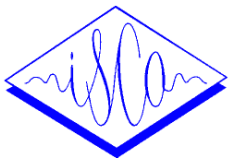

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

PROSODIC TRANSFER IN SPOKEN LANGUAGE INTERPRETATION

Dieter Huber

ATR Interpreting Telephony Research Laboratories
Seika-cho, Soraku-gun, Kyoto 619-02, Japan
and
Department of Information Theory, Chalmers University of Technology
S-412 96 Gothenburg, Sweden

ABSTRACT

This paper presents a unified approach to the description and classification of prosodic phenomena in continuous speech, and evaluates its applicability to interpreting telephony for a limited transfer task between equivalent samples of Japanese and English dialogue. An algorithm is proposed which uses the F_0 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. A tentative set of transfer roles for the "translation" of prosodic features between Japanese and English is introduced, and directions for further research are indicated.

1. INTRODUCTION

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 [8], 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 proposes a unified approach to the description and classification of prosodic phenomena in continuous speech, and evaluates its applicability to automatic spoken language interpretation for a limited transfer task between equivalent samples of Japanese and English dialogue. An algorithm is presented which uses the F_0 tracings of connected speech dialogue as input and performs speaker independent segmentation into prosodically defined information units. Detailed descriptions of the algorithm and its application to text-to-speech synthesis, automatic speech recognition, spoken language parsing (integrating speech processing and natural language processing techniques), disambiguation, and speaker adaptation have been published earlier [2-6]. The present study is aimed to adapt the same method to the problem of prosodic transfer in automatic spoken language interpretation between Japanese and English. The primary research goals at this initial stage are (i) to assess the feasibility of the material and of the suggested approach, and (ii) to formulate a first tentative set of transfer rules suitable for further simulative research. Given these initial goals, the scope of this report has been restricted both with respect to the amount of data (only one dialogue) and the number of suprasegmental features (only intonation, pausing, and some duration measures) investigated.

2. DATA

The material chosen for this study was selected from the ATR bilingual dialogue database [7], 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.

Table I Dialogue 1 (Japanese Version)

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)

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.

