CORPUS OF SPONTANEOUS JAPANESE: ITS DESIGN AND EVALUATION

Kikuo Maekawa

The National Institute for Japanese Language, Dept. Language Research, Tokyo, Japan
kikuo@kokken.go.jp

ABSTRACT

Corpus of Spontaneous Japanese, or CSJ, is a large-scale database of spontaneous Japanese. It contains speech signal and transcription of about 7 million words along with various annotations like POS and phonetic labels. After describing its design issues, preliminary evaluation of the CSJ was presented. The results suggest strongly the usefulness of the CSJ as the resource for the study of spontaneous speech.

1. INTRODUCTION

Recently, there is a growing consensus among linguists and speech scientists that spontaneous speech (SS hereafter) should be one of the most important research targets in the coming decades. Also, it is a part of the consensus that the study of SS should be based, at least partly, upon the analysis of reliable corpus. This is necessary because SS is inherently much more complex and diverse than the read-speech.

Corpus of Spontaneous Japanese, or CSJ, is a large-scale database of spontaneous Japanese that has been developed by the collaboration of National Institute for Japanese Language (NIJLA) and Communications Research Laboratory (CRL) as a part of the *Spontaneous Speech: Corpus and Processing Technology* project (1999-2003) supported by the Ministry of Education and Science. Project supervisor is professor Sadaoki Furui of Tokyo Institute of Technology [1].

2. DESIGN ISSUES

2.1 Aims

The aim of the CSJ is two-fold: CSJ is the resource for the study of speech recognition on the one hand, and linguistic and natural language processing on the other. An important problem we encountered at the beginning of the project was the trade-off between the requirements of speech recognition and that of linguists. Broadly speaking, speech recognition puts more emphasis on the quantity of data rather than the quality, while linguists put more emphasis on the quality and richness of annotation. To throw oil on this troubled water, we designed a corpus that had dual structure.

2.2 Duality of the corpus structure

The CSJ was designed to contain spontaneous speech of at least 7 million words. We considered this to be a very minimum size to meet the requirement of speech engineers to construct a workable language model for SS recognition.

At the same time, there is a true subset of the CSJ that we call the *Core*, whose size was designed to be 500k words. And this is the *Core* to which we concentrated the cost of annotation. In addition to transcription and POS analysis, annotations described in 2.4.4 and 2.4.5 below were provided for the *Core*.

2.3 Sources

The speech recorded for CSJ is so-called common, or standard, Japanese, a variety shared widely by educated people and used in more or less public circumstances. Speakers who had clear dialectal features in his/her morphology or segmental phonology were excluded.

There are two main sources of SS for the CSJ: ‘academic presentation speech’ (APS) and ‘simulated public speech’ (SPS). APS is the live recording of academic presentations in 9 different academic societies covering the fields of engineering, social science, and humanities. APS is indispensable for CSJ because it is this type of speech that is the most immediate target of the SS recognition studies.

SPS, on the other hand, is studio recording of paid layman speaker’s speech, of about 10-12 minutes, on everyday topics like ‘the most delightful/saddest memory of my life’ presented in front of small audience and in relatively relaxed atmosphere. We tried to balance as much as possible the age and sex of SPS speakers.

SPS is necessary for two reasons. For one thing, SPS is expected to correct skewed distribution of APS in terms of the speakers’ age, sex, and vocabulary, since typical APS is spoken by young male graduate student and characterized by many field-specific technical terms. For another, we planned to introduce difference of speaking style between APS and SPS. Comparison of relatively higher (APS) and lower (SPS) speaking styles will be helpful for the study of linguistic variations, which is one

of the most important areas of linguistic research to which CSJ is expected to be applied (see 4.3 below).

Lastly, though CSJ is basically a corpus of monologue, it contains some dialogue speech for the sake of comparison. CSJ contains 15 hours of dialogue speech consisting of A) Interview with the APS/SPS speakers about the content of their talks, B) Task-oriented dialogue, and, C) Free conversation.

2.4 Annotations

2.4.1 Speech signal

Speech signal was recorded using a head-worn close-talking electric condenser microphone and a DAT, and down-sampled to 16kHz, 16-bit, before being stored in the corpus.

