ON THE DISCOURSE FUNCTION OF INTONATION

Dieter Huber

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

ABSTRACT

This study explores the differences between discourse intonation and the kind of pitch contours typically found in isolated sentences. Three kinds of material are evaluated systematically: (1) orally read lists of semantically unrelated sentences, (2) orally read narrative texts, and (3) dialogues. The material consists of equivalent samples of Swedish, English and Japanese speech, produced by native speakers (both female and male) of the respective languages. It will be shown that discourse intonation differs from intonation in semantically unrelated sentences with respect to practically all F_o parameters investigated in this study.

1. INTRODUCTION

Human speakers typically associate their verbal speech utterances with intricate patterns of voice fundamental frequency. This phenomenon has been widely attested, and is acknowledged as a universal, innate quality of speech, common to all speakers, in all languages, and in all kinds of spoken utterances. Numerous scientific studies within a variety of disciplines have been undertaken to investigate the form and function of these fundamental frequency patterns, to establish their communicative status, and to disentangle the seemingly infinite variety of linguistic and paralinguistic conditioning factors that human speakers so aptly and without apparent effort combine into one single contour. Most of these studies have been restricted to the domain of the sentence as maximal unit of linguistic processing, thus adhering to the traditional view that larger units like paragraphs, text and discourse are formed by mere juxtaposition of autarchic, independently prefabricated sentences. There is, however, convincing evidence that human speakers use variations in voice fundamental frequency in a systematic way to signal *cohesion*, *structure* and *prominence* in connected speech according to criteria other than purely syntactic, and that listeners at the other end of the speech communication chain are able to detect and to decode these prosodic messages, and to make use of them in order to gain information about the intended meaning of the utterance in its situational and co-textual context. The purpose of this study is to investigate these differences, i.e. between discourse intonation and the kind of pitch contours typically found in isolated sentences.

2. DATA

Three kinds of material are evaluated systematically: (1) orally read lists of semantically unrelated sentences, (2) orally read narrative texts, and (3) dialogues. The material has been selected from the ATR [7],[8] and the CTH [2] speech databases and comprises equivalent samples of Swedish, English and Japanese speech. The English and Japanese dialogues consist of simulated telephone conversations conducted within the applications domain of conference registration, whereas the Swedish dialogues were conducted spontaneously.

Ten native speakers of the respective languages participated in the recordings selected for this study: 3 speakers of Standard Swedish (2 male, 1 female), 3 speakers of American English (2 male, 1 female) and 4 speakers of Standard Japanese (3 male, 1 female). Registration of the speech samples was conducted in anechoic, sound-insulated recording studios both at ATR in Kyoto (Japan) and at CTH in Gothenburg (Sweden), using high-quality digital recording equipment.

3. ANALYSES

Approximately one minute of recorded speech per speaker and speech style was analysed for this study. Pitch extraction was performed using the DWAPIT pitch determination algorithm presented earlier in [3]. Pitch estimates were obtained at 16-ms intervals for both periodic and aperiodic (laryngealized) stretches of speech. Segmentation of the F_o tracings into *intonation units* (IU) was performed following the approach published in [4]. According to this approach, two global declination lines which approximate the trends in time of the peaks (topline) and valleys (baseline) of F_o across the utterance, are computed by the linear regression method. Computation is reiterated every time the *Pearson 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 F_o 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.

4. RESULTS

4.1 Number of Intonation Units

A total of 586 intonation units has been established in the accumulated material for all ten speakers. The distribution of these intonation units per language and speech style is summarized below in figure 1.

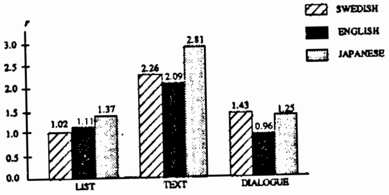

Figure 1. Intonation units per language and speech style. The bar heights *r* depict the ratio between the number of intonation units and the number of sentences contained in the respective material.

These distributions reveal a clear and consistent tendency, observable in each of the three languages, to subdivide orally read texts into a larger number of prosodically cued *chunks* than both the list and the dialogue material. All ten speakers produced predominantly one intonation unit per sentence in the list reading task, as predicted by most studies of sentence intonation, whereas in the text reading task the individual sentences were processed on the average in terms of between 2 and 3 intonation units.

Quite obviously, differences in sentence structure and informational content need to be taken into account for a comprehensive assessment of these ratios. This is particularly relevant with regard to the *r* values obtained for the dialogues which clearly reflect (1) the comparatively larger proportion of short and incomplete sentences included in the material, and (2) the more frequent use of intonation units that stretch over the time extent of several consecutive sentences (cf. [6] for a more detailed discussion). Also, the dialogue material investigated in this study contains significantly less subordination than the texts and sentences, and only few examples of it-clefts and wh-clefts that typically occur as separate, prosodically cued chunks in read narrative.

Considering the higher degree of interlanguage variability found in the dialogues, it must also be appreciated that the Swedish material consists of spontaneous conversations, i.e. including a larger proportion of hesitations, false starts, fragmentary constructions, etc. than the simulated dialogues in the English and Japanese samples.