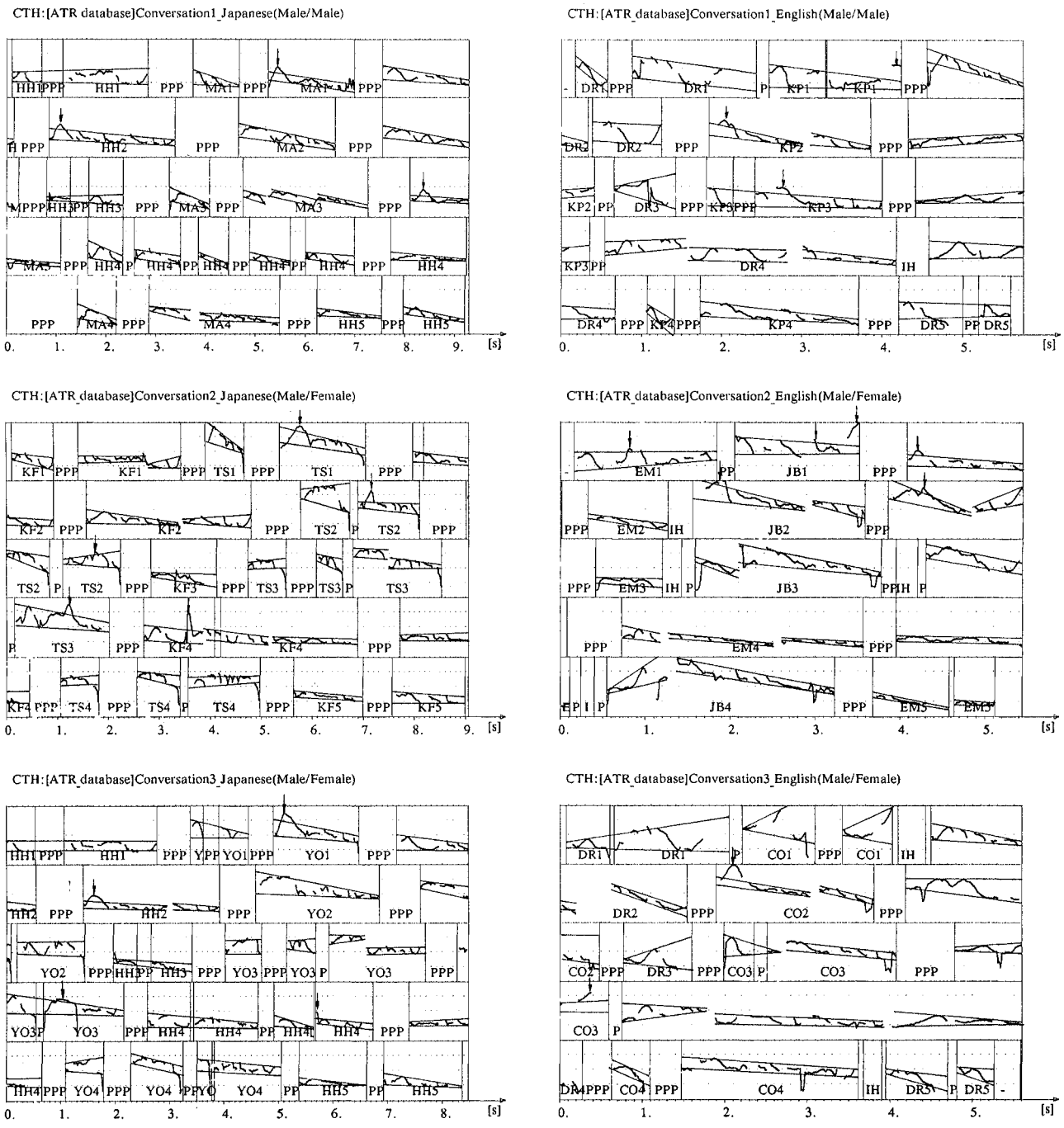

Figure 1 F_0 contours of the first dialogue segmented into *intonation units*. The graphs in the left column depict the three Japanese recordings, the graphs in the right column the three English ones. The individual units are indicated by their respective baseline/topline configurations. Arrows identify 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).

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 conversations were conducted monolingually, engaging pairwise combinations of speakers within the same language group. 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). Further details concerning the material, subjects, equipment and recording procedures are described in reference [7].

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 [1]. Segmentation of the pitch contours into prosodically defined information units (henceforth refered to as *intonation units* or IU) was performed both in the Japanese and in the English speech material following the approach published earlier [2-3]. Thus two global declination lines were computed by the linear regression method, which approximate the trends in time of the peaks (topline) and valleys (baseline) of F_0 across the utterance. Computation was reiterated every time the Pearson correlation coefficient dropped below a preset level of acceptability ($r > 0.5$). Segmentation was 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. Figure 1 shows the F_0 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 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. Given the limited scope of this interim report, no further parameterisation, quantification and statistical evaluation is attempted at this stage. A more comprehensive report based on the full set of data contained in the ATR bilingual dialogue database is under preparation.

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.

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

Recording		Japanese	English
Conversation_1 (male-male)	*n*	24	18
	%	18.2	13.6
Conversation_2 (male-female)	*n*	27	20
	%	20.5	15.1
Conversation_3 (male-female)	*n*	24	19
	%	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 in [5].

4.2 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 *l*, 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 (male-male)	*l*	46224(22)	28464(15)
	u	31536	23456
Conversation_2 (male-female)	*n*	45104(21)	26720(12)
	u	32656	21856
Conversation_3 (male-female)	*n*	42288(26)	27584(14)
	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 $\bar{x}$ state the means, lines *s* the standard deviations.

Recording		Japanese	English
Conversation_1 (male-male)	$\bar{x}$	1314	1303
	s	621	732
Conversation_2 (male-female)	$\bar{x}$	1209	1093
	s	543	601
Conversation_3 (male-female)	$\bar{x}$	1301	1198
	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 shows, 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 in [5].

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 dialogue.