During the recording time, one of the recording staff evaluated the way the speech being recorded was spoken. All speech materials were impressionistically rated with respect to their spontaneity, perceived speaking rate, clearness of articulation, speaking style, and so on.

2.4.2 Two-way transcription

Transcription of SS is perhaps the most important and laborious part of the compilation work.

In CSJ, speech signal was divided into physical 'utterances' at the locations of longer-than 200 ms pause. The resulting 'utterances' were then transcribed using a scheme designed specially for CSJ [2]. The scheme provides two independent transcriptions called 'orthographic' and 'phonetic'. The orthographic transcription is mixture of Kanji (Chinese logograph) and Kana (Japanese syllabary) like ordinary Japanese writing, but it differs from the ordinary writing in that it does not permit any variation, which is a notorious characteristic of Japanese writing system. Phonetic transcription, on the other hand, uses Kana exclusively in order to represent the phonetic details of speech as narrowly as possibly within the limit of syllabary.

Various tags were embedded in these transcriptions to mark SS-specific phenomena like filled-pauses, word-fragment, reduced articulation, mispronunciation, and so forth. In addition to these speech events, non-speech events like laughter and coughing were tagged also.

2.4.3 POS information

Part-of-speech (POS) analysis was applied then to the transcription. Two types of 'word' were recognized reflecting the high degree of freedom in Japanese word-formation. 'Short-unit word (SUW)' approximates dictionary item of ordinary Japanese dictionary, and, 'long-unit word (LUW)' represents various compounds.

Manual POS analysis was applied to a subset of the corpus, and the resulting hand-labeled data was used as the learning data for the development of automatic POS tagging software [3].

2.4.4 Phonetic labeling

In addition to the annotations described so far, segment and intonation labels were provided for the materials in the *Core*. The segment label set was basically phonemic, but some phonetic events were also labeled. The latter included devoicing of vowels, timing of closure release in stops, palatalization of consonants before /i/, and so forth.

Phoneme labels were generated from the phonetic transcription, aligned automatically to speech signals using an HMM-based alignment technique, and adjusted by human labelers [4].

As for intonation labeling, traditional J_ToBI [5] was extended for SS. In the new X-JToBI scheme, inventories of tonal events as well as break indices were enriched considerably [6]. X_JToBI labels, when coupled with the segment labels, could provide rich information about the phonetic variations observed in SS. More importantly, it is expected that the phonetic labels of the *Core* provide opportunity for the study of paralinguistic information like speakers' attitudes and intentions [7,8].

2.4.5 Miscellaneous annotation

Annotations described so far were all in the blueprint of the corpus [1]. Prompt progress of the corpus compilation, however, enabled us to provide more annotations for the *Core*. New annotations included A) discourse structure labeling [9], B) locations of important syntactic boundaries [10], and C) dependency structure analysis of the clause as defined by B above. These annotations are currently underway.

3. STATUS-QUO OF THE CORPUS

Recording of APS and SPS material came to an end in December 2001. By that time, 805 and 590 different speakers provided 1007 APS and 1715 SPS materials respectively.

Currently, the whole CSJ contains speech signal of about 661 hours including 299 hours of APS, 330 hours of SPS, and 32 hours of miscellaneous material including dialogues, reading of short written passages, and the reproduction of transcribed speech by the same speaker. All materials have already been transcribed and the error correction is underway. The total number of SUW in these materials is estimated to be about 7.2 million.

Manual POS analysis has applied to 74 hours of APS and SPS speech and resulted in 884k SUW and 750k LUW. The accuracy of the manual analysis was estimated by random sampling to be about 99.9% for SUW and 97% for LUW. Pilot automatic tagging of SUW was conducted at the CRL and achieved accuracy of about 94% [3]. All transcription text should be POS analyzed by the end of the whole project.

Phonetic labeling of the *Core*, whose size is estimated to be about 44 hours, has been in due course. Segment and intonation labels have been applied to 40 and 22 hours respectively at the time this manuscript is being written. The whole labeling is expected to be done by the end of May 2003.

4. EVALUATIONS

In the rest of this paper, the usefulness of CSJ will be evaluated from a point of view of linguistics. Unless specified otherwise, the 884k word POS-analyzed sub corpus mentioned above was used for evaluation.

