SPEECH STYLE VARIATIONS OF F_0 IN A CROSS-LINGUISTIC PERSPECTIVE

Dieter Huber

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

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: (i) orally read lists of semantically unrelated sentences, (ii) orally read narrative texts, and (iii) 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_0 parameters investigated in this study.

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 by linguists 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 contours, to establish their communicative status, and to correlate their properties with features of linguistic structure, mental computation, situational context, and speaker idiosyncrasy. Models of intonation have been proposed for a number of languages and with varying claims as to their universal applicability and predictive power. Some of these models have been implemented in the prosodic components of speech synthesis-by-rule systems, and have been subjected to systematic perceptual evaluation and assessment (e.g. Bacri 1987). Given the vast amount of research efforts and the wealth of literature on the subject, the results achieved so far, however, are inconsistent: while producing satisfactory F_0 patterns in the synthesis of isolated words and short, unrelated sentences, the prosodic quality deteriorates rapidly in text-to-speech conversion of authentic and textually coherent material, rendering an overall impression of wooden, often unintelligible, and "dreadfully monotonous reading - the same pitch contour sentence after sentence" (Umeda 1982), which listeners after a while find tedious and ultimately unacceptable to listen to.

At least three possible reasons can be proposed to account for these shortcomings:

1 - Most studies of human intonation have been restricted to the domain of the sentence as the maximal unit of linguistic processing, thus adhering to the traditional view that larger units like paragraphs, texts, and discourse, are formed by mere juxtaposition of autarchic, independently pre-arranged sentences.

2 - Intonation studies have traditionally been carried out within the framework of one or another of a number of competing syntactical, metrical, or conceptual theories, and tend to ignore features of communicative and/or informative importance that lie outside the focus of the ruling linguistic paradigm.

3 - One of the main difficulties in intonation research is 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. In order to isolate the property under scrutiny for separate investigation, while keeping all other parameters constant, researchers often employ experimental methods such as the use of reiterant speech, nonsense words and syllables, all-sonorant utterances, inverted recordings, humming, fixed carrier sentences, carefully constructed structural

ambiguities, and restrictions as to speaker type (mostly male linguists, often the experimenters themselves), listener type (mostly graduate students), and experimental conditions (explicit instructions to speak in a "natural" tone of voice, and to avoid "undue" emphasis and emotionality).

Based on these prerogatives, speech scientists have over the years accumulated an impressive body of experimental data on how properties of age, sex, personality, socio-economic status, speech rate and rhythm, segmental phonetics and phonology, syntax, semantic content, affective meaning, speaker height, weight, and attitude, and many more factors, each contribute and mutually interact with one another to create the complex patterns of F_0 we encounter when studying human speech. The almost exclusive use of "lab-speech" together with the confinement to mostly isolated and sometimes quite awkward sentences implies, however, that these data have as yet to be tested in unrestricted, unelicited and textually coherent speech produced by non-linguists under non-laboratory conditions.

There is today a strongly felt need among scientists in both theoretical and applied speech research to transcend the traditional linguistic preoccupation with syntactic structures, and to shift the focus of attention from distinctiveness in the linguistic sense to considerations of communicative function, informational content, and interactional structure. Instead of starting from the primacy of linguistic distinctivity and searching in the acoustical signal for cues that can be used to confirm or refute the validity of one or another of a multitude of conflicting linguistic theories, it may be wiser to start the other way round, viz. to search in pitch curves of authentic speech utterances for salient and perceptual relevant features in such a way that first the question is answered *whether* they occur, and only subsequently *where* and *why*. Such an outlook most necessarily take into account possible influences of context, derived from textually and semantically coherent speech material. Research in the newly emerging science of text and discourse has indicated that important aspects of sentence formation are both directly and indirectly influenced by factors outside and beyond the sentential domain. Studies of textual aspects of prosody in a variety of languages have shown among other things that the placement and magnitude of pitch accents depend on the organisation of adjacent text units, and thus cannot be reliably predicted within the limited scope of the sentence in which they occur. Prosodic boundaries have been claimed to demark discourse structure rather than syntactic units, and the functions of declination, pitch range, and stress have been suggested by several researchers to cue pragmatic rather than grammatical categories within the larger realm of textual information processing.

