parts subjects recognize that speaker's intention, attitude, and emotion are expressed by voice pitch height, voice loudness, and speech rate. Direct aspects of paralanguage, such as speaker's attitude or emotion, however, are difficult to observe in cut out utterances but should be observed in the flow of entire dialogue. In order to keep the amount of work for labelers not too large, speech materials are excised part of dialogue.

3.2.3 Instructions to Labelers

We do not specify or ask subject to describe any contents of paralanguage. Subjects are asked to put pitch height, loudness, and speech rate labels where paralanguage is interpreted as being transferred as those features' changes. Generally speaking, paralinguistic features presence across much longer period of time than prosodic features that is specified on each prosodic word. We instructed to labelers to pay attention to voice height, loudness, and speech rate continuing across multiples of clauses, phrases, and further longer period.

3.2.4 Environments

Subjects could hear any part of the dialogue they wish on the SPARC Station 10 with a headphone. Subjects performed four sessions. From the first to the third session, subjects transcribed the training dialogue. In the fourth session, subjects transcribed the test dialogue.

Subjects use the Hiragana (Japanese syllable oriented characters) to transcribe both verbal sounds and nonverbal sounds, e.g., laugh. The TEI tags and entities were used to transcribe nonverbal features [3]. All of subjects attended to meetings, to learn about TEI notation and how to transcribe the dialogue, before first session. We had meetings at the end of each session. At the meeting, we answered questions from subject about criteria for transcription.

3.2.5 Labels Used

Known labels for suprasegmental features, ToBI is one for prosody [4]. There is Japanized version J-ToBI also. These ToBI's intends to label pitch raise/lower within a word or a phrase. The descriptions of ToBI are defined for local pitch variation and local feature pattern based descriptions, therefore inappropriate for paralanguage, which bears global features. TEI (Text Encoding Initiative) advocates various information of utterances including paralinguistic features, kinesics, and wide ranges of information allowing free description of a transcriber [3]. We adopted some of TEI labels shown in Table 1, which does not cover all of TEI labels.

A subject is asked to place a label at the beginning of a clause to make it easy to postprocess comparison between different labelers. But this request is not

mandatory.

Table 1. Adopted TEI paralinguistic labels.

p/+, p/++	Tone is raised, very much raised.
p/-, p/--	Tone is lowered, very much lowered.
l/+, l/++	Voice becomes loud, very much loud.
l/-, l/--	Voice lowered, very much lowered.
t/+, t/++	Tempo up, very much.
t/-, t/--	Tempo down, very much.

3.2.6 Labeling Example

Table 2 shows labeling example. The left most number is a time point in second within the speech sample. The labeling program automatically gives time stamp. The second column is an utterance label, and the right most column is the paralanguage.

Table 2. Labeling example.

Time	Utterance	Paralanguage
1.040	<bs>shi&t</bs>pau/s	
1.728	<bs>yama	l/-,t/+,p/-
2.264	<bs>ko-mo-	
2.732	<bs>aredana-	p/+
3.346	</bs>pau/s,<bs>tozo-<bs>N	
4.211	<bs>ji-saNno	l/+,p/+
4.681	<bs>are</bs>pau/m	

3.2.7 Labelers

Subjects involved in this experiment are three including one of authors of this paper. Two of them had no experience of labeling, so they practiced training and meeting for three times. Three times of training is sufficient to get used to this sort of work from our previous experience. Subjects can learn how to work and produce consistent and stable outputs.

3.3 Consistency of Labels

3.3.1 Crosscheck between Labelers

In this paper, we evaluated errors and inconsistencies of transcriptions based on crosscheck between transcribers. In spontaneous dialogue, the correct transcription does not exist, because dialogues do not have any scripts [5]. Number of labels transcribed by subjects is 1008. About 40% of labels are coincided for all subjects, 50% of labels are scattering but never conflicting. Whether to write or not to write labels is arbitrary for subjects, then some is labeled and some is skipped. There are about 10% conflicting transcription such as pitch raised/depressed, loudness increasing/decreasing, speech rate accelerated/braking.

3.3.2 Review of Other Labeler's Labels

There is uncertainty of labels for labeler because this labeling process is free and no control. A labeler can arbitrary mark a label at some point or skip that point.

After finishing labeling, a labeler reviewed other labeler's labels while listening to the speech materials to identify whether they can agree with the other labeler's labels or not. We asked subjects to judge for the labels those they labeled but the other does not labeled, or those other labeler labeled but they themselves did not labeled. The result as shown in Table 3, 60% to 70% of labels are agreed. This means that if plurality labelers worked on the same speech material and crosschecked other labeler's labels then consistency of labels can be increased.

Cross check test of conflicting labels (about 10%of labels) between labelers, a labeler insist on his labels and only 10% to 40% agreed other labeler's label. This means some case a labeler had made mistake, but most case is subjective judgement from different view of labelers.

Table 3. Percent of agreement after cross review.

	Pitch	Loudness	Speed
Mark or not	64.8	76.3	59.9
Collision	44.4	33.3	14.3

4. PHYSICAL MEASURE OF LABELS

4.1 Physical Measure of Voice Characteristics

We have observed restricted aspects of paralanguage. These are intrinsically correlated known acoustic parameters.