4.1 Speaking rate

Speaking rate was compared between the CSJ and the read speech in the ATR speech database (phonemically balanced 503 sentences read professional speakers)[11]. Speaking rate was defined as the number of morae spoken per second, and computed for all pause-separated 'utterances' in the case of CSJ, and for all sentences in the case of ATR database. The averaged speaking rate and the S.D. was 8.01 and 2.07 respectively in the case of CSJ, and 7.11 and 0.96 in the case of ATR. Under the supposition of normal distribution, about 0.1% of CSJ 'utterance' has speaking rate faster than 14.2 mora/sec. This extremely high speaking rate could cause problem in the recognition of SS [12,13]. Figure 1 compares the speaking rate and S.D. of three speech types in the CSJ. APS in general was faster than SPS, but the APS in the engineering fields was faster than that in humanities.

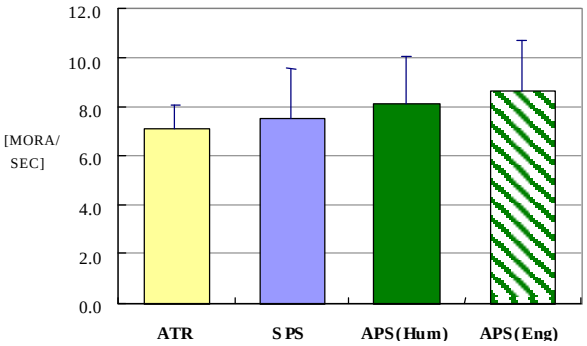

Figure 1. Comparison of speaking rate across speech types of CSJ. Error bar shows standard deviation.

4.2 Disfluency

Disfluency is the most salient feature of SS. Among the tags used in the transcription of CSJ, three tags were concerned deeply with disfluency: Tag (D) marked word fragments, tag (W) marked incorrect or reduced pronunciations, and tag (F) marked filled-pauses [1,2]. Figure 2 shows the ratio of these tags to the total number of SUW in a speech as a function of speech type (APS vs. SPS) and speaker's sex. Males had more disfluency than females regardless of the speech type. At the same time, SPS contained more disfluency than APS, as long as (D) and (W) were concerned.

Figure 3 examined the correlation between the ratio of disfluency and the impressionistic rating of the spontaneity of speech (cf. 2.4.1). Spontaneity was rated using the scale of 1(virtually read) to 5(really spontaneous). Although all (D), (W), and (F) were positively correlated with the perceived spontaneity, the correlation of (F) was not as clear as in (D) and (W).

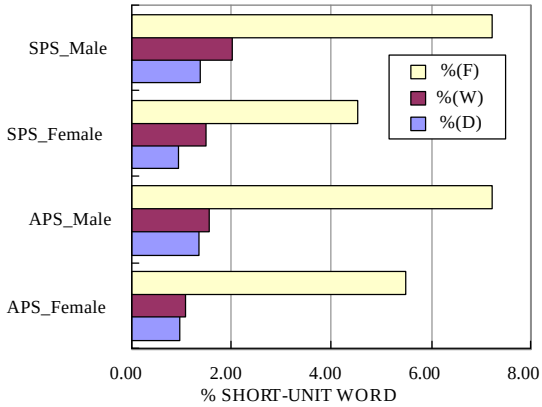

Figure 2. Distribution of disfluency tags in CSJ.

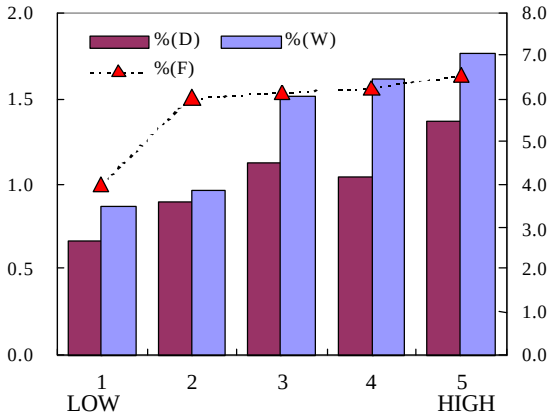

Figure 3. Correlation between the ratio of disfluency tags and the impressionistic rating of spontaneity (abscissa). The left ordinate is for the ratio [%] of (D) and (W), while the right one stands for the ratio of (F).

