Acoustic Analysis and Transcription of Linguistic and Paralinguistic Features in Dialogue Speech

Shigeyoshi Kitazawa*, Satoshi Kobayashi†,
Takao Matsunaga‡, Hideya Ichikawa‡ and Junichi Nishiyama‡

Abstract

Nonverbal features of spontaneous speech are important in human dialogue. And transcription of dialogue including the paralanguage is a useful technique for analysis of the dialogues. But, the paralinguistic features are transcribed by transcriber's subjective now. Some features are measurable as concrete acoustical features. We made measurement of the speech rate through spectrogram transformation of the wave envelope and automatic estimation of such features is helpful in transcribing the dialogues. In this paper, we examine an automatic method of speech rate estimation and studied consistency of descriptions between different transcribers.

1 Introduction

Nonverbal communications play an important role in human dialogue where participants use natural speech so called spontaneous speech. That nonverbal information including the vocalizations and the voice qualities is called as the "paralanguage"[1]. There are number of aspects of paralanguage. Some of them are sensational features, and difficult to realize measurements. The others, however, are measurable as concrete acoustical features such as loudness, pitch, and speech rate. Here we described how to measure the speech rate. Another aspect of dialogue research is the transcription into text of the paralinguistic features such as voice loudness, tone, and speech rate as well as utterances. We made various measurements of validity and consistency of descriptions between different transcribers.

2 Definition of Speech Rate and Measurement Method

A mora is a prosodic term that is a conjunction of a consonant and a short vowel. A mora is another phoneme or a phoneme conjunction of the equivalent length. A mora consists a structure of /V/, /vV/,/CV/,/CvV/,/N/, and /Q/ (V: a vowel, v: a semivowel, and C: a

*Shigeyoshi Kitazawa (北澤 茂良): Professor, Department of Computer Science, Faculty of Engineering, Shizuoka University

†Satoshi Kobayashi (小林 聡): Doctor Cource, The Graduate School of Electronic Science and Technology, Shizuoka University

‡Takao Matsunaga (松永 隆雄), Hideya Ichikawa (市川 英哉), Junichi Nishiyama (西山 淳一): Master Cource, Department of Computer Science, Faculty of Engineering, Shizuoka University

consonant). In studying Japanese phoneme, a mora is a more realistic and convenient unit than the segmental phoneme. The speech rate is the number of moras per second.

One technical measurement of the rate of speech is achievable through the phoneme recognition, that is, the point of time of each phoneme is marked along the time line and the resulting phoneme (hence the mora) lengths are averaged along some intervals resulting a number of moras per second. This definition, however, is difficult to realize with the automatic speech recognition technology, but possible only by hand labeling. There are several possible hypotheses of perception of the rate of speech. We estimate the tempo without identification of the phoneme and do not exploit speech recognition, hence we assume we can recognize the rate of speech without recognizing the content of speech. Furthermore we can perceive, for example, the utterance speed from the narrow band filtered speech sound. This fact suggests that our sense of a tempo can be perceivable from the envelope of the waveform. In the preceding study, the rhythm is correlated to the interval of the center of energy between adjacent syllables[2].

Because Japanese has the CV-syllable-timed feature, downswings of the envelope appear at every consonant segment almost at the same interval. A speech envelope changes dynamically from a consonant to a vowel and then to the next consonant forming the peaks and the valleys. The intervals between peaks and valleys are expected to be approximately equal or periodic because of the syllable-timed feature. If this is true, then we can extract this periodicity through the following procedure: the DFT (Discrete Fourier Transform) of the Hamming windowed envelope pattern. We employed window size about one second through our experiment. The window includes local pauses and non-lexical voicing due to non-verbal or paralinguistic expressions.

The speaking rate is observed as a dominant spectral peak in a frequency domain, where the speaking rate is visualized in the frequency-time plane like the formantic pattern of spectrograms. The frequency and the time are scaled downward to one two hundredth of the 8 kHz sampling rate of the normal spectrogram. We could observe gray gauged monochrome patterns in the 20 Hz frequency region with one second of time window.

