

COMPREHENSION OF PROSODY IN SYNTHESIZED SPEECH

J.P. Vonwiller*, R.W. King*, K. Stevens+ and C.R. Latimer+

*School of Electrical Engineering
University of Sydney

+Department of Psychology
University of Sydney

ABSTRACT - An experiment to determine the extent to which prosodically controlled synthesized speech can convey interactional meaning is described. The results reveal that most forms of interactional meaning can be comprehended, and that the formal basis for the definitions of meaning can be used to derive computational rules for prosody in text-to-speech systems.

INTRODUCTION

Prosody performs a number of functions in human speech. The controlled variation of fundamental pitch, duration and amplitude provide, in turn, the characteristic intonation, rhythms and stress patterns of natural utterances. These prosodic variations provide the only distinctions between some word pairs and, at a higher level, enable listeners to focus on key content words. At still higher linguistic levels, prosody conveys personal and interactional meaning and aspects of emotion.

Demonstrations of 'copy synthesis' reveal that modern speech synthesis hardware is potentially good enough for a wide range of applications. The acceptance of synthesis systems incorporating real-time text-to-speech algorithms is, however, limited largely by their relative inability to provide natural sounding and accurate prosody. Furthermore, it has been reported that the specific lack of accurate prosody detracts significantly from utterance comprehension (Waterworth & Thomas, 1985). There is thus considerable interest in computing better prosody in text-to-speech systems, and this forms the underlying engineering motivation of the present work.

This paper reports an investigation into the contribution of compatible suprasegmental prosodic features to the comprehension of synthesized speech. The model focuses on the following prosodic features: tonic pitch contour, pretonic pitch contour, pitch range, intensity and duration. These features are derived from the intonation component of Halliday's Theory of Systemic Functional Grammar (Halliday, 1985). The role of intonation in this model is to realise - and differentiate - certain semantic options, or 'principal meanings'. In the case of statements, for example, these meanings include quite subtle distinctions between 'neutrality', 'reservation', 'contradiction', 'assertion' and 'modality' (implying that the speaker is uncommitted to the utterance).

In our experiment these intonation patterns were imposed onto the output of a speech synthesizer. We examined the extent to which listeners perceived intonation to carry the same subtlety of meaning in synthesized speech as it does in natural speech. We describe the experiments and their results (including some of the problems of definition and methodology), and conclude with a brief comment on their implication for automatic prosody assignment in speech synthesis systems.

INTONATION AND MEANING

Halliday (1970, 1985) has described how intonation patterns determine the meaning of a given utterance. Halliday's system is based around the concept of the 'tonic contour', the intonation associated with the main informational focus of the utterance. Halliday recognizes a system that has a basic opposition between falling and rising pitch. Falling pitch indicates certainty, and rising pitch indicates uncertainty. The system of intonation patterns branches out from this primary distinction by

neutralising the fall/rise opposition to introduce a level pitch; and by combining the two poles to produce fall-rise and rise-fall categories. This gives five 'tone' distinctions for the pitch contour:

Tone 1 - *fall* Tone 2 - *rise* Tone 3 - *level* Tone 4 - *fall-rise* Tone 5 - *rise-fall*

Applied to simple utterances, this system assigns a basic meaning to each pitch contour. For more complex utterances, Halliday defines meanings in terms of the pitch contours for pretonics and tonics.

The meanings used in this research are variations of the major speech functions in the semantic categories of *statements*, *questions*- (yes/no and wh-), and *commands*. Within each of these three categories we have established five distinctions, giving a total of fifteen distinctive sample meanings. The neutral form represents a tone choice that would always be appropriate for expression of the particular semantic category. The other forms elaborate this neutral form, as described in Table 1.

