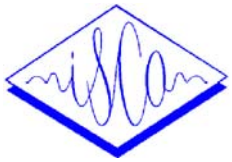


EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

CAN WE PERCEIVE ATTITUDES BEFORE THE END OF SENTENCES? THE GATING PARADIGM FOR PROSODIC CONTOURS

V. Auberg , T. Gr pillat and A. Rilliard
Institut de la Communication Parl e, ESA CNRS 5009
Universit Stendhal/INPG, Domaine Universitaire, 38040 Grenoble Cedex - France
Tel: +33 04 76 82 41 17 Fax: +33 04 76 82 43 35 Email: auberge@icp.grenet.fr

ABSTRACT

In previous works, we proposed, on the basis of acoustic analysis and synthesis, that intonation can be cognitively and linguistically described by a lexicon of prosodic contours. The aim of the present work is to validate on the French language such an approach in perception processing. Two experiments are described hereafter. The first one consists in evaluating on complete utterances the identification of the six attitudes currently implemented in the ICP TTS system. The second one is a gating experiment which aims at showing a stable and early identification of attitudes. Perception results are commented and compared to the acoustic forms to analyse the relation between the perceptive and prosodic points of unicity if any.

1. INTRODUCTION

The hypothesis which motivates this work is essentially that the morphology of intonation might be described by mental lexicons of stored contours, hierarchically organized. Each class of contours (i.e. set of variant contours) would directly carry specific linguistic values, without any phonological interface [1]. Such an hypothesis was often introduced as a possible linguistic description, for example for French, by F nagy [2], and was proposed as a cognitive processing by Cl ment and G rard [3], who even propose a specific window devoted to the perception of intonative contours. Perception experiments described below use the gating paradigm, often used in the study of lexical access, with or without priming effects. Previous experiments [4, 5, 6] and especially Thorsen [7] showed that anticipatory behaviour can be observed in prosodic tasks. Similarly to Thorsen, we focus here on the sentence level of intonation for which 6 attitudes values were selected. The first experiment evaluates listeners performance on full versions of sentences pronounced with these 6 attitudes. The results are used as a reference for the second experiment using gated signals .

2. THE REFERENTIAL EXPERIMENT

2.1. Experimental settings

These six values are surely not equivalent in terms of reference to a universal code classification, but we will not discuss here the distinction and relations between attitudes and modalities: all codes will here be referenced as attitudes.

The aim of this first experiment is (i) to evaluate if the attitudes we had retained were conventionally known and were specific to current and natural speech (ii) to measure the pertinence of the chosen definition for the explanation of the attitudes to the listeners (iii) to measure the perceptual distance between the different acoustic morphology of these attitudes. Six various values were selected, in accordance with the attitude values used in the TTS ICP system [8]. These values, mixing modalities and attitudes, are usually defined in pragmatics and didactics of French [9] as: "simple declaration" (DC) , "simple question" (QS) , "exclamation of surprise" (EX), "evidence" (EV), "incredulous question" (DI), "suspicious irony" (SC). A set of simple sentences, varying from 1 to 5 syllables, was built with strong syntagmatic constraints: each sentence consists of a nominal group (NG) varying from 0 to 5 syllables followed by a verbal group (VG) varying from 5 to 0 syllables (that means one sentence of 1 syllable, three of 2, four of 3, five of 4, six of 5). This volountary expected modulation of the contours was introduced in order to determine if the prediction effect must be, as expected, interpreted like a global effect, independent of sub-levels modulating sentence contours, or if it must be parameterized by the sub-level content. In order to avoid specific desambiguisation strategies, we did not choose homophone sentences. However some phonetic constraints were applied, and the phonetic content of sentences was very restricted. Semantic weight of sentences was chosen to be as neutral as possible as regards to the attitude values. Each sentence was recorded by one speaker with the 6 attitudes. The recording was devoid of list effects, but acoustic forms are quite coherent for each attitude, sometimes with a clear

modulation of carried contours as we already noticed for the production of such attitudes [8].

To evaluate (i) and (ii), one group of twelve listeners was trained (with corrective feed back using 12 stimuli not included in the test) and another group of twelve listeners was not. A welcoming message pronounced by the speaker was played first to familiarize listeners with the speaker's voice and intonation. He could then read the 6 code values introduced by a description of a possible situation involving such attitudes. Each acoustic stimulus could be played as many times as the listener needed. He was forced to associate each stimulus to one code from the six that were proposed to him. Each stimulus was presented twice (that means 19x6x2=228 stimuli) in order to avoid a comparison at the low level of perception loop and to ensure a code access. Each session lasts 45 minutes. Finally, the listeners answered to a list of questions asking if the definition of attitudes seemed well chosen and if these attitudes looked familiar and natural.