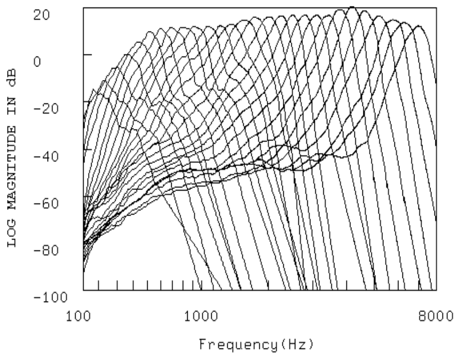


Figure 1: 28 channel auditory filter.

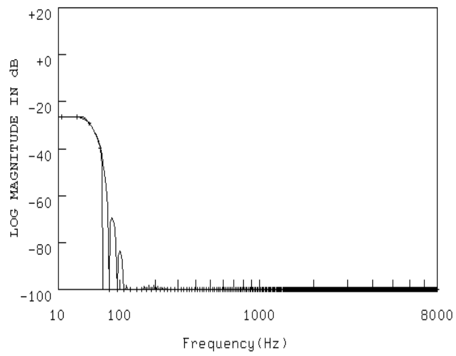


Figure 2: 80Hz FIR low pass filter.

We employed the bandpass filtered speech with the auditory model for the source of the speech envelope. The characteristic of the auditory filter is shown in Fig¡ure 1. In order to obtain an envelope of the speech waveform, we first rectified the wave to obtain a half-wave, on which then we lowpass-filtered to obtain an approximate envelope. We designed a low pass filter of the cutoff frequency at 80 Hz to keep the envelope details and to eliminate high frequency fluctuations due to pitch and formant frequencies as well. This filter deals about ten times of the average rate of spontaneous speech, 8 mora per second. The characteristic of the low pass filter is shown in Figure 2. In both the speech waves of bandpass filtered and fullband, we could observe in the spectrogram of the wave envelope such concentrated spectral energy around the frequency corresponding to the speech rate.

3 Speech Rate Estimation on Synthesized Waveform

We examined our method with synthesized envelope waveforms as a preliminary experiment. We use four envelope waves as follows. All of them are half wave sinusoids rectified to simulate the envelope of vowel segments. The half waves are arranged so as to be proportional to the inverse of the speech rate or arranged along the corresponding to the period as long as 2 second.

A: Stationary period with a short consonantal gap between the adjacent half waves.

B: Gradually decreasing period by decreasing the gaps of A.

C: Stationary period. But frequencies of a half wave increase gradually, therefore the gaps increase gradually.

D: Adjacent half waves make an overlap. And the period gradually decreases, hence overlap increases while the shape of half wave sinusoids is kept.

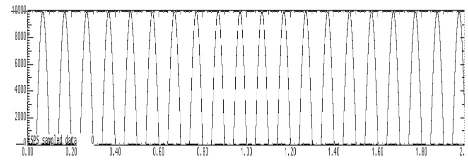
Figure 3:A∼D show these envelopes and Figure 4:A∼D show the spectrograms. In Figure 4, we could observe a dark bar corresponds to the speech rate. From Figure 3 and Figure 4, this method could catch wide ranges of periodicity, therefore this will estimate speech rates based on the intervals between peaks of the envelope. The result was independent of the duty ratio of the half wave sinusoid, i.e., the peak and the dip ratio.

In the end of this preliminary experiment, we use a synthesized wave based on the Japanese rhythm rule [2]. Envelope and spectrogram are in Figure 5and 6. Figure 7 shows manually calculated speech rate of Figure 5. The spectrogram of it shows a more complicated texture than the simple test signals in Figure 3 and 4.

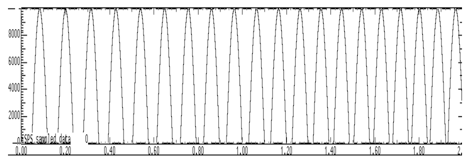
4 Speech Rate Estimation on Real Dialogues

Then we examined real dialogue speeches taken from TV programs. The test set has 13 utterances. Seven utterances were spoken by a male speaker, six utterances were spoken by a female speaker. The sampling frequency is 8000Hz and the durations are three to seven seconds. The real speech rates of the test set are measured by hand labeling of individual phonemes. Contents of the test set utterances are on Table 1. And an overview of automatic estimation process is shown in Figure 8. We rectified the wave to obtain a half-wave at first(b). Then, we lowpass-filtered to obtain an approximate envelope(c). At last, we compute DFTs of extracted envelope(d).