type	meaning	tone	description
statements			
1	neutral	1	falling tonic, with flat or uneven pretonic
2	reserved	4	fall-rise tonic, plus falling pretonic
3	contradiction	2	sharp fall-rise, with falling pretonic
4	assertion	5	rise-fall tonic, with fall-rise pretonic
5	modality	3	level-rise tonic, with an even pretonic
questions			
6	yes/no, neutral	2	rising tonic, with a falling pretonic
7	yes/no, forceful	1	steep falling tonic, with mixes rising and falling pretonics
8	wh-, neutral	1	falling tonic, with falling pretonic
9	wh-, tentative	2	rising tonic, with falling pretonic
10	wh-, echo	2	rising tonic, with a rising tail, no pretonic
commands			
11	instruction	1	falling tonic, with fall-rise pretonic
12	forceful	5	falling tonic, with fall-rise pretonic, (all intensified relative to Type 11, such that the tonic rises to at least twice the pitch height, falls twice as far, and the tonic syllable is about one third longer)
13	inviting/request	3	level-rising tonic, with flat pretonic
14	persuade	1,3	falling tonic for the first part, followed by level-rising tonic in the second part, no pretonic
15	concession	4	fall-rise tonic, with a falling pretonic

Table 1. Intonation Patterns Associated with Meanings

EXPERIMENTAL DESIGN

The Synthesizer System

The aim of the experiment is to determine the extent to which listeners to synthesized speech perceive the fifteen variations of meaning described above. Our underlying interest in using the results of this work to improve prosody computation in text-to-speech systems led us to use a synthesizer designed for phonetic level input.

The synthesizer (made by Loughborough Sound Images, Ltd.) is on a plug-in PC card, and is a digital signal processor implementation of a parallel formant synthesizer (Quarmany and Holmes, 1984). In its 'formant' mode it can reproduce male human speech to very high accuracy. For this investigation, however, the synthesizer was driven in its 'phonetic' mode. The entry data for this mode is a string of phoneme codes, and duration, pitch and amplitude may be assigned to each phoneme. The synthesizer software includes a stored 'speaker table' to translate the input phonemes into formant

data frames, and performs interpolation between phonemes. The system is supplied with a 'phonetic phrase editor' for phoneme code entry and adjusting phoneme pitch and duration.

Sentence Selection

Using Halliday's prosodic descriptions for the tonics and pretonics, the fifteen meanings (as in Table 1) of the carrier sentence, "The jumper slides off the chair", were prepared for synthesis using the phonetic phrase editor. The synthesized sentences were then recorded onto a digital audio tape. The pitch contours for three of the sentences are shown in Figure 1.

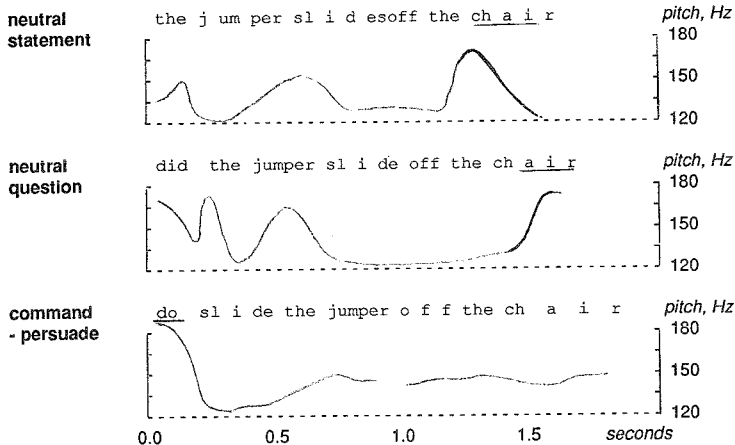


Figure 1. Pitch Contours of Three Test Sentences

The Paired Comparison Experiment

Within each of the three functional categories - statements, questions and commands - there are five different prosodic patterns. After considering the complexity of the listening and meaning labelling task, and the anticipated problems of data analysis, a paired comparison test was considered to be the most effective experimental method.

Within each functional category, the sentences were presented in pairs, and subjects were asked to judge each pair against a stated quality. For example, a *neutral* statement and an *assertive* statement were paired and subjects were asked to judge which "sounds more neutral". All forms of the statement and command sentences were tested with respect to the five prosodic patterns within that category. The question comparisons omitted that between the two neutral forms - *Wh-* and *Yes/No* - on the grounds that listeners would be confused. The presentation order of sentences within each pair was randomised with respect to the prediction that the sentence *constructed* as the exemplar of the particular prosodic pattern would be consistently *judged* as an exemplar of that type. Further, to avoid serial order and fatigue effects the sentence pairs were randomised according to the methods described by Ross (1934) and Wherry (1938). The complete experiment consisted of 58 sentence pairs.