4.3 Language variations

Study of language variation is one of the most important fields of linguistics to which CSJ is expected to be applied. It is also important for the study of speech recognition and natural language processing, because variations could be serious obstacles to speech and/or language processing.

4.3.1 Vowel devoicing

Phonemic analyses of Japanese vowels describe devoiced vowels as the conditional allophones that occur when /i/ and /u/ are preceded and followed both by voiceless consonants. But it is well known that the devoicing rate of the close vowels does not reach 100%. 427973 vowels involved in 23 hours of segment-labeled *Core* material uttered by 29 females and 56 males were analyzed to examine the variation in vowel voicing [14]. As expected, the devoicing rate of close vowel was 89.2 and 84.3% for /i/ and /u/ respectively. This fact suggested the presence of devoicing-preventing factors.

One such factor was the avoidance of devoicing in the environment of consecutive devoicing, i.e., the sequence of morae each consisting of a voiceless consonant and a close vowel. Figure 4 shows the devoicing rate of the two close vowels in the environment of consecutive devoicing. There was a clear tendency to avoid consecutive devoicing, and, the combination of consonant manners played a crucial role there. When a fricative was involved in the combination, it was always the vowels associated with the fricative that showed higher devoicing rate, and if both consonants were fricatives, the last vowel showed higher devoicing rate.

Although the avoidance of consecutive devoicing had been well acknowledged by previous introspective analyses, the systematic nature of the phenomenon was captured for the first time by the analysis of CSJ.

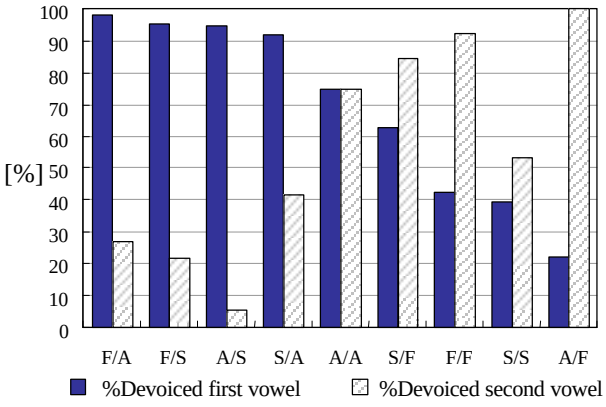

Figure 4. Devoicing rate (ordinate) of two close vowels in the environment of consecutive devoicing. The abscissa is the combination of consonant manners. 'F', 'A', and 'S' stand respectively for fricative, affricate, and stop.

Figure 5 shows the effect of speaking rate on close vowel devoicing. The numbers in the abscissa denote speaker-normalized speaking rate classes; 1 and 4 mean the slowest and fastest 25% of utterances involved in one speaker's speech. Devoicing rate increased monotonically as the function of speaking rate.

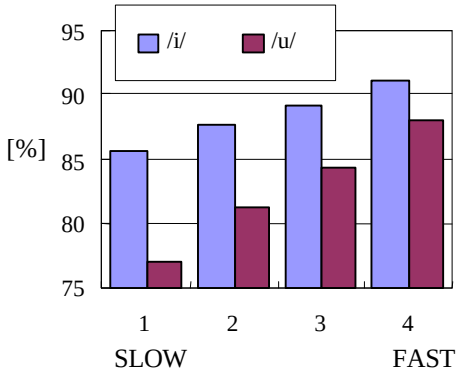

Figure 5. Influence of speaking rate on the rate of close vowel devoicing.

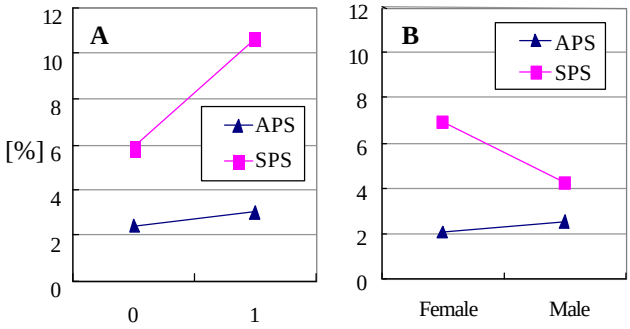

Figure 6. Interaction between laughter and speech type (panel A) and speaker's sex and speech type (B). '0' and '1' in panel A denote absence and presence of laughter. The ordinate shows shortening rate.