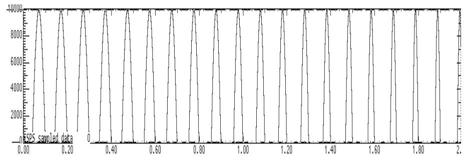
We employed the bandpass filtered speech with the auditory model for the source of the speech envelope. Cut off frequency of the low pass filter is 40Hz.



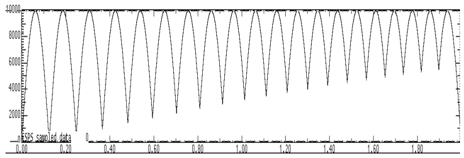
A: Stationary period.



B: Decreasing period.

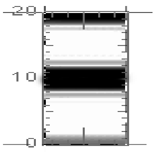


C: Increasing silent gaps.

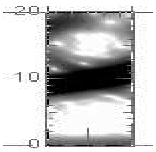


D: Decreasing period and dips.

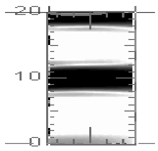
Figure 3: Synthesized envelopes. (each wave continues 2 seconds)



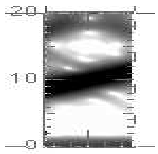
A: Stationary period.



B: Decreasing period.



C: Increasing silent gaps.



D: Decreasing period and dips.

Figure 4: Spectrograms of synthesized envelopes in the range of 20Hz and 2 seconds.

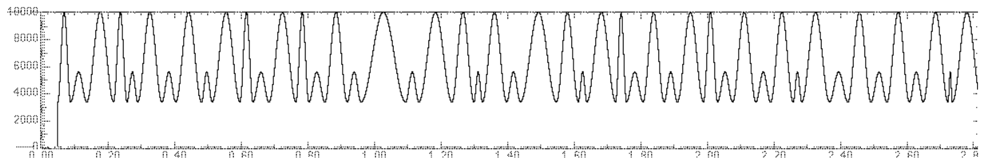


Figure 5: Rule-based synthesized envelope in 2.8 seconds.

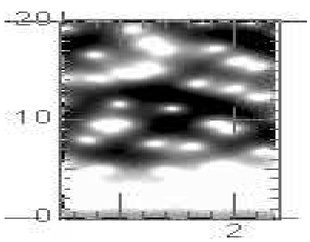


Figure 6: Spectrogram of the rule-based synthesized envelope (Figure 5) in the range of 20Hz and 0.5~2.4 seconds.

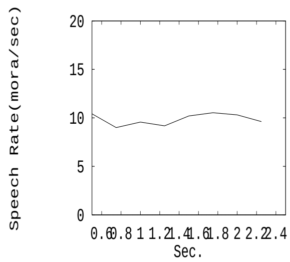


Figure 7: Period of synthesis data(Figure 5).

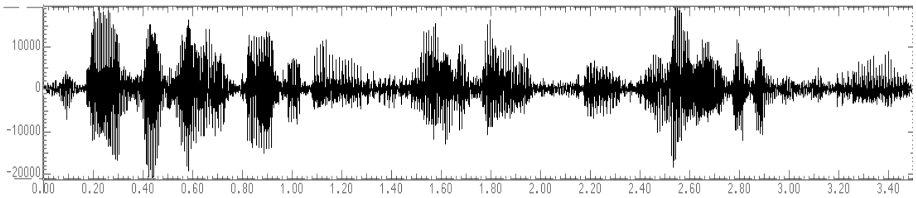
Table 1: Utterances of the test set.