The recorded synthesized sentences were edited together with (non-synthesized) experimental instructions onto a single audio cassette tape. Each test sentence pair was preceded by an instruction, such as, "Which utterance sounds more like the speaker is being forceful?" The two synthesized sentences followed after a pause of about 1-2 seconds, with a pause of the same duration between them. A pause of about 3-4 seconds followed before the next instruction.

Conduct of the Experiment

The subjects were University staff and students, all native Australian English speakers, who were approached indirectly through announcements made in lectures, and directly by personal contact. The sample consisted of 20 males and 21 females, the majority of whom were students, and all were naive to the experiment.

The tests were carried out in the Language Study Centre of the University. The facilities can accommodate 22 listeners simultaneously, each having his/her own cassette machine onto which the test material was downloaded from the master tape, and good quality headphones.

At the start of the test, the subjects were given a written explanation of the project and an answer sheet. They were then advised on how to use the cassette machines. Before the formal test began, the subjects were played five sample sentence pairs to familiarise them with the task. The complete test of 58 sentence pairs was one of concentrated listening, so the subjects were instructed to stop the tape after Pair 20 and again after Test Pair 40, take off the headphones and do some sitting stretching exercises for a few seconds before carrying on with the task. At the end of the test the subjects were invited to write any comments they wished to make. The test took about 14 minutes.

DATA ANALYSIS

The data obtained from the paired comparisons consisted of frequency tallies indicating the number of times the constructed pattern was judged as predicted. The data analysis sets out to test two general hypotheses:

1. That there is an association between the constructed (objective) prosodic patterns and the judgements of the prosodic patterns made by subjects;
2. That the objective ordering of sentences with respect to prosody is reflected in the frequency of responses made by subjects.

In this paper we concentrate on the analysis and results regarding the first of these hypotheses. Table 2 presents these data as percentage occurrences of association of obtained judgements of meaning for each constructed (predicted) meaning, for the three categories of sentence types.

The high values along the diagonals of these data indicate that there is substantial association between the obtained judgments of meaning and those constructed. Two statistical tests, the χ^2 test of association and the calculation of contingency coefficient, have been applied to the data. These tests substantiate that in all cases there is a significant association between the predicted and obtained score. The Mann-Whitney-U test revealed that there were no significant differences between the responses of males and females.

The second hypothesis is also supported by the experimental data. The frequencies with which perceived meanings are associated with ones other than that predicted reduces as the perceived intonation pattern become increasingly dissimilar to that of the constructed pattern.

Consider, for example, the perception of the *neutral statement*, as presented in Table 3. The frequency of a choice of a judgement decreases as the intonation pattern associated with that judgement becomes less similar to that of the neutral statement. Examination of the ranking for other sentence types reveals that the similarity between pitch gradients may also be involved in the judgement of meaning. The steeper the pitch movement the less neutral, or the more committed the meaning. These aspects of our experiment will be discussed further in a subsequent paper.

statements		% obtained				(number of
<i>predicted</i>	neutral	reserved	contra.	assert.	modal	<i>samples</i>)
neutral	78	4	5	2	10	(164)
reserved	10	47	15	11	18	(164)
contradictory	17	14	35	21	13	(164)
assertive	6	4	0	89	1	(135)
modal	12	4	4	5	76	(164)

questions		% obtained				(number of	
<i>predicted</i>	Y/N	neut.	forceful	tentative	echo	wh-neut.	<i>samples</i>)
Y/N neutral	52	20	15	13	-	-	(119)
forceful	10	64	10	7	9	-	(164)
tentative	11	12	60	10	7	-	(164)
echo	1	2	12	82	2	-	(164)
wh-neutral	-	12	4	5	79	-	(123)

commands		% obtained				(number of
<i>predicted</i>	instruct	forceful	inv/req.	persuade	concess.	<i>samples</i>)
instruction	76	9	9	4	2	(164)
forceful	9	85	1	4	0	(164)
inviting/request	10	4	60	17	10	(164)
persuade	4	5	3	84	4	(164)
concession	7	11	12	21	49	(164)

Table 2. Frequencies (%) of Associations between Obtained and Predicted Judgements of Meaning






<i>type</i>	% assoc.	tone	intonation contour
neutral	78	1	
modal	10	3	
contradictory	5	2	
reserved	4	4	
assertive	2	5	

Table 3. Comparisons of Association Frequency and Intonation Contours for the Neutral Statement

DISCUSSION

The principal result of this work so far has been to confirm that Halliday's intonation patterns do convey correct interactional meanings, even when embedded in synthesized speech of moderate quality, and in sentences tested outside an interactive context.