4.3.2 Shortening of lexical long vowels

Japanese has phonological contrast of the vowel length, and, there are many minimal pairs like /obasan/ ('aunt'), /obaHsan/ ('grand ma'), and /oHbasan/ (a surname), where 'H' stands for a long vowel. But long vowels could be shortened sometimes. Taking examples from English loans, /deHtaH/ and /deHta/ both mean 'data', and /pataHN/ and /pataN/ (where 'N' is a moraic nasal) both mean 'pattern'.

Analysis of 55282 phonological long vowels [15] revealed influences of both extra-linguistic and linguistic factors on the rate of shortening. The former included speaker's sex (Female 4.20% shortened, Male 2.77%), speech-type (SPS 5.97%, APS 2.44%), and 'laughter' (With 7.83%, Without 3.18%). Very interestingly, however, 3-way ANOVA revealed that none of the 3 extra-linguistic factors was significant at the 0.01 level. What really mattered were the interactions between sex and speech-type ($P<.003$), and, speech-type and laughter ($P<.022$).

Figure 6 shows the nature of these interactions. This figure suggests that long vowels tended to be shortened

when a speaker was mentally relaxed (as indicated by the presence of laughter and/or situational difference between SPS and APS). Moreover, and interestingly, females tended to be more relaxed than males under the situation of SPS.

Influence of linguistic factors was complex. Important factors involved word class (Loan 8.38%, Sino-Japanese 2.06%), position in a SUW (Initial 0.09%, Medial 3.77%, Final 4.71%), and, accentedness of the long vowel (Accented 2.58%, Unaccented 4.87%), among others. 3-way ANOVA revealed, moreover, all main factors and all interactions of the first and second order were significant with at least 0.001 level. See [15] for the interpretation of linguistic factors.

4.3.3 Morphological variations

As an example of morphological variation, word coalescence of /de/ and /wa/ into /zya/ was analyzed [16].

There were two types of /de/+/wa/ differing in the POS of /de/, --case particle of 'place' and *rentaikei* of the auxiliary verb /da/--, and this diff erence affects strongly the coalescence. As for particle, coalescence rates were 0.8 and 4.7% in APS and SPS respectively, while the corresponding rates of auxiliary verb were 28.2 and 59.0%.

3-way ANOVA of sex, speech-type, and laughter, applied separately for particle and auxiliary verb revealed main effects of speech-type and laughter, and their interaction, for both particle and auxiliary verb.

Coalescence rate of particle without laughter was 1.6% when pooled over APS and SPS, but the rate jumped up to 12.5% when there was laughter. The corresponding rates in the case of auxiliary verb were 41. case of particle, the effect of laughter appeared only in SPS, while in the case of auxiliary verb (panel B), the effect appeared in both speech-types. In addition, figure 8 shows the relationship between the impressionistic rating of speaking style and the observed coalescence rate. Despite the large difference of coalescence rate due to the POS status, coalescences of particle and auxiliary verb both correlated well with the impressionistic rating.

Moreover, figure 8 seems to provide a key to clarify the nature of the interactions shown in figure 7. In figure 8, the line standing for the particle showed step-like change between the low (1 and 2) and high (3-5) speaking styles, while the change in auxiliary verb was gradual, the consequence being that coalescence of particle could be a strong signal of the descent in speaking style.

Put differently, people perceive keenly the change in speaking style (descent in this case) when they listened to a variant of low occurrence probability (coalesced particle in this case), but they do not necessarily perceive the change when they listened to a variant of high occurrence probability (coalesced auxiliary verb).

This hypothesis is supported by the analysis reported in the literature [17] where moraic nasalization of two /no/ particles were analyzed. The rate of moraic nasalization differed considerably between the case particle (0.5%) and nominalization particle (51%). And the observed curves of nasalization as a function of impressionistic rating of speaking style showed expected difference, i.e., step-like and gradual changes in case and nominalization particles respectively.