2.2. Results

2.2.1 Learning effects

Except for one listener we did not observe any specific listener strategy: answers are similarly shared among listeners.

Results show that, neither the linguistic structure nor the length – except on the 1-length sentences (see table 1) – influences the score.

Attitudes are quite well recognized (better than a random choice which stands at 17%). Trained subjects (Fig. 2a) have better results than untrained subjects (Fig. 2b), especially on the worst identified attitudes (globally, with training = 90 % right scores; without training = 78%).

	1-length	2-length	3-length	4-length	5-length
% ident.	67	82	83	84	85

Table 1: identication scores for each length

	DC	IQ	EV	EX	QS	SC	confusion
DC	99,6	0	0	0	0,9	0	0,9
IQ	0,4	82	0,4	8,3	3,5	25,4	38
EV	0	0,4	98,2	3,1	0	0,4	3,9
EX	0	3,1	1,3	88,6	0,4	0,4	5,2
QS	0	0	0	0	93,9	0	0
SC	0	14,5	0	0	0,4	72,4	14,9

(a) with training

	DC	IQ	EV	EX	QS	SC	confusion
DC	96,9	1,3	2,2	0	0	1,8	5,3
IQ	0	69,3	1,3	24,6	7,9	39,5	73,3
EV	2,6	0,4	86,8	0,0	0,0	2,2	5,2
EX	0,4	4,8	7,0	70,6	1,8	3,9	17,9
QS	0	3,5	0,9	0,9	90,4	1,3	6,6
SC	0	20,6	1,8	3,9	0,0	51,3	25,8

(b) without training

Table 2a&b: confusion matrix for the 6 attitudes. Results are given with percentage. The last column is the sum of the carry forward percentage, i.e. the attraction power of an attitude.

But we cannot find out if the training period teaches the code definition to the listener, or the attitude itself or, and that is the risk, prosodic contours they have never heard in speech. Their remarks following the experiment were mainly about the non-adequate definitions of attitudes. After this experiment we held a socio-linguistic enquiry (50 subjects) which confirms this interpretation [9], added to the fact that some attitudes can not be related to spontaneous speech but to reading and theatrical speech. Training seems also fast: the answers to the first presentation of a stimulus are much lower than those of the second presentation.

Table 3 confirms that the training effect can rather be related to the task learning than to an artefactual acquisition of new acoustic patterns. It shows that the trained subjects unlearn less but learn less between the first and the second presentation whereas untrained listeners dislearn more but learn more. We verified that this improvement is coherent for each stimulus: there is no exchange between learned and unlearned sentences [10].

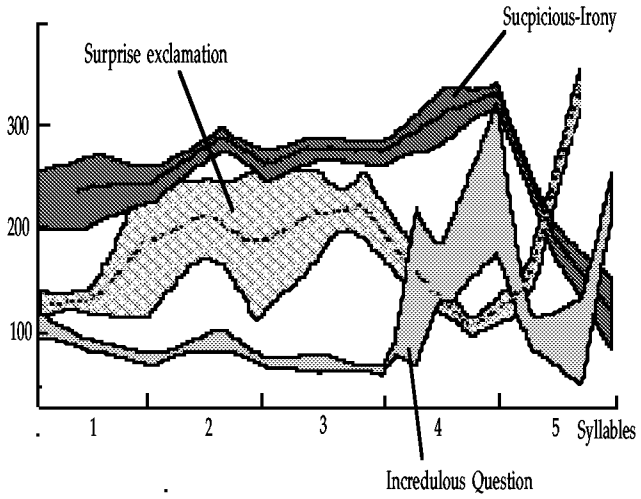
% difference 1->2	DC	IQ	EV	EX	QS	SC
with training	-1 (100)	-22 (93	-2 (99)	+18 (80)	+9 (89)	+27 (59)
without training	-4 (99)	-30 (84)	-10 (91)	+15 (64)	+12 (84)	+34 (34)

Table 3: the first number in cells is the difference between the percentage of good answers for first presentation and the percentage for second. The second number in brackets is the percentage of good answers for first presentation.