DATA

This study explores the differences between discourse intonation and the kind of F_0 contours typically found in isolated sentences. Three kinds of material are evaluated systematically: (i) orally read lists of semantically unrelated sentences, (ii) orally read narrative texts, and (iii) dialogues. The material has been selected from the ATR (Kurematsu et al 1989, Huber 1990) and CTH (Hedelin & Huber 1990a) speech databases and comprises equivalent samples of Swedish, English and Japanese speech. The English and Japanese dialogues consists of simulated telephone conversations conducted within the applications domain of conference registration, whereas the Swedish dialogues were conducted spontaneously, i.e. without the subjects knowing that their conversations were being recorded. 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). Approximately one minute of recorded speech per speaker and speech style was analysed. Registration of the speech samples was conducted in anechoic, sound-insulated recording studios both at ATR and CTH, using high-quality digital recording equipment (SONY DTC-1000ES and PCM-F1) set to 16-bits quantization at a fixed sampling rate of 44.1 kHz. Pitch extraction was performed using the DWAPIT pitch determination algorithm (Hedelin & Huber 1990b). Pitch estimates were obtained at 16-msec intervals and calculated to the first decimal. Segmentation of the F_0 tracings into prosodically defined *information units* (henceforth refered to as intonation units or IU) was performed following the approach published in Huber (1989). Thus, two global declination lines are 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 is reiterated every time the *Pearson correlation coefficient* drops below a preset level of acceptability. Segmentation is 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_0 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.

FUNDAMENTAL FREQUENCY DISTRIBUTIONS

Histograms of the voice fundamental frequency distributions (FFD) were computed separately for each of the ten speakers and each of the three speech styles investigated in this study. Figure 1 exemplifies the results obtained for one of the male Swedish subjects. Also listed are the F_0 means ($\bar{x}$), modes ($\bar{m}$), standard deviations (s), and the number of 16-msec analysis windows (n) containing voiced speech.

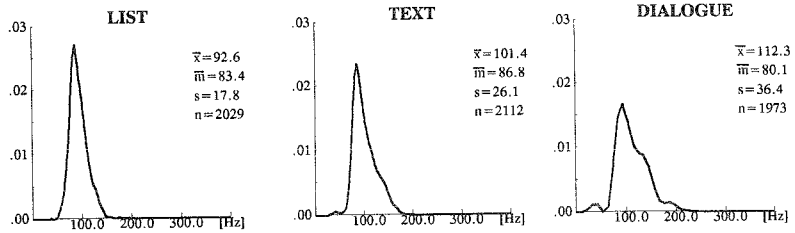

Figure 1. Fundamental frequency distributions for one male Swedish subject.

As can be seen in these FFDs, the F_0 mean and variability range values are distinctly higher both in the text and in the dialogue as compared with the sentence material. It should be noted, however, that the modal values of F_0, i.e. depicting the dominant, most frequent pitch values in the data as indicated by the peaks in the FFD histograms, are significantly lower and less variable than the calculated means. This discrepency is caused by the skewness of the F_0 frequency distribution. Clearly, the modal values ($\bar{m}$) are more representative of a speakers actual F_0 processing behaviour than the simple arithmetic means given by $\bar{x}$.

Apart from obvious interspeaker and interlanguage differences, the same global tendencies, i.e. higher average F_0 and larger F_0 ranges in the text and dialogue as compared with the isolated sentence data, can be observed for all speakers across all languages investigated in this study. Most importantly, the distinctly broader and more irregular appearance of the discourse FFDs clearly indicates a higher degree of F_0 variability that needs to be taken into account in practical computer speech applications.