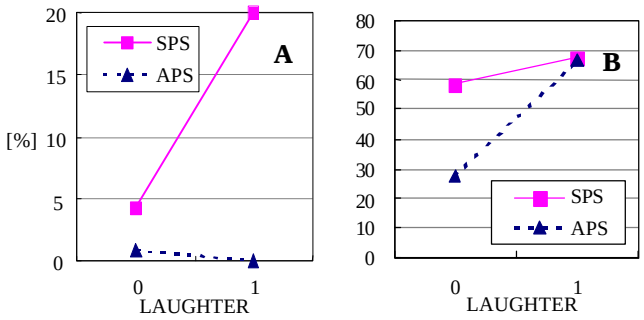

Figure 7. Interaction between laughter and speech type in particle (panel A) and auxiliary verb (B). The ordinate stands for the coalescence rate [%].

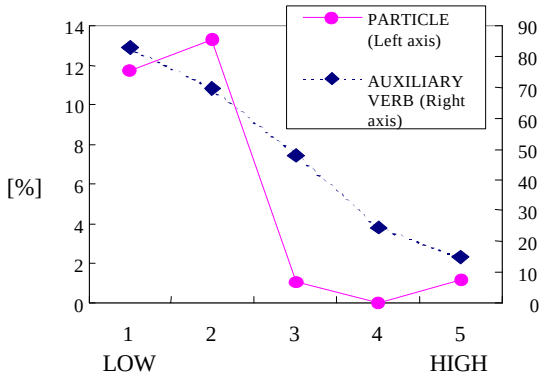

Figure 8. Correlation between the impressionistic rating of speaking style (abscissa) and the coalescence rate.

Under this hypothesis, interaction captured in figure 7 could be interpreted in the following way. Coalescence of particle in APS was suppressed regardless of the presence of laughter, because the coalescence might be a strong cue for the perception of low speaking style, which was not welcomed in APS. The effect of laughter was visible in SPS because there was no need to suppress low speaking style in SPS. On the other hand, the coalescence of auxiliary verb was not suppressed in APS because this coalescence does not run the risk of being a cue for the perception of style lowering.

Here, it is interesting to note that the relation between the laughter and speech-style in the shortening of long vowels (panel A of figure 6) is similar to the relation in

the particle coalescence (panel A of figure 7), and, the range of probability of vowel shortening was similar to that of particle coalescence.

4.4 Intonation and discourse structure

The intonation and other miscellaneous labeling of the *Core* are underway. Figure 9, taken from an ongoing study at the NIJLA, shows the relationship between the F0 shape and discourse structure [18]. 'DSP' and 'P' stand respectively for the strength of discourse segment boundaries and F0 peak height of lexically accented syllables. 'P-2' and 'P-1' are the lexical accents preceding the discourse boundary in question and 'P+1' and 'P+2' are the ones following the boundary. The relationship between the peak height and boundary strength were reversed before and after the boundary. Strong boundary (i.e., DSP2) was marked with lower preceding peaks and higher following peaks, and, weak boundary (DSP0) was marked with higher preceding peaks and lower following peaks.

This figure is a convincing demonstration of both the potential of CSJ as the resource of discourse research and the validity and usefulness of the X-JToBI labeling scheme.

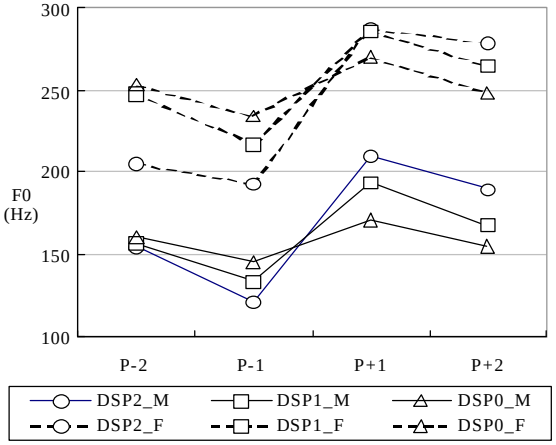

Figure 9. Comparison of the peak F0 height of accents across discourse segment boundaries (see text). Real and broken lines stand for female and male data respectively.