M1	itai dokogaitaitoka nandakandaiwanai katadesitakara
M2	soudesuneh anoh nagaisihnnankadatoh zuttomottenakyanannaindeh
M3	owarinohouni naruto tega sibiretekityaundesuyoh
M4	eh demoitijikan itijikanguraidesukaramah nijikanwa sitijika osokutomo hatijiguraidesune
M5	ndeh kujiguraikara norihajimeteh nijiguraimade desukaneh
M6	notte desaigo mata tyokotto umano teiresite kaerun
M7	nitiyoubiga yasumikato omottandesukedoneh umanokeikoga arimasite tyoubaga
F1	ma imawa gendaiteki gendaigekio yaritaimonodato omotterassyarusoude gozaimasuga
F2	demo sonosanninno ohdisyonno hokanohitotatimo minna neoagezuni soredemo
F3	utini kityattandesuyoneh hagakimitainamonde sorede nangatu nannitini
	oatumarikudasaitteittara moukonaihitoga iruwakejanai
F4	uhn dakara souyunotte sugoku yokuwakarujan annanikataku oyakusokusitanonitteh
F5	ma yoru gohanga owatte otyato nitibutte yukotonandesukedo sokode mata anosehza
F6	taihendesu sorede oyasumiga itinitimonaitte yukotowa nitiyoubiha

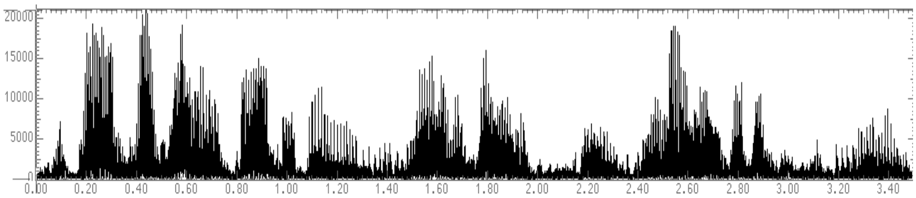
Table 2: Measurements of utterances in Table1.

	A	B	C	D	E	F	G	manual computed			DFT computed(Hz)		
	sec.			dips		%	rate	head	mid.	tail	head	mid.	tail
M1	3.505	48	28	50	–2	4.2	8.0	12.0	8.8	9.6	6.7	11.5	8.6
M2	4.000	48	34	54	–6	12.5	8.5	8.4	6.9	13.9	6.8	8.7	10.7
M3	1.685	28	23	24	4	14.3	13.6	15.6	13.4	17.1	11.4	12.3	15.1
M4	4.977	69	44	61	8	11.6	8.8	11.9	8.0	12.9	10.3	7.6	10.4
M5	4.161	47	30	57	–10	21.3	7.2	6.8	9.0	13.5	6.4	8.2	8.3
M6	3.537	39	25	57	–18	46.2	7.1	6.9	7.2	9.6	5.1	6.0	6.1
M7	4.257	55	38	48	7	12.7	8.9	9.3	9.5	9.0	7.1	8.9	8.0
avg.	3.731	47.7	31.7	50.1	9.2*	17.5	8.9	10.1	9.0	12.2	7.7	9.0	9.6
F1	4.257	62	42	58	4	6.5	9.9	10.6	12.3	11.9	8.5	9.4	9.4
F2	3.697	66	36	53	13	19.7	9.7	12.1	11.1	9.7	10.5	8.6	8.6
F3	6.929	102	65	94	8	7.8	9.4	11.6	12.0	14.4	11.6	7.8	9.6
F4	4.273	62	41	56	6	9.7	9.6	7.3	10.4	9.5	8.5	7.6	10.4
F5	5.521	67	39	81	–14	20.9	7.1	8.9	8.9	6.1	8.8	9.7	7.8
F6	3.665	54	35	49	5	9.3	9.5	6.0	11.5	11.6	11.9	10.0	7.1
avg.	4.724	68.8	43.0	65	9.2*	12.3	9.2	9.4	11.0	10.5	10.0	8.9	8.8

