PROSODIC STRUCTURES IN JAPANESE AND ENGLISH DIALOGUES

Dieter Huber
Department of Information Theory
Chalmers University of Technology
S-412 96 Gothenburg, Sweden

1. INTRODUCTION

This paper presents a comparative study of prosodic structures in equivalent Japanese and English dialogues. A unified approach to the description and classification of intonation is proposed, which uses the F0-tracings of connected speech dialogue as input and performs speaker independent segmentation into prosodically defined information units. The time-alignment of these units with linguistic structure is established separately for each language, which permits both monolingual classification and bilingual comparison of the prosodic data. The primary research goals at this stage are (i) to assess the feasibility of the suggested approach, and (ii) to provide a detailed quantitative description of prosodic struc-tures and their underlying communicative significance in a comparative Japanese-English perspective. Three prosodic parameters are investigated systematically: intonation, pausing and laryngealization.

2. DATA

The material chosen for this study was selected from the ATR Bilingual Dialogue Database (Kurematsu 1990, Huber 1991b) and consists of six recordings of the first of seven simulated telephone dialogues conducted within the applications domain of conference registration. The corresponding Japanese and English texts of this dialogue are listed in tables I and II.

Ten speakers participated in the recording of the material: five native speakers of Standard Japanese (2 female, 3 male) and five native speakers of British (1 male) and American (2 female, 2 male) English. The speakers were selected from the research staff employed

Table I Dialogue 1 (Japanese Version)

役割	日本語
質問者	もしもし。 そちらは会議事務局ですか?
事務局	はい。 そうです どのようなご用件でしょうか?
質問者	会議に申し込みたいのですが。 どのような手続をすればよろしいのでしょうか。
事務局	登録用紙で手続きをして下さい。 登録用紙は既にお持ちでしょうか。
質問者	いいえ。 まだです。
事務局	分かりました。 それでは、登録用紙をお送り致します。 ご住所とお名前をお願いします。
質問者	住所は大阪市北区茶屋町二十三です。 名前は鈴木真弓です。
事務局	分かりました。 登録用紙を至急送らせて頂きます。
質問者	よろしくお願いします。 それでは失礼します。

Table II Dialogue 1 (English Version)

Speaker	Utterance
Questioner	Hello. Is this the Conference office?
Office	Yes. That's right. May I help you?
Questioner	I'd like to apply for the conference. How can I apply?
Office	Please apply with a registration form. Do you already have a registration form?
Questioner	No. Not yet.
Office	All right. We'll send you a registration form. Your name and address, please?
Questioner	My address is 23 Chayamachi, Kita-ku, Osaka My name is Mayumi Suzuki.
Office	All right. We'll send you a registration form immediately.
Questioner	Thank you very much. Good-bye.

at ATR and counted, at the time of the recording sessions, between 24 and 35 years of age. None of the subjects reported any history of speech and hearing disorder.

Registration of the speech samples was carried out under optimal conditions (anechoic studio, no face-to-face interaction) at the ATR Auditory and Visual Perception Research Laboratories, using high-quality digital recording equipment (SONY DTC-1000ES). During the recording sessions, the conversations were conducted monolingually, engaging pairwise combinations of speakers within the same language group. Scripted versions of the dialogues were handed to the subjects upon arrival at the recording studio. Only minimal instructions were given on how to render the material during the recordings. Speakers were encouraged to deliver the dialogues in a relaxed, conversational style, and to avoid as much as possible falling into a normal text reading mode. The English subjects were permitted to choose for themselves the names and addresses (English or Japanese) they wanted to use in the dialogues, in order to avoid any potential pronunciation problems. No instructions were given concerning the inclusion of pauses or the placement of stress (emphatic or contrastive) at any place in the utterances. If speakers produced false starts, hesitations, or departed from their intended rendering of the conversation in any way, they were free to decide themselves whether and how to include any corrections. Further details concerning the material, the subjects, the equipment and the recording procedures are described in Huber (1991b).