2.2.2 Error analysis

The analysis of confusions in Table 2 gives clear repartitions. The modalities DC and QS and well identified and wrong results are distributed among the other attitudes. That is the same for the attitude EV. Let us notice that QS, more EX, and

much more SC, are attracted by IQ; whereas IQ is confused with SC. These tendencies do not vary from one listener to the other but they are global tendencies. Rhythmic contours of sentences are similar for all these three attitudes. If we compare the F0 contours (fig 1), we can see that sentences have, for each attitude, varying contours depending on the syntagmatic structure (carried contours in the sense of the ICP prosodic modelization)



!Figure 1: Spaces of varying ontours for EX, SC, IQ in 5-length sentences

3. THE GATING EXPERIMENT

3.1. Experimental settings

The 3 sentences of 2 syllables and the 6 sentences of 5 syllables are selected for the second experiment to cover the two maximum syllabic lengths in the corpus. The chosen gate is the syllable. The truncated stimuli are completed by white noise of constant duration (taken 300 ms longer than the maximum length of the left-out signals in order that listeners could not capture the length of the sentence from the noise duration). The two sets of stimuli are not mixed: 6 subjects (named "normal order") listen first to the stimuli – 2L – extracted from the 2 length sentences (i.e. 3x2x6=36 stimuli) and then listen to the stimuli – 5L –extracted from the 5-long sentences (i.e. 6x5x6=180 stimuli); 6 other subjects listen to the stimuli in the reverse order (they are named "reverse order"). They are not informed of the separation of these two sets. The 12 subjects are not trained, to avoid learning effect. The protocol of this experiment is the same as the first, but since the choice is forced, we added a scale of confidence (completely sure, mean sure, not sure at all) for each answer.

3.2. Results

3.2.1 . Comparison with experiment 1

The identification score for 2L and 5L on the last gate (Table 4) are significatively inferior to the score they obtain in the first experiment (70% vs. 82% for 2L; 78% vs. 84% for 5L). Scores of 2L on the whole stimuli are inferior to the scores of 5L on the whole stimuli, which is not the case in the first experiment.

% identification	DC	IQ	EV	EX	QS	SC
2L on last gate	86	55	75	67	89	53
5L on last gate	94	69	95	55	94	64

Table 4: ident. for each attitude on the whole sentence

3.2.2. The order effect

An unexpected result is an order effect. Table 5 shows that 5L scores are not influenced by order, and that 2L are less recognized than 5L if they are heard before 5L and are better recognized than 5L when they are heard after 5L. This is particularly true for IQ, EX and QS.

% right scores	2L	5L
normal order	43	55
reverse order	61	56

Table 5: percentage of recognition for all the gates of 2L and 5L as a function of the order of presentation.

This order effect appears locally (e.g. 1/2 for SC, EV; 4/5 for IQ), it is globally neutralized (see average values on table 6).

We could think that when the listener learns global contours from 5L , he is able to use it better on 2L, and on the contrary, has greater difficulties to expand from 2 to 5 syllables. But it can not be explained only as a training effect of 5L for 2L , because the scores during the 5L and the 2L stimuli, in both order, do not increase. Perhaps it is a wrong priming implied because the 5L stimuli include 36 2-length and 36 1-length stimuli. The results (table 4) for each attitude show that the difference between reverse and normal order is bigger for EV in absolute value and proportionaly to the percentage of recognition in normal order.

	2L(begin)	5L				2L(end)
Attitude	1/2.	1/5	2/5	3/5	4/5	5/5	2/2
DC nor	61	78	89	86	92	97	89
DC inv	61	81	↘ 78	↘ 72	89	92	83
IQ nor	17	22	39	53	↘ 25	69	50
IQ inv	22	17	31	67	↘ 47	69	61
EV nor	0	19	61	81	↘ 69	92	50
EV inv	11	19	44	53	78	97	100

EX nor	22	17	48	↘ 36	42	53	56
EX inv	61	25	42	58	↘ 53	58	78
QS nor	17	8	39	50	75	94	78
QS inv	28	17	56	69	83	94	100
SC nor	11	8	36	56	56	69	61
SC inv	39	11	44	↘ 36	44	58	44
aver/norm	21	25	52	60	60	79	64
aver/inv	38	27	50	59	66	79	79
aver/tot	29	27	51	60	63	79	71

Table 6: percentage of identification gate/gate for 5L .First and last columns are first and final gate of 2L. The 5 columns inside are the 5 gates of 5L.;"norm" means normal order; "inv" means inverted order. ‖ is the limit of random order. ↘ indicates a decreasing score.