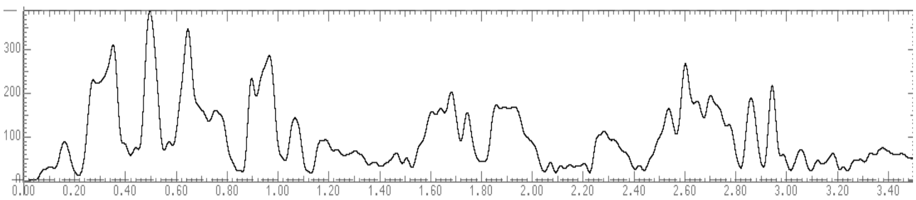
A:duration, B:phonemes, C:moras, D:dips detected,
E=B–D:difference, F=$|E|/B \times 100$:errors[%], G=C / A:global speech rate in mora/sec.,
manual computed:manually estimated speech rate,
DFT computed:computationally estimated speech rate,
head:speech rate in beginning 1 sec., mid.:speech rate in the middle 1 sec.,
tail:speech rate in 1 sec. at the end,
$*$:standard deviation of differences.



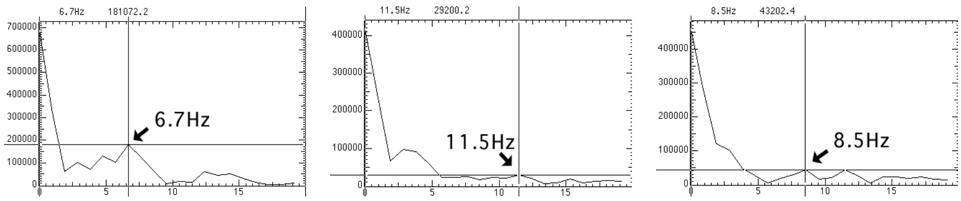
a: Original sound wave of the utterance M1.



b: Rectified wave.



c: Low pass filtered envelope.



d: Head DFT. d: Middle DFT. d: Tail DFT.

Figure 8: Process of speech rate estimation proceeds from the top(a) downward to obtain the DFT peaks.

We computed DFTs of the envelopes to find the frequency of peak energy as an estimated speech rate with one second hamming window.

The speech rate computed as the mora per second fluctuates due to paralinguistic features. The speech rate estimated as the inverse of period between the center of power of the adjacent vowels is dependent of pauses and nonlexical items in the dialogue. We compared the local speech rate by hand labeling and computing.

Table 2 shows manually estimated speech rates and computed speech rates around "head", "middle" and "tail" of data. The manual estimation and the DFT estimation correlated with coefficient 0.55. Figure 9 shows this correlation.

Figure 10 and 11 are envelopes and spectrograms of the real speech taken from Table 1. We can see traces of continuous changes of speech rate in the spectrograms, however, the traces are not unique for each spectrogram due to aliasing of the harmonic components. The spectrograms show more complicated texture than the test signals in Figure 4, and look similar to Figure 6.

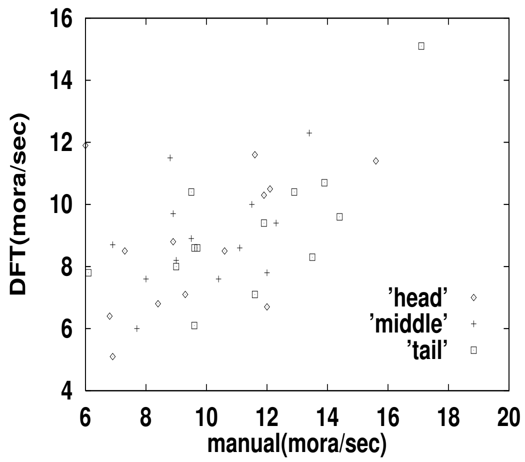
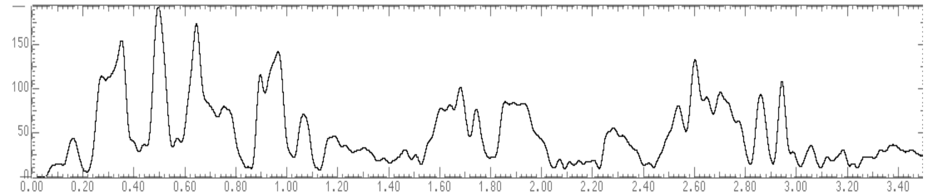