3. ANALYSES

For purposes of analysis and storage, the DAT mastertapes produced at ATR were downsampled to 8 kHz (maintaining 16-bits quantization) and transfered to auxiliary disk storage running under a DEC GPX work station at the Department of Information Theory, Chalmers University of Technology. Pitch extraction was performed using the DWAPIT pitch determination algorithm presented earlier (Hedelin & Huber 1990). Pitch estimates were obtained at 16-ms intervals for both periodic and aperiodic (laryngealized) stretches of speech. Segmentation of the F0-tracings into *intonation units* (IU) was performed following the approach published in Huber (1989a). According to this approach, two global declination lines, which approximate the trends in time of the peaks (topline) and valleys (baseline) of F0 across

the utterance, are computed by the linear regression method. Computation is reiterated every time the *Pearson product moment correlation coefficient* drops below a preset level of acceptability. Segmentation is thus performed without prior knowledge of higher level linguistic information, with the termination of one unit being determined by the general resetting of the intonation contour wherever in the utterance it may occur. The F0 onsets (intercepts) and offsets (endpoints), durations, declination line slopes and key values of these intonation units, as well as their time-alignment with features of linguistic structure were established individually for each of the speakers participating in this study.

Detailed descriptions of the algorithm and its application to automatic speech recognition, spoken language parsing (integrating speech processing and natural language processing techniques), disambiguation, and automatic spoken language interpretation (interpreting telephony) have been published earlier (see list of references). Figure 1 shows the F0-contours of the six conversations (i.e. representing each of the ten speakers at least once) segmented into intonation units.

4. RESULTS

In this section, only the results refering to the number of intonation units, their average durations, their alignment with linguistic structure, and the occurrence of dialogue internal pauses are reviewed. In addition, the occurrence of laryngealization at boundary locations within the dialogues is investigated. Given the limited socpe of this report, no further parameterization, quantification and statistical evaluation has been attempted at this stage.

4.1 Number of Intonation Units

A total of 132 intonation units was established in the six conversations. 75 of these units (56.8 %) pertain to the Japanese recordings, the remaining 57 (43.2 %) to the corresponding English material. The exact figures per conversation (viz. speaker constellation) and language are listed below in table III.