5. CONCLUSION

The results reported in this paper, together with the works reported elsewhere in this workshop, suggests strongly the very useful nature of the CSJ. The compilation of CSJ is still going on aiming at the final public release in the spring of 2004. Although there still remain problems to be hurdled, I believe firmly that the release of CSJ opens up a new era in the study of spontaneous speech, both home and abroad.

ACKNOWLEDMENT

I appreciate deeply the courtesy of the CSJ speakers who kindly gave us permissions to make their speech publicly available.

REFERENCES

[1] K. Maekawa, H. Koiso, S. Furui, and H. Isahara. "Spontaneous speech corpus of Japanese," *Proc. 2nd LREC*, Athens, 947-952, 2000.

[2] H. Koiso, N. Tsuchiya, Y. Mabuchi, M. Saito, T. Kagomiya, H. Kikuchi and K. Maekawa. "Transcription criteria for the *Corpus of Spontaneous Japanese*," *Japanese Linguistics*, Nat'l Inst. for Japan. Lang., 9:43-58, 2001.

[3] K. Uchimoto, C. Nobata, A. Yamada, S. Sekine, and H. Isahara. "Morphological analysis of Corpus of Spontaneous Japanese," This volume.

[4] H. Kikuchi and K. Maekawa. "Performance of segmental and prosodic labeling of spontaneous speech," This volume.

[5] J. Venditti. Japanese ToBI Labeling Guidelines." *OSU Working Papers in Linguistics*, 50, 127-162, 1997.

[6] K. Maekawa, H. Kikuchi, Y. Igarashi, and J. Venditti. "X-JToBI: An extended J_ToBI for spontaneous speech," *Proc. 7th ICSLP*, Denver, 3: 1545-1548, 2002.

[7] H. Fujisaki. "Prosody, Models, and Spontaneous Speech" In Y. Sagisaka, et al., (Eds.), *Computing Prosody: Computational Models for Processing Spontaneous Speech*, 27-42, New York: Springer.

[8] K. Maekawa and N. Kitagawa. "How does speech transmit paralinguistic information?" *Cognitive Studies*, 9:3, 46-66, 2002.

[9] K. Takeuchi, K. Takanashi, I. Morimoto, H. Koiso, and H. Isahara. "Committee based discourse purpose assignment: Discourse structure annotations of spontaneous Japanese Monologue". This volume.

[10] K. Takanashi, T. Maruyama, K. Uchimoto, and H. Isahara. "Identification of 'sentence' in spontaneous Japanese – Detection and modification of clause boundaries". This volume.

[11] Y. Sagisaka, et al., "A large-scale Japanese speech database," Proc. ICSLP, Kobe, 1089-1092, 1990.

[12] H.Nanjo and T.Kawahara. "Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition," *Proc. IEEE-ICASSP*, 725-728, 2002.

[13] T. Shinozaki and S. Furui. "Error analysis using decision trees in spontaneous presentation speech recognition," *Proc. ASRU2001*, Madonna di Campiglio, Trento, 2001.

[14] K. Maekawa and H. Kikuchi. Corpus-based analysis of vowel devoicing in spontaneous Japanese –An interim report," To be published in J. van de Weijer, K. Nanjo, and T. Nishihara (Eds.) *Japanese Voicing*.

[15] K. Maekawa. "Shortening of lexical long vowels in spontaneous speech – Analysis of *Corpus of Spontaneous Japanese*." *Proc. 2002 Spring Meeting of the Society of Japanese Linguistics*, 43-50, 2002.

[16] K. Maekawa. "Study of language variation using *Corpus of Spontaneous Japanese*." *Journal of the Phonetic Society of Japan*, 6:3, 48-59, 2002.

[17] H. Koiso, M. Saito, Y. Mabuchi and K. Maekawa. "Hanashikotobaniokeru zyoshino hatsuonkagensyouno zittai" *Proc. 10th Meeting of Shakaigengo-kagakukai*, 215-220, 2002.

[18] K. Yoneyama, H. Koiso, and J. Fon. "A corpus-based analysis on prosody and discourse structure in Japanese spontaneous monologues." This volume.