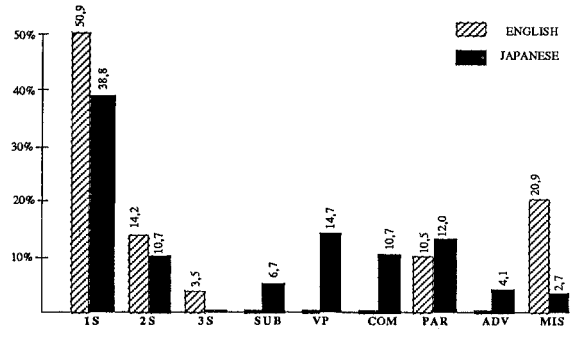

Figure 2 Correlations between intonation units and features of linguistic structure. Percentages are calculated separately for each language. The preposed *n* in category (**S**) states the number of complete, consecutive sentences covered by the time extent of one single intonation unit.

Most commonly in our accumulated dialogue material (43.9%) intonation units correspond in a regular fashion with single sentences. 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 (84.5%) of sentences associated with a separate intonation unit in the dialogues constitute single-clause sentences.

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 units. 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.

No 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.

5. TRANSFER RULES

Based on these findings. the following tentative set of transfer rules for the "translation" of prosodic features from English (source language) to Japanese (target language) is introduced [1] :

(1) Reduce average intonation unit length and variability by a factor of 0.9;
(2) Transfer intonation units extending over *n* complete, consecutive sentences ($n>1$) into n-1 intonation units;
(3) Translate intonation units pertaining to the **MIS** category into regular "grammatical" intonation units by moving the IU resetting to the closest constituent boundary;
(4) Process single constituents in terms of separate intonation units if they comprise ≥ 7 syllables;
(5) Process sentence initial adverbials in terms of separate intonation units regardless the number of syllables they comprise;
(6) Maintain intonation unit structure for constituents in the **PAR** category;
(7) Insert pauses between all intonation units.

Application of these rules to our three English conversations effectively transforms:

- the number of intonation units from 57 to 73 (instead of 75);
- the number of pauses from 41 to 70 (instead of 69), and
- the average overall duration per conversation (without pauses) from 22.794 ms to 28.579 ms (instead of 31.802 ms)

90.7 % of the syntax-prosody alignments actually found in the recorded Japanese conversations are correctly predicted, however, missing shorter constituents (e.g.登録用紙を), irregular resettings (cf. section 4.3), and the interspeaker variability displayed by our Japanese subjects with respect to intonation unit processing in the subsentence domain.

6. SUMMARY AND CONCLUSIONS

This paper constitutes a first interim report of a pilot study on prosodic transfer in spoken language interpretation. An algorithm for the segmentation and broad classification of continuous speech into prosodically defined information units has been presented. It has been demonstrated that this algorithm reliably segments connected speech dialogue into linguistically meaningful units both in Japanese and English. A tentative set of transfer rules for the "translation" of prosodic features between Japanese and English has been proposed, based on the limited set of duration, pausing and alignment data investigated in this study.

As shown earlier [2-6], 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 F_0 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 Gothenburg, Sweden, was made possible in part by support from the Swedish Board of Technical Development (STU).

NOTES

[1] For a first approximation, the same set of rules is taken to apply *mutatis mutandis* to the transfer from Japanese to English.

REFERENCES

[1] P Hedelin and D Huber, "Pitch period determination of aperiodic speech signals", Proc. ICASSP 90, Albuquerque, 1990

[2] D Huber, "Aspects of the Communicative Function of Voice in Text Intonation", Ph.D. Dissertation, Gothenburg/Lund,'1988

[3] D Huber, "A statistical approach to the segmentation and classification of continuous speech into phrase-sized information units", Proc. ICASSP 89, Glasgow, 1989

[4] D Huber, "Parsing speech for structure and prominence", Proc. Intern. Workshop on Parsing Technologies, Carnegie Mellon University, Pittsburgh, 1989

[5] D Huber, "Voice Characteristics fo female speech and their representation in computer speech synthesis and recognition", Proc. ESCA 89, Paris, 1989

[6] D Huber, "Prosodic contributions to the resolution of ambiguity", NORDIC PROSODY V, Åbo, 1989

[7] D Huber, "A Bilingual Dialogue Database for Automatic Spoken Language Interpretation between Japanese and English", ATR Technical Report, (forthcoming)

[8] J K Myers and T Toyoshima, "Known Current Problems in Automatic Interpretation: Challenges for Language Understanding", ATR Technical Report TR-128, 1989