3.2.3. Prediction effect

The value "not sure" of the scale of confidence is much less used than the two others values, except for the first gate of the 2L (that is not the case for the gate 1 of the 5L).

Prosodic priming effects are clear (Table 6): more than 60 % of right identification at 1/2 (that is first gate of 2L) for DC and EXinv and 80% for DC at 1/5 (that is first gate of 5L). The differences of scores between 1/2 and 1/5 are difficult to explain from the analysis of the acoustic F0 and duration contours. Identification performance is not monotonous as a function of larger gates. As shown in table 4, 3/5 and 4/5 have sometimes a decreasing score, even if the following syllable is higher than the preceeding one. We can not explain this phenomenom by syntagmatic (lexical or syntactic boundaries) neither by acoustic similarities. The scale of confidence shows that they chose more often "not sure" and "mean sure" for the inverted 2L than for the normal 2L, and, more generally, for the 2L than for the 5L. For DC, IQ and EV 1/2 scores are inferior to 1/5 scores. In other cases, scores of 1/2 are lower 1/5 (except for EXinv). This confirms that information is not uniformly distributed but concentrated in key points, even if the contour is global (Gestalt principles).

A more precise acoustic analysis is surely necessary to point out common prefixes of the intonative forms (to discriminate the contour inside a supposed set of contours), key points on acoustic or acoustic indices related to the prediction locations (for identification). Perhaps in the morphology of intonation contours we will understand why DC is identified very early (though it is not an attractive value, since it is quite never used as a wrong "magnet" answer like IQ) and why identification performance is not monotonous.

4. CONCLUSIONS

The only hypothesis we tried to verify is the priming capabilities of listeners for intonative values relevant of a global level, that is in this experiment the sentence level for read utterances. As Thorsen [7] has shown before, we can clearly observe such a prediction effect on a complex task (6 different values), which could be an indication of a lexical organization of intonation. We do not try in this preliminary work to give a precise interpretation of results, but the influence of order and the non linearity of prediction must be discussed later. The results of this work was used for evaluating the TTS system of ICP. Consequently, after this work, it was decided to refine the definition of attitudes (SC and IQ), to change and increase the set of attitudes we initially retained [9].

Acknowledgements

Many thanks to Jean-Pierre Orliaguet, Yann Morlec and G. rard Bailly who spent a lot of time to held this experiments in the prosody group at ICP, and especially to G rard who tried to rewrite this paper in "not french english".

References

[1] Auberg V. (1991) La synth se de la parole, des r gles aux lexiques, PhD Thesis, Un. P M France, Grenoble.

[2] Fonagy Y. (1982), Prol gom nes un dictionnaire des nonc s en situation, Amsterdam Benjamins Eds.

[3] Cl ment J. et G rard J. (1996), Programmation de la production et anticipation de l'identification des formes prosodiques. Etude d veloppementale, JEP, 199-202.

[4] Grosjean, F. (1980), Spoken word recognition processes and the gating paradigm, Percept. Psychophys. 28, 267-283.

[5] Lickley et Bard (1993), Processing disfluent speech::recognising disfluency before lexical access, Eurospeech proc, 3, 935-937.

[6] Blaauw, E. (1995), On the Perceptual Classification of Spontaneous and Read Speech, PhD Thesis, Utrecht University .

[7] Thorsen, N.G. (1980), "A study of perception of sentence intonation-evidence from Danish," J. Acoust. Soc. Am. 67 (3), 1014-1030.

[9] Callamand, M. (1987), Aspects prosodiques de la communication, Edtudes de Linguistique Appliqu e, Didier Ed. Paris.

[8] Morlec Y., Bailly G. & Auberg V. (1997), Synthesising attitudes with global rhythmic and intonation contours, this volume.

[9]Moroni V. (1997) Enqu te sur les attitudes du français spontan : d finition et interpr tation, Master thesis in Language Engeenering, Un Stendhal, Grenoble

[10]Gr pillat T. (1997) Perçoit-on, par l'intonation, l'attitude d'un locuteur avant la fin de l' nonc ?, Master thesis in Language Engeenering, Un Stendhal, Grenoble