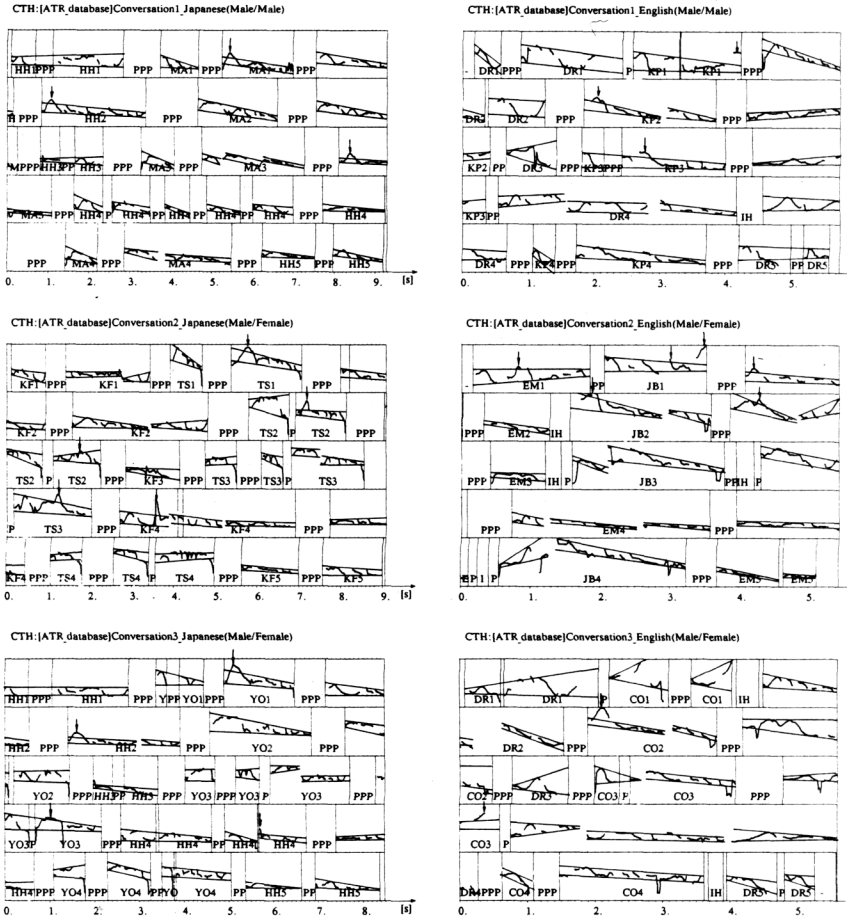

Figure 1 F_0 contours of the first dialogue segmented into *intonation units*. The graphs in the left column depict the three Japanese recordings, the graphs in the right column the three English ones. The individual units are indicated by their respective baseline/topline configurations. Arrows identify areas of prominence outside the F_0 range defined by topline-baseline. The dotted horizontal lines serve as calibration marks at 100, 200 and 300 Hz, the displayed range thus covering 0-400 Hz. Vertical lines identify turn and/or pause boundaries. Individual turns are denoted by indicating the speaker label and the sequence number of his or her turn within the respective dialogue. Speaker labels **TS** and **YO** represent the two female, **HH**, **MA** and **KF** the three male Japanese subjects. The five English speaking subjects are identified as **JB** and **CO** (American females), **EM** and **KP** (American males) and **DR** (British male).

Table III Number of Intonation Units per Conversation and Language. Percentage figures are based on the accumulated material.

Recording		Japanese	English
Conversation_1	n	24	18
(male-female)	%	18.2	13.6
Conversation_2	n	27	20
(male-female)	%	20.5	15.1
Conservation_3	n	24	19
(male-female	%	18.2	14.4

As can be observed from these data, the Japanese speakers participating in this study display a consistent propensity to subdivide their dialogue utterances into a larger number of prosodically cued chunks than their English counterparts. There also appears to be a slight tendency for the female speakers to produce more intonation units than their male dialogue partners, as indicated by the larger number of units established in the male-female conversations. This tendency is more pronounced in the English material, however, and conforms with the data for Swedish discourse published earlier in Huber (1989c).

4.2 Intonation Unit Durations

The overall durations of the six conversations, both with and without dialogue-internal pauses, are listed in table IV.

Table IV Durations (in ms) per conversation and language. Figures in line 1 state the total overall durations t, figures in line 2 the actual utterance durations u, i.e. the length of the actual speech sequences without dialogue-internal pauses. The number of pauses contained in the respective conversation are added in brackets.

Recording		Japanese	English
Conversation_1	t	46224(22)	28464(15)
(male-female)	u	31536	23456
Conversation_2	t	45104(21)	26720(12)
(male-female)	u	32656	21856
Conservation_3	t	42288(26)	27584(14)
(male-female	u	31216	22762

Table V summarizes the average durations (i.e. means $\bar{x}$ and standard deviations s) of the intonation units established in each of the six conversations.

Table V Intonation unit durations (in ms) per conversation and language. Lines state the means, $\bar{x}$ lines s the standard deviations.

Recording		Japanese	English
Conversation_1	$\bar{x}$	1314	1303
(male-female)	s	621	732
Conversation_2	$\bar{x}$	1209	1093
(male-female)	s	543	601
Conservation_3	$\bar{x}$	1301	1198
(male-female	s	517	683

These figures reveal a clear and consistent tendency of the Japanese speakers to produce:

(1) longer overall utterance durations;
(2) more pauses per conversation;
(3) both longer and less varied intonation unit durations.

Hypothesis testing with χ^2 at 5% significance level reveals, however, that only the differences summarized under (1) and (2) are statistically significant. Regarding potential cues for systematic speaker variability between the female and male subjects participating in this study, it is interesting to note the distinctly shorter average intonation unit durations found in the conversations involving one female dialogue partner. This tendency is again more pronounced in the English material and conforms with the findings on female Swedish speech reported earlier (Huber 1989c).

4.3 Time-alignment with Linguistic Structure

The time-alignment of the 132 intonation units with sentences (**S**), nounphrase/subjects (**SUB**), verbphrases (**VP**), complements (**COM**), adverbials (**ADV**) and parentheticals or other kinds of structural parallelism (**PAR**) is summarized in the bar chart in figure 2. In addition to these labels I found it necessary to include even a category (miscellaneous - **MIS**) to capture occurrences of intonation units that begin at or terminate somewhere *within* a constituent and thus cannot be classified in terms of established grammatical theory.

As can be seen from these data, the overwhelming majority (88.2%) of intonation units identified by the segmentation algorithm correspond in a clearly defined way with units of syntactic structure. This regular syntax-prosody correspondence, however, is significantly more prevalent in the Japanese (97.3%) than in the English (79.1%) conversations. Intonation units pertaining to the **MIS** category were found only once in the Japanese material (cf. HH2 in conversation 3), whereas in the English recordings, this kind of essentially non-grammatical intonation unit was produced not only considerably more often, but by all five subjects, in a rather consistent fashion, at various places of the dialogues.