4.2 Prosody-Syntax Alignment

The overwhelming majority (84.6%) 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 (98.2%) than in English (82.2%) and Swedish (79.9%) material. It is also slightly more pronounced in the orally read texts (85.5%) as compared with the dialogues (83.8%).

Most commonly in our accumulated dialogue material, intonation units correspond in a regular fashion with single sentences (40.3%), whereas in the text material the results are more inconsistent between the three languages investigated in this study. In 36.6% of the English and 32.4% of the Swedish texts, intonation units time-align with clauses. In the Japanese text material, on the other hand, only about one tenth (10.1%) of the intonation units pertain to the clause correspondence class, thus indicating a markedly different prosodic processing behaviour.

Larger structures beyond the sentence domain (i.e. stretching over two or three consecutive sentences) are almost exclusively found in the dialogues, with only 1.1% 3-sentence occurrences in the English and 2.1% 2-sentence occurrences in the Swedish texts. Conversely, intonation units corresponding to single constituents in the subsentence domain (i.e. nounphrase-subjects, verbphrases, adverbials, parenthetical constructions, etc) occur more often in the text (41.9%) than in the dialogue (24.9%) material, with a significant prevalence in the Japanese (60.3%) as compared with both the English (35.9%) and Swedish (29.5%) speech samples.

Only the discourse material has been scrutinized at such a detailed level of linguistic analysis. For the speech samples produced in the list reading task, a predominant one-to-one relationship between isolated sentences and single, coherent intonation units has already been established in the previous section.

4.3 Declination Line Parameters

The declination line parameters onset (intercept), offset (endpoint), duration, slope and key were calculated separately for each of the 586 intonation units investigated in this study. Statistical evaluation of these data revealed the following tendencies:

(1) Intonation units aligning with the isolated sentences from the list reading task are on the average shorter, steeper, less varied, and start with higher baseline onsets and substantially lower topline intercepts than in the discourse material;

(2) Important features of prosodic variation such as for instance rising baselines, "bi-modal" toplines, and narrow *versus* wide key (cf. [5]) do not occur in the list material at all, but are frequently used in discourse;

(3) The only parameter for which no statistically significant differences could be established between the different kinds of material is the baseline endpoint, which thus appears to provide a common point of reference, marking the bottom of a speakers voice range for both discourse and isolated sentence production.

Separate investigation of both the IU initial and IU final peaks and valleys, in order to account for the potential status of these points as independently controlled linguistic variables (e.g. [1]) revealed:

(4) significantly higher measures of variability for both the very *first* and the very *last* peaks and valleys in the intonation unit contours of the dialogue as compared with both the sentence and text material;

(5) the consistent use of categorical

distinction by all ten speakers with respect to both the first and the last peak/valley of the IU contour in the discourse but not in the list material.

4.4 Laryngealization

Patterns of aperiodic voice vibration (laryngealization) were observed to occur at various kinds of textually, syntactically and prosodically induced boundaries in our material. The acoustical characteristics of these patterns and their function as complementary/compensatory boundary cues have been discussed earlier in [3]. It has also been claimed that female speakers differ in a systematic way from male speakers in their use of laryngealization in connected speech [5]. This claim, based originally on Swedish text material, is further substantiated by the results of the present investigation, which show that the three 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 13.4% *versus* 8.1%);

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

There are, however, significant differences in the frequency of occurrence of these patterns between the three languages, as reflected in the following percentages:

SWEDISH	26.8%
ENGLISH	33.4%
JAPANESE	39.8%

Even more importantly, the use of laryngealization as a boundary cue differs markedly between the three kinds of material, where it occurs least frequently in the lists of semantically unrelated sentences (Swedish 7.3%; English 10.1%; Japanese 13.5%) and most frequently in the narrative texts (Swedish 60.4%; English 54.2%; Japanese 49.3%). The respective figures for the dialogue material (Swedish 32.3%; English 35.7%; Japanese 37.2%) reveal a somewhat intermediary status for the conversational speaking mode.

In summary, laryngealization as a boundary marker (either alone or together with other juncture cues such as for instance pause, declination resetting, F_0-fall-rise patterns, devoicing, phonological blocking, etc) displays its strongest potential in the highly structured and optimally controlled text reading mode, whereas it is used to a significantly lesser degree in the other two speaking styles, i.e. 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.

REFERENCES

[1] BRUCE,G. (1982), "Textual aspects of prosody in Swedish", *Phonetica* 39, 274-287

[2] HEDELIN,P.& D.HUBER (1990), "The CTH speech database: An integrated multilevel approach", *Speech Communication* 9(4), 365-374

[3] HEDELIN,P.& D.HUBER (1990), "Pitch period determination of aperiodic speech signals", *Proc. ICASSP*-90, 361-364

[4] HUBER,D. (1989), "A statistical approach to the segmentation and broad classification of continuous speech into phrase-sized information units", *Proc. ICASSP*-89, 600-603

[5] HUBER,D. (1989), "Voice characteristics of female speech and their representation in computer speech synthesis and recognition", *Proc. EUROSPEECH*-89, 477-480

[6] HUBER,D. (1990), "Speech style variations of F_0 in a cross-linguistic perspective", *Proc. SST*-90, 186-191

[7] HUBER,D. (1990), "A bilingual dialogue database for automatic spoken language interpretation between Japanese and English", *ATR Technical Report*

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