In addition to the higher average F_0 and larger F_0 ranges in the discourse data, the FFD histograms depict one more apparently systematic speech style variation of F_0, i.e the existence of a small but clearly demarcated area of low F_0 values that lies distinctly below the "normal" range of pitch variability in the discourse data, but is almost completely absent in the sentence histograms. This difference reflects the use of various patterns of laryngealization, which has been discussed earlier in Hedelin & Huber (1990b).

INTONATION

Number of intonation units

A total of 586 intonation units has been established in the accumulated material for all speakers. The distribution of these intonation units per language and speech style is summarized in figure 2.

These data reveal a clear and consistent tendence, 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 large proportion of short sentences included in the material, and (2) the more frequent use of intonation units that stretch over the time extent of several complete, consecutive sentences (cf. following section).

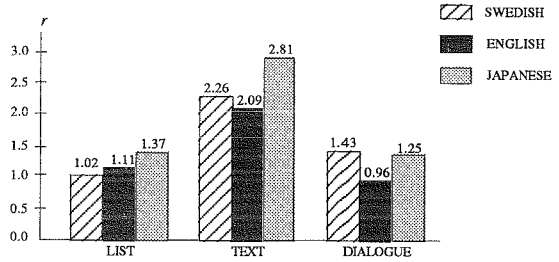

Figure 2. Intonation units per language and speech style. The bar heights r depict the ratio n/s between the number of intonation units n and the number of sentences s contained in the respective material.

Considering the higher degree of interlanguage variability found in the dialogues, it must also be remembered 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.

Time-alignment with linguistic structure

The time-alignment of the intonation units contained in the text and dialogue material with sentences (**S**), clauses (**C**), nounphrase/subjects (**SUB**), verbphrases (**VP**), complements (**COM**), adverbials (**ADV**), and parentheticals or other kinds of parallel structure (**PAR**) is summarized separately for each language in figure 3. 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. In addition to these constituent structure labels I found it necessary to include even a category (miscellaneous - **MIS**) to capture occurrences of intonation units that begin with or terminate somewhere *within* a constituent and thus cannot be classified in terms of established grammatical theory. This also includes intonation units that time-align with non-grammatical, fragmentary sentence structures, i.e. in spontaneous, unelicited speech.

As can be seen from these data, 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%). Regarding the lowest correspondence ratio found in the Swedish dialogue material (72.6%) it must be remembered once more that it consists of spontaneous speech, i.e. comprising the highest proportion of fragmentary constructions (cf. previous section).

Most commonly in our accumulated dialogue material (40.3%) intonation units correspond in a regular fashion with single sentences, 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. These figures are almost exactly in agreement with the correlation data reported by other researchers and thus confirm the general tendency reported in the literature (e.g. Altenberg 1987). 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. It must be appreciated in this context, however, that the majority of sentences associated with a separate intonation unit in the Japanese text material actually constitute single clause sentences.

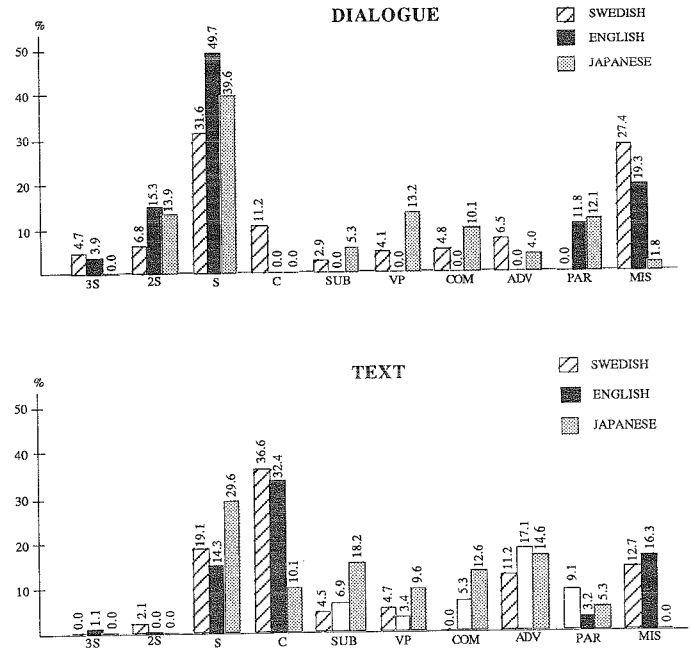

Figure 3. Time-alignments between intonation units and features of linguistic structure per language and speech style. Percentages are calculated separately for each language. The preposed *n* in category (**S**) states the number of complete, consecutive sentences covered by the time extent of one single intonation unit.