Most commonly in our accumulated dialogue material, intonation units correspond in a regular fashion with single sentences (43.9%). This general tendency can be observed both in the Japanese and in the English material, however, with a significantly higher percentage of co-occurrences in the English conversations. It must be appreciated in this context that the majority of sentences associated with a separate intonation unit in the dialogues constitute occurrences of single-clause sentences (84.5%). Regarding larger structures beyond

the sentence domain, the English subjects display a distinctly greater tendency to process two or even three consecutive sentences in terms of one single intonation unit. Conversely, intonation units corresponding to single constituents in the subsentence domain (i.e. **SUB, VP, COM, ADV** and **PAR**) were almost exclusively found in the Japanese material. However, only 41 (32.5%) of the 126 "bunsetsu" phrase units contained in the accumulated Japanese conversations were actually associated with separate intonation units.

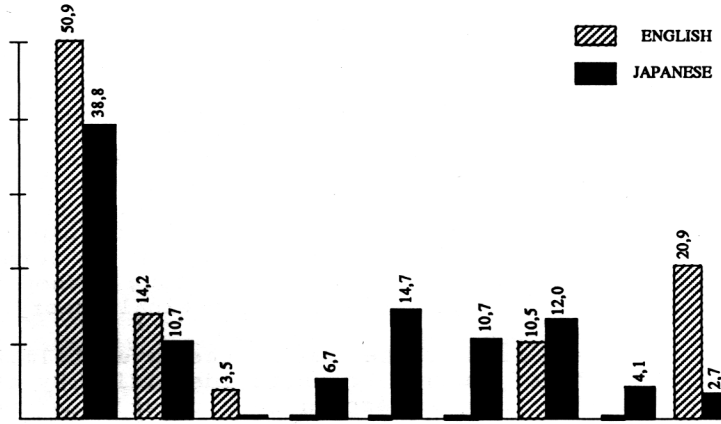

Figure 2 Correlations between intonation units and features of linguistic structure. Percentages are calculated separately for each language. The preposed n in category (**S**) states the number of complete, consecutive sentences covered by the time extent of one single intonation unit.

No statistically significant difference between the two language groups is evident with respect to the **PAR** category, viz. both the Japanese and the English speakers processed the seventh turn of the dialogue in terms of between 4 and 6 separate, largely identical intonation units. This congruity appears to be specially relevant in view of the fact that the English speakers were permitted to exchange the Japanese name and address for a more familiar English one in order to avoid any potential pronunciation problems.

4.4 Pausing

The number of dialogue-internal pauses per language and conversation has already been listed earlier in table IV together with the overall duration measures. Table VI below summarizes the average durations (i.e. means $\bar{x}$ and standard deviations *s*) of the pauses established in each of the six conversations.

Table VI Pause durations (in ms) per conversation and language. Lines $\bar{x}$ state the means, lines s the standard deviations..

Recording		Japanese	English
Conversation_1	$\bar{x}$	668	335
(male-female)	s	119	143
Conversation_2	$\bar{x}$	593	405
(male-female)	s	168	193
Conservation_3	$\bar{x}$	426	344
(male-female	s	201	196

These figures reveal a clear and consistent tendency of the Japanese speakers to produce not only more pauses per conversation (compare table IV and discussion in the earlier subsection on intonation unit durations) but also to use both (1) longer and (2) less varied pause durations than their English speaking counterparts.

As in the case of the number of dialogue-internal pauses, hypothesis testing with χ^2 at 5% significance level confirms the statistical significance of these differences. Regarding potential cues for systematic speaker variability between the female and male subjects participating in this study, it is interesting to note the distinctly shorter average pause durations established in the Japanese conversations involving one female dialogue partner, whereas the opposite tendency can be observed in the corresponding English material. On the other hand, both the Japanese and the English data reveal larger standard deviations in the male-female as compared with the male-male conversations, thus indicating a higher degree of pause duration variability in conversations involving a female dialogue partner.

Clearly, both larger and more varied discourse data need to be studied in order to arrive at any firmer conclusions regarding these speaker sex differences in a cross-linguistically valid way. Further study based on more speakers and a larger material also needs to be undertaken in order to establish the correlations between speech pauses and various kinds of textually, syntactically, semantically and/or physiologically induced boundaries. We find it encouraging to note, however, that the limited set of results included in this present study is in agreement with the general tendencies of Japanese pausing behaviour reported lately by among others Sugito (1990).