The subjects' written comments and the data analysis do, however, point to problems in making judgements of meaning. Some subjects registered difficulty with knowing what was meant by 'contradictory', 'inviting', 'neutral' and to a lesser extent 'persuading'. The use of these terms is to some extent context dependent; and these sentence types are amongst the lower scoring results. Some of the possible reasons for this are examined in more detail below.

Statements: 'contradictory' and 'reserved' scored lowest

- (a) The intended meaning of 'contradictory' was that "the speaker is responding to something contrary to what he expected." The pattern heard to be most like it was 'assertive', and to contradict

implies having an opinion. That 'contradictory' performed poorly across all the comparisons supports the notion that it was not well explained.

- (b) 'Reserved' was confused mostly with 'modal' and 'contradictory'. Possible interpretive problems exist on two planes: one the confusion of meaning, where to be reserved could be taken to be similar to the meaning of contradictory; the other an intonation pattern confusion since the 'modal' pattern is similar to the 'reserved' pattern.

Questions: 'yes/no-neutral' and 'tentative' scored lowest

- (a) 'Neutral' seemed to be a difficulty for 'Yes/no neutral' questions. Perhaps a pitch rise in this pattern is not uniformly expected.
- (b) 'Tentative' was confused with the 'Yes/no' questions. There is a possibility that the intonation pattern overrides all else in questions.

Commands: 'inviting' and 'concede' scored lowest

- (a) An utterance asking for a concession from someone else implies a context already understood. This may have increased the difficulties for the subjects in recognizing the difference between 'concede' and 'persuade'.
- (b) 'Inviting' and 'persuade' both have an element of asking which could lead to a potential confusion.

CONCLUSIONS

The results of this experiment indicate that interactional meaning can be comprehended from synthesized speech with prosodic control. Furthermore, the Halliday intonation patterns used in this experiment provide a systematic basis for computing appropriate prosody in many text-to-speech applications, particularly interactional ones. In such applications, determination of the information focus, and hence assignment of the tonic contour could be part of the text-generation component. Apparent confusions between intonation patterns and meanings, however, indicate that further attention is needed on two fronts. We are currently developing methods for automatic computation of prosody, and are examining the extent to which the synthesized intonation patterns should be exaggerated to enhance their dissimilarities. Improvements in future experimental methodology, by including context and better descriptors to distinguish meaning, will also be addressed.

ACKNOWLEDGEMENTS

The authors acknowledge Mr. J Taylor's assistance in preparing and editing the experimental tape, and for setting up the experiments in the Language Study Centre. The financial support of a Sydney University Research grant and the Norman I. Price Scholarship fund are gratefully acknowledged.

REFERENCES

- Halliday, M.A.K.(1975), *A Course in Spoken English: Intonation*, (Oxford University Press: Oxford)
- Halliday, M.A.K.(1985), *An Introduction to Functional Grammar*, (Arnold: London)
- Quarby, D. and Holmes, J (1984), *Implementation of a Parallel-formant Speech Synthesizer using a Single-chip Programmable Signal Processor*, Proc. IEE, 131-F, 6, 563-569
- Ross, R.T. (1934), *Optimum Orders for the Presentation of Pairs in the Method of Paired Comparisons*, J. Educational Psychology, 25, 5, 372-382
- Waterworth, J.A. and Thomas, C. (1985), *Why is Synthetic Speech Harder to Remember than Natural Speech*, 201-206, Proc. of Human Factors in Computing Systems II, (North-Holland: Amsterdam)
- Wherry, R.J. (1938), *Orders for the Presentation of Pairs in the Method of Paired Comparison*, J. Experimental Psychology, 23, 631-660