Larger structures beyond the sentence domain (i.e. **2S** and **3S**) are almost exclusively found in the dialogues, with only 1.1% 3S-occurrences in the English and 2.1% 2S-occurrences in the Swedish texts. Conversely, intonation units corresponding to single constituents in the subsentence domain (i.e. **SUB**, **VP**, **COM**, **ADV** and **PAR**) 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.

Declination line parameters

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

(1) Intonation units aligned with 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, narrow *versus* wide key, and "non-grammatical" (fragmentary) intonation units 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. However, this assumption might need to be revised in view of the distributional data for different patterns of *laryngealization* at boundary positions which have so far only been investigated for the Swedish text material (cf. Hedelin & Huber 1990b).

Intonation is a global phenomenon that stretches typically over domains larger than a single word. Intonation units as defined in this study have accordingly been described in terms of two global declination lines which approximate the trends in time of the peaks and valleys of F_0 in a linear relationship. However, a full description of F_0 contours in running speech not only has to specify the separation of these contours into intonationally coherent *chunks* that are marked by an integral melodic pattern and optionally delimited by pauses. In addition, local F_0 phenomena both within the intonation unit (accentuation, prominence, etc) and at the boundary between two adjacent IUs (juncture signals, continuation rises, onset peaks, turn taking cues, etc) need to be specified in both linguistic and phonetic terms.

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 (cf. Bruce 1982, Liberman & Pierrehumbert 1984) revealed further:

(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.

SUMMARY AND CONCLUSIONS

This study presented some differences between discourse intonation and the kind of pitch contours typically found in isolated sentences, in a cross-linguistic perspective. It has been shown that discourse intonation differs from intonation in semantically unrelated sentences with respect to practically all parameters investigated in this study. Quite obviously, an accurate model representation of the voice source that captures not only linguistic structure and speaker variability but also the inherent differences between different speech styles, is of paramount importance for all aspects of speech signal processing (analysis, synthesis, transmission, coding, enhancement, compression, etc) and computer speech applications (text-to-speech, speech recognition, speaker identification and verification, etc), and will have to account for these differences.

REFERENCES

Altenberg, B. (1987) *Prosodic patterns in spoken English*, (Lund University Press, Lund)

Bacri, N. (1987) *Perceptual spaces and the identification of natural and synthetic sentences*, Proc. XIth ICPhS, Vol.2, pp.219-222

Bruce, G. (1982) *Textual aspects of prosody in Swedish*, Phonetica **39**, pp.274-287

Hedelin, P., Huber, D. (1990a) *The CTH speech database: An integrated multilevel approach*, Speech Communication Vol.9, No.4, pp.365-374

Hedelin, P., Huber, D.(1990b) *Pitch period determination of aperiodic speech signals*, Proc. ICASSP 90, pp.361-364

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

Huber, D. (1990) *A bilingual dialogue database for automatic spoken language interpretation between Japanese and English*, ATR Technical Report (forthcomming)

Kurematsu, A., Takeda, K., Kuwabara, H., Shikano, K. (1989) *ATR Japanese speech database as a tool for speech recognition and synthesis*, Proc. ESCA Workshop on Speech Input/Output Assessment and Speech Databases, Noordwijkerhout, The Netherland

Liberman, M., Pierrehumbert, J. (1984) *Intonation invariance under changes in pitch range and length*, in: M.Aronoff & R.Oehrle (eds) Language Sound and Structure, pp.157-233

Umeda, N. (1982) *F_0 declination is situation dependent*, JASA **70**, pp.350-355