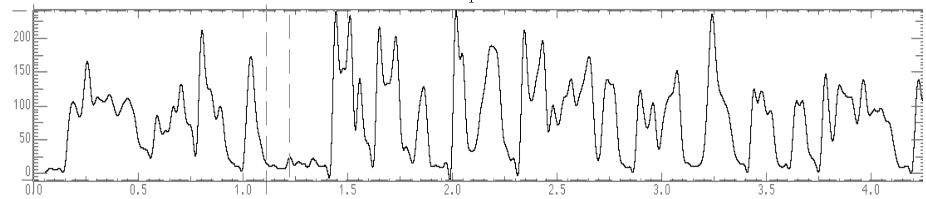
Figure 9: Correlation between manual and DFT computed speech rates.

It was difficult to recognize the speech rate as a unique dark bar pattern.

Figure 12 shows spectrograms of auditory filtered results. Some channels are show better result than the non-filtered one.

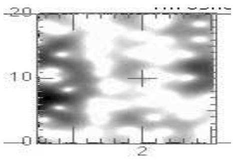


A: Envelope of M1.

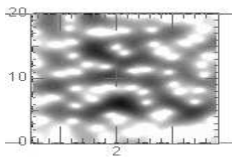


B: Envelope of F1.

Figure 10: Envelopes of the utterance drawn from the test set in Table 1.

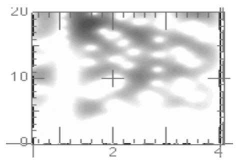


A: Spectrogram of M1.

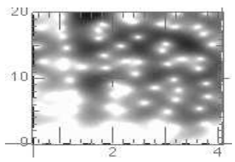


B: Spectrogram of F1.

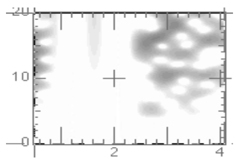
Figure 11: Spectrograms of real speech.



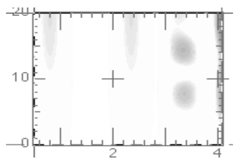
Channel 3



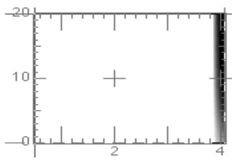
Channel 8



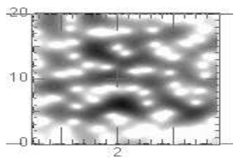
Channel 11



Channel 16



Channel 21



Not Filtered

Figure 12: Spectrograms of outputs from the auditory filter (F1).

5 Inter-transcribers consistency of transcription

Text encoded dialogue is a useful form to analyze dialogue phenomena. However, paralinguistic features are classified by transcribers' subjective. Inconsistencies between transcribers, omissions and misleading, are unavoidable. Such inconsistency is the problem for analysis of dialogues. We estimated those errors and inconsistencies based on the cross-check between transcribers.

We use a real speech data that has about 40 seconds' durations from a TV program of interviewing between a female interviewer and a male interviewee. In the program, speakers talked freely each other. Subjects heard the recorded dialogue on SPARC Station 10 with replay program and headphone. The program can replay designated part by the user.

Subjects use the Hiragana (Japanese syllable oriented characters) to transcribe verbal sounds and nonverbal sounds, e.g., laugh. The TEI notations transcribe nonverbal features [3]. Subjects are five people. They attended to meetings to learn about the transcription, but they didn't have any practices.

We used the following formula to estimate consistency rates.

$$R = \frac{a+b}{A+B} * 100[\%]$$

A: Number of transcription items (Hiraganas, Tags, and Entities) occurred in one of transcriptions.

B: Number of transcription items in another transcription to be compared with A.

a: Number of matched items in transcription A.

b: Number of matched items in transcription B.

(a = b)

6 Results

6.1 Consistency of inter-transcribers

Table 3 shows consistency rates. In Table 3, "Char" means Hiragana for verbal sounds, and "Tag" means nonverbal features, i.e., Tags and Entities. Criteria for each consistency estimation are followings:

A: Count only matched transcription units.