4.1.1 Speech Tone Height

Tone raise/fall is psychologically correlated to pitch or fundamental frequency averaged for certain interval to eliminate prosodic fluctuations.

4.1.2 Loudness

Loudness of voice is measured with average power rms. (root mean square) Averaging is necessary to eliminate phonological variations.

4.1.3 Speech Rate

Speech rate is measured in mora per second in Japanese, that is number of mora uttered with in a second. In microscopic view, an instantaneous speech rate is an inverse of a mora duration. In practice, this instantaneous measure is very variable; therefore we needs to average individual mora rate within a suitable interval.

4.2 Psychophysical Measure

One of possible hypothesis is that when a labeler marks a label, the subject noticed significant change in acoustic domain. In prosody, the maximum value of fundamental frequency and magnitude of accentual command are employed. Here we for paralanguage, assume to have to observe longer period to time, i.e. peak values in a clause or a utterance are not necessarily a suitable values since these are parameters for prosody.

We investigated suitable observation time window sizes

and suitable physical measures. Here "suitable" we mean that physical measure and subjective labels matches best.

4.2.1 Window Size

The most suitable window size is shown in the following Table.

Table 4. Observation window size.

Feature	Window size	Hitting rate
Height	300 ms.	68.9 %
Loudness	500 ms.	82.1 %
Speed	400 ms	74.3 %

4.2.2 Differential Measure

There may be various possibilities, but here we adopted differential score as psychophysical measure. Common usage for sensuous measure of noticeable difference is log-scaled ratio measure: *20*log(current value/previous value).*

4.2.3 Hitting Score

Psychophysical measure is evaluated for correspondence with paralinguistic labels. Increasing measure should correspond to raised voice tone, louder voice, and faster utterance. While decreasing measure should correspond to lowered voice tone, lowered voice, and slow down utterance. In case of these matching, we count as a hit, and in case of opposite matching, we count as a blank.

5. RESULTS AND DISCUSSION

5.1 Consistency between Labelers

One of reliability measure of the label is that different labelers put the same label at the same point. Table 5 shows how many labelers among three put the same label at that point and stable paralinguistic feature is detected.

Table 5. Consensus and psychophysical hitting.

(a)Voice tone height.

Coincidence	# of Labels	% of Hitting
sole label	159	65.6
2 of 3 labeled	130	76.6
3 of 3 labeled	12	75.0
Conflicting labels	47	51.1

(b) Voice intensity or loudness.

Coincidence	# of Labels	% of Hitting
sole label	122	74.6
2 of 3 labeled	90	86.4
3 of 3 labeled	63	100.0
Conflicting labels	16	50.0

(c) Rate of talking.

Coincidence	# of Labels	% of Hitting
sole label	118	69.2
2 of 3 labeled	86	83.3
3 of 3 labeled	54	94.4
Conflicting labels	40	52.5

Larger changes showed higher possibility to be noticed and labeled by transcribers, and also showed higher possibility to be labeled as coincident labels. However larger change is not necessary sufficient cue for paralinguistic features: at the largest change 100% of tempo, 60% of loudness and 30% of pitch height were perceived and labeled.

5.2 Evaluation of Psychophysical Measure

In what extent psychophysical measure and subjective labels matched evaluates the psychophysical measure we proposed here. As shown in Table 5 along the right most column, the psychophysical measure and labelers' judge have hit. Obviously, conflicting labels are half to half for of three psychophysical measures. While where labels are in agreements, hitting rates are significantly higher as the number of agreeable coincident labels increases. Therefore the proposed psychophysical measure reflects paralinguistic features.

6. CONCLUSION

We investigated the relation of paralinguistic labels and physical quantity, that is voice tone height is correlated with fundamental frequency, loudness is correlated with short-term power, and rate of talking is correlated with number of moras within a second. The physical changes are represented differentially as a contrast between proceeding 400 ms quantity and succeeding 400 ms quantity. About 80 % of labels coincided with physical changes. Whether subject hear paralinguistic event in a speech is partially dependent on the magnitude of differential change of physical measure, however, our results also suggest there are unknown factors that may affect to perception of paralinguistic features.

7. REFERENCES

[1] G. L. Trager (1958), "Paralanguage: A First Approximation", *Studies in Linguistics*, 13, Univ. of Buffalo, pp. 1-12.

[2] H. Fujisaki (1996), "Prosody, Models, and Spontaneous Speech", *Computing Prosody*, Springer, pp. 27-42.

[3] C. M. Sperberg-McQueen, L. Burnard (eds.) (1994), "Base Tag Set for Transcription of Spoken Texts", *TEI P3*, Chapter 11, Text Encoding Initiative, Chicago, Oxford.

[4] K. Silverman, M. Beckman, J. Pirelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg (1992), "ToBI: A Standard for Labeling English Prosody", Proc. OCSLP, 2, pp.867-870.

[5] S. Kobayashi and S. Kitazawa (1995), "Consistency of Inter-Transcribers' Transcription", *Proceedings of the European Conference on Speech Communication and Technology*, pp. 1263-1266.