4.5 Laryngealization

A total of 57 occurrences of larygealization (i.e. patterns of aperiodic voice vibration) was observed at various pre-boundary and post-boundary locations in the accumulated dialogue material. The locations of these patterns in the six dialogues can easily be identified in the pitch tracings in figure 1, where they appear as unstable, irregular F0 fluctuations and/or spikes, typically outside the range of the baseline-topline configuration. Figure 3 shows one of these occurrences of laryngealization as an example taken from the Japanese material.

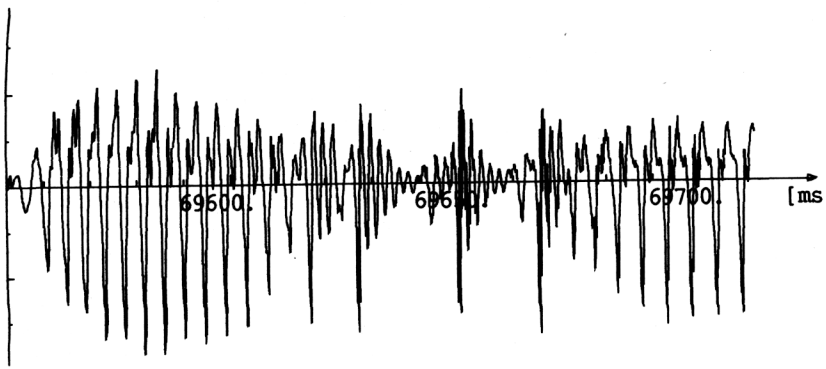

Figure 3 Laryngealization (i.e. aperiodic voice vibration) produced by the female Japanese speaker **TS** at the second turn boundary of Conversation 2

The acoustical characteristics of these patterns and their function as complementary/compensatory boundary markers have been discussed earlier in Hedelin & Huber (1990). It has also been claimed that female speakers differ in a systematic way from male speakers in their use of laryngealization as a boundary cue in connected speech (Huber 1989c). This claim, based originally on orally read Swedish texts, is further substantiated by the results of the present investigation, which show that both the English and the Japanese female speakers participating in this study:

(1) make distinctly more frequent use of laryngealization as a boundary marker than their male counterparts (on the average 59.6% *versus* 40.4%)

(2) apparently prefer to employ *creak (vocal fry)* patterns at pre-boundary locations where the men - in as far as they use any laryngealization at all - produce predominantly *creaky* voice.

Compared with these rather distinct speaker sex differences, the variability in the frequency of occurrence of these patterns between the two languages appears to be comparatively minor, as reflected in the following percentages:

ENGLISH	48.9%
JAPANESE	51.1%

It must be appreciated in this context, however, that laryngealization as a boundary marker (either alone or together with other juncture cues such as for instance pause, declination resetting, F0-fall-rise patterns, devoicing, phonological blocking, etc) displays its strongest potential in the highly structured and optimally controlled text reading mode (cf. Huber 1991a), whereas it is used to a significantly lesser degree in both orally read isolated (i.e. semantically unrelated) sentences and in dialogues, viz. where the boundaries are signaled by other linguistic (e.g. semantic incoherence between the sentences on the list) or paralinguistic (e.g. changes in voice quality at conversational turn boundaries) means.

5.APPLICATIONS

It is today widely acknowledged that the accurate representation of the prosodic characteristics of speech is of paramount importance for all aspects of speech signal processing (synthesis, coding, transmission, compression, enhancement, etc). In speech coding, for instance, the quality of the vocoded speech deteriorates rapidly as a function of imprecise F0 estimates. In speech synthesis, considerable effort is dedicated today to the development and implementation of prosodic models for the generation of natural sounding pitch contours in text-to-speech. Equally important, modern speech recognition systems make increasingly use of the F0 information inherent in the pitch contours for the segmentation of the continuous speech utterance into semantically meaningful "chunks" and for the automatic detection of the stressed parts of the utterance. Finally, in automatic interpretation of spoken language, unlike in machine translation of written texts, it is important not only to correctly translate the verbal contents of the utterance, but also to transform the prosodic characteristics of the source language input into an equivalent representation at the target language output. This implies that prosody must be parsed, understood, transfered, and generated (Myers 1989) in order to enable the system to make intelligent use of suprasegmental information during the various constituent stages of spoken language interpretation: automatic speech recognition (ASR), machine translation (MT), and speech synthesis (SS).