B: Criterion A or the same contents transcribed around the near position in are included.

C: Criterion A and the similar contents transcribed at the same position are included.

D: Criteria A, B, and C are included.

E: Criterion D and the similar contents transcribed around the near position are included.

Here, "similar" means

1. Differences in length notation, e.g., "e-"("-" represents a long vowel) and "ee" are similar.
2. A transcriber remarked the sound is not definite but can be heard to other sound. There are suggestions that match in both transcriptions.
3. Differences in degree of change of nonverbal features, e.g., in case of tempos, "aa" and "a" are similar.

Table 3: Consistency rates.

Criterion:A

	Consistency(%)	Inconsistency(%)	TOTAL
ALL	6,698(69.9%)	2,891(30.1%)	9,589
Char	5,750(86.8%)	877(13.2%)	6,627
Tag	948(32.0%)	2,014(68.0%)	2,962

Criterion:B

	Consistency(%)	Inconsistency(%)	TOTAL
ALL	6,874(71.7%)	2,715(28.3%)	9,589
Char	5,750(86.8%)	877(13.2%)	6,627
Tag	1,124(37.9%)	1,838(62.0%)	2,962

Criterion:C

	Consistency(%)	Inconsistency(%)	TOTAL
ALL	7,294(76.1%)	2,295(23.9%)	9,589
Char	6,130(92.5%)	497(7.5%)	6,627
Tag	1,164(39.3%)	1,798(60.7%)	2,962

Criterion:D

	Consistency(%)	Inconsistency(%)	TOTAL
ALL	7,470(77.4%)	2,119(22.1%)	9,589
Char	6,130(92.5%)	497(7.5%)	6,627
Tag	1,340(45.2%)	1,622(54.8%)	2,962

Criterion:E

	Consistency(%)	Inconsistency(%)	TOTAL
ALL	7,627(79.5%)	1,962(20.5%)	9,589
Char	6,130(92.5%)	497(7.5%)	6,627
Tag	1,497(50.5%)	1,465(49.5%)	2,962

Table 4: Inclusion of corresponded nonverbal features by loosing the criteria.

Criterion	Corresponded(%)
Criterion B	176(5.9)
Criterion C	216(7.3)
Criterion E	157(5.3)
Total	549(18.5)

Table 5: Inconsistencies of sound.

Contents	Inconsistencies(%)
Long vowel	
at the end of word	70(14.1%)
in interjection	26(5.2%)
long vowel in the word	42(8.5%)
Overlap	
Misleading	110(22.1%)
Double consonant	
at the end of word	27(5.4%)
in interjection	6(1.2%)
double consonant in the word	7(1.4%)
Laugh	80(16.1%)
Others	129(26.0%)
Total	497(100%)

In Table 3A, 87% of sounds in Hiragana corresponded. But, only 32% of Tags and Entities corresponded. In Table 3E, the loosest criteria, 92% of sounds in Hiragana corresponded, and 50% of nonverbal features corresponded.

These results say transcriptions of nonverbal features are affected by transcribers' subjective.

Table 4 shows how many nonverbal features correspond newly included in each new criterion (B, C, E), and those percentages against all nonverbal features.

6.2 Analysis of inconsistencies

Table 5 shows inconsistencies of transcribed sound with Hiragana in Table 3E. "%" means each percentage against whole sounds inconsistencies in Table 3E. Here, 36% are caused by differences in sound length, long vowels and double consonants. Many of other inconsistencies are caused by misleading. However, there are sounds that are difficult to transcribe in Hiragana, e.g., laugh.

In previous section, we see the problem is transcriptions of nonverbal features. Almost inconsistency of nonverbal features consists of omissions and insertions. Those are 1,436 items against 1,465 items of inconsistency of nonverbal features. The 29 remainders are contradicted items between transcriptions. Table 6 shows contents of omissions and insertions. Here, upper elements are paralinguistic features and lower elements are conversational analysis information. "%1" means percentages against number of each nonverbal feature.

Table 6: Omissions and insertions at nonverbal features.

	Count	Omissions & Insertions(%1:%2)
Pitch	734	461(62.8%:32.1%)
Voice Quality	397	277(69.8%:19.3%)
Loudness	424	259(61.1%:18.0%)
Speech Rate	204	142(69.6%: 9.9%)
Nonverbal Sounds	130	49(37.7%: 3.4%)
Others	132	89(67.4%: 6.2%)
Overlap	348	78(22.4%: 5.4%)
Pause	331	73(22.0%: 5.1%)
Speaker Exchange	262	8(3.0%: 0.6%)
TOTAL	2962	1436(48.5%:100%)

"%2" means percentage against all of omissions and insertions.

Paralinguistic features, those are pitch, voice quality, loudness and speech rate, marked high inconsistency rates. However, utterance overlaps, pauses and speaker exchanges, these are conversational analysis information, have lower inconsistency rate than paralinguistic ones. Those features, unlike paralinguistic features, don't have any degree. Those are judged by transcribers' subjective.

In case of prosody transcription, 80% consistency was achieved [4]. The research used a graphical interface, e.g., a speech waveform. In this paper, we could see paralinguistic features had high inconsistency rate. Those features, however, are measurable as concrete acoustical features. This means that those features could be estimated automatically, like in earlier part of this paper and [5]. Therefore, if we present such information graphically to transcribers, then consistency rates could become higher than in this paper.

Another way, which will make much better consistency, is to become expert and to verify transcription. We made a program to verify TEI based transcription. It interprets some TEI Tag and Entities, then makes simple sinusoid wave sounds. Interpretable Tags and Entities are shown at Table 7. Speakers are transcribed with 'u'-tag and represented with pitch of sounds. High pitch is assigned to a speaker, and low pitch is assigned to another. Overlaps are transcribed with 'anchor'-tag and also interpretable, the program can output overlapped sound. 'Inhale' and 'exhale' are transcribed with 'vocal'-tag and are represented as pause now. Hence, transcribers can hear transcribed paralinguistic features and can verify own transcription.

7 Conclusion

In preliminary experiment of speech rate estimation, the results show availability of the method in this paper to estimate speech rates automatically. However, in the experiment with real speech data showed complicated spectrograms. We will research how to track speech rates from a complicated spectrogram.

In experience of consistency of inter-transcribers' transcription, we got a consistency rate of over 80% to verbal sound. Nonverbal features, however, are affected by transcribers' subjective. Then, there are many omissions and insertions. We will examine changes of

Table 7: Interpretable Tags & Entities by Verification Program.

Tags	means
anchor	overlap
pause	pause
shift	changes of loud,pitch and tempo
u	utterance
vocal	inhale,exhole

Entities	means
&stress;	stress
&trunc;	truncated syllable
&lf;	low fall intonation
&lr;	low rise intonation
&fr;	fall rise intonation
&rf;	rise fall intonation

consistency rate cause by transcribers' skill, a graphical interface, and a verifying program.

References

[1] George L. Trager: "Paralanguage: A First Approximation", Studies in Linguistics, 13, 1-11, 1958.

[2] Masayo Katoh, Mitsuo Komura, Shin'ichiro Hashimoto: "Rhythm rule in Japanese based on the center of energy gravity of vowels", The Journal of the Acoustical Society of Japan, pp. 888-896, vol. 50 no. 11, 1994.

[3] C.M.Sperberg-McQueen, Lou Burnard: "Base Tag Set for Transcription of Spoken Texts", TEI P3 chapter 11, Text Encoding Initiative, Chicago, Oxford, 1994.

[4] Kim Silverman, Mary Beckman, John Pirelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, Julia Hirschberg: "TOBI: A Standard for Labeling English Prosody", Proc. of ICSLP 1992 pp. 867–870.

[5] Shigeyoshi Kitazawa, Satoshi Kobayashi, Takao Matsunaga, Hideya Ichikawa, "Tempo Estimation by Envelope for Recognition of Paralinguistic Features in Spontaneous Speech", Proc. of ICSLP 1994 pp. 1691–1694.