Ideally, the same framework of linguistic-prosodic description that is used in source language ASR for segmentation, classification, and the automatic detection of stress, should also be applicable in target language SS to synthesize intonation contours, accentuation patterns, and durational variations. Moreover, it should continuously supply relevant information to the MT module of the interpreting system to support NLP parsing, disambiguation, anaphoric resolution, etc. To operate in such a fully integrated mode, prosodic transfer requires (a) an adequate internal representation of prosody that can be used in a unified manner during the ASR, MT and SS stages of the automatic interpretation process, and (b) detailed knowledge of prosodic phenomena and their underlying communicative significance in a comparative, multilingual perspective.

This paper presents such a unified approach to the des-

cription of F0-based prosodic phenomena, and evaluates its applicability in a comparative study of prosodic structures in equivalent Japanese and English dialogues. An algorithm for the segmentation and broad classification of continuous speech into prosodically defined information units has been introduced. It has been demonstrated that this algorithm reliably segments connected speech dialogue into linguistically meaningful units both in Japanese and English.

As shown earlier (cf. Huber 1989a), the segmentation algorithm not only aims to unearth the underlying information/intonation structure of the utterance, but permits the description and quantification of individual intonation units in terms of 10 parameters (i.e. duration, declination line slope, onset, offset and resetting, for the baselines and toplines respectively). In addition, once the extent of an intonation unit has been established both in the time and in the frequency domain, areas of prominence indicating the semantically most important parts of the utterance can easily be identified (and quantified!) as overshooting F0 excursions (cf. figure 1) that provide valuable points of departure for further linguistic analyses and island parsing strategies.

Clearly, the data published in this report represent only to a very limited degree the full range of linguistic-prosodic information present in the acoustical speech signal. Further research is under way and will be pursued (i) following a simulative approach, (ii) based on a large amount of data, viz. covering the total of 280 conversations contained in the ATR bilingual dialogue database, and (iii) exploiting the full range of parametric description provided by the model (i.e. declination line slope, onset, prominence, etc) thus permitting the internal representation of prosodic features in quantitative terms.

ACKNOWLEDGEMENTS

The research reported on in this paper was initiated during my stay as invited researcher at the ATR Interpreting Telephony Research Laboratories in Kyoto, Japan. I wish to thank Dr. Akira Kurematsu, Dr. Tsuyoshi Morimoto, Shigeki Sagayama and Dr. Yoshinori Sagisaka for their support and valuable comments. The research conducted at Chalmers University of Technology in Sweden, was made possible in part by support from the Swedish Board of Technical Development (STU).

REFERENCES

HEDELIN, P. & HUBER, D. (1990) Pitch period determination of aperiodic speech signals, **Proc. ICASSP-90**, pp.361-364, Albuquerque, U.S.A.

HUBER, D. (1989a) A statistical approach to the segmentation and broad classification of continuous speech into phrase-sized information units, **Proc. ICASSP-89**, pp.600-603, Glasgow, Scotland

HUBER, D. (1989b) Parsing speech for structure and prominence, **Proc. International Workshop on Parsing Technologies**, pp. 115-125, Carnegie Mellon University, Pittsburgh, U.S.A.

HUBER, D. (1989c) Voice Characteristics fo female speech and their representation in computer speech synthesis and recognition, **Proc. EUROSPEECH-89**, pp. 477-480, Paris, France

HUBER, D. (1990a) Prosodic transfer in spoken language interpretation, **Proc.ICSLP-90**, pp.509-512, Kobe, Japan

HUBER, D. (1990b) Speech style variations of F_0 in a cross-linguistic perspective, **Proc.SST-90**, pp.186-191, Melbourne, Australia

HUBER, D. (1991a) On the discourse function of intonation, **Proc.XIIth ICPhS**, Vol.5, pp.190-193, Aix-en-Provence, France

HUBER, D. (1991b) A Bilingual Dialogue Database for Automatic Spoken Language Interpretation between Japanese and English, ATR Technical Report TR-I-0196, Kyoto, Japan

KUREMATSU, A. *et al.*(1990) ATR Japanese speech database as a tool for speech recognition and synthesis, **Speech Communication** 9(4), pp.357-363

MYERS, J.K. & T. TOYOSHIMA (1989) Known Current Problems in Automatic Interpretation: Challenges for Language Understanding, ATR Technical Report TR-128, Kyoto, Japan

SUGITO, M. (1990) On the role of pauses in production and perception of discourse, **Proc.ICSLP-90**, pp.513-516, Kobe, Japan