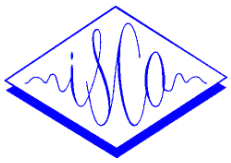

The ESCA Workshop
on Speech Synthesis
Autrans, France
September 25-28, 1990

CONTEXTUALLY APPROPRIATE INTONATION IN SPEECH SYNTHESIS

Jill House* and Nick Youd**

* Department of Phonetics and Linguistics, University College London
Gower Street, London WC1E 6BT
(also Infovox AB, Solna, Sweden)

** Logica Cambridge Ltd
Betjeman House, 104 Hills Road Cambridge CB2 1LQ

ABSTRACT

A dialogue information system under development in the SUNDIAL project incorporates an interface between a message generator and a British English synthesis-by-rule system, to ensure that knowledge relevant to the determination of prosody is passed on in the form of textual markers. Syntactic or pragmatic in origin, the markers receive their prosodic interpretation on the synthesis side of the interface.

1. INTRODUCTION

The work described here forms part of the overall design of a voice-driven dialogue information system, intended to operate over telephone lines, initially in a flight enquiries application. The voice output subsystem consists of three components: a message planner, a linguistic generator, and a synthesis-by-rule system. The message planner and linguistic generator expand a skeleton message, consisting of a dialogue act label and a number of parameters, into a pragmatically and syntactically annotated surface string, using a variety of knowledge sources: lexico-grammatical information, contextual knowledge, and knowledge of the linguistic resources available for making dialogic function explicit. This surface string is then used to drive a synthesis by rule system, where pragmatic and syntactic markers are converted into explicitly prosodic annotations. We are principally concerned here with how the annotations on it are derived and processed across the interface.

2. GENERATING ANNOTATIONS FOR THE INTERFACE

Four major kinds of annotation to the *interface string* (input to the synthesis system) are generated along with the text: focus markers, "dialogue act" markers, boundary markers, and syntactic labels. Additional markers bracket certain alphanumeric 'formulae' such as flight numbers. All are motivated by their potential prosodic repercussions.

2.1 Focus domains

Focus marking is motivated by its impact on accent placement. The belief model and linguistic history together determine **focal domains** (Gussenhoven 1984), which delineate certain subexpressions as follows:

notation	description	text annotation[1]	
[!...]	emphatic focus	#!#	
[*rf*...]	referring focus	#rf#	
[(+)...]	normal focus		(default)
[(-)...]	negative focus	#l:# ... #:l#	

Focal domains are assigned according to the status of semantic objects in the belief model, or as a result of the re-use of surface expressions. During generation, objects are assigned negative, referring or normal focus according to their status vis-à-vis the *attentional space* (Grosz & Sidner 1986) and the *attentional centre*, which is a register which maintains those objects which are the most recent of their type to be mentioned.

Emphatic focal domains may be assigned if a recent expression or subexpression from the linguistic history can

[1]To distinguish textual markers from real text, they are enclosed by # marks. By convention, single markers are normally placed to the right of their domain, while double markers surround a domain.

be adapted to achieve the semantic content currently required; those subexpressions corresponding to modifications are marked for emphatic focus. In (1), re-use of an entire utterance results in marked focus in a wh-question, where the second question may be derived from the first by substituting "from" for "to".

Dialogic function may override information status in determining focal domains. For example, the dialogue act *corrective* requires emphatic focus on the expression corresponding to the parameter in need of correction, as in (2). Similarly, the act *confirm-parameter* requires that subexpressions corresponding to parameter values in need of confirmation receive non-negative focal domains (3). For this reason, dialogue act definitions, as seen by the message planner, are accompanied by requirements for focal markings on parameters (see Table 1).

Referring focus is assigned to expressions which need to be singled out -- when negative focus is blocked, when an item from within a given set requires focus, or when there is the need to mark the functional dependency between descriptions or their referents, as in (4).

Examples (showing focal domains explicitly):
(1) Where are you travelling to ... And where are you travelling [! from]
(2) You can't fly from [! Gatwick] on Sunday
(3) From [(+) Heathrow] to [(+) Manchester]
(4) BUA flight 123 [*rf* leaves London] at seven, and [*rf* arrives in Cairo] at ten fifteen

2.2 Signalling dialogic function intonationally

There is a trade-off between the employment of melodic and lexical/syntactic resources in signalling dialogic function. For example, in (5b), clarification of dialogic function takes place at a meta-level, whereas (5a) requires intonation to render this unambiguous. A variety of strategies may be employed to determine the correct balance of intonational and non-intonational resources in signalling dialogic function. When required, a "dialogue act" marker, to be interpreted in contour terms, may be assigned to the annotated string; so (5a) acquires the label #pq# ('polar question') sentence-finally, which will ensure a rising intonation.

Example:
(5) a From Heathrow to Manchester (--> From Heathrow to Manchester #pq#)
b From Heathrow to Manchester ... have I got that right

Some examples of such markers, together with the dialogue act labels which may select them, are shown as part of Table 1. The labels themselves are mnemonic, and intended to convey constraints on possible contour assignment, which will be developed empirically.

Table 1: Dialogue act labels, marked focus domain labels and "dialogue act" markers

dialogue act	marked focus domains	"dialogue act" markers
seek-parameter	param description: (+)	#wh#
give-parameter(s)	param value: (+)	
seek-confirm-default(s)	param value: (+)	#pq#
seek-confirm-parameter(s)	param value: (+)	#pq#
summarise-task	all task parameters: (+)	#.#
corrective/suggestive	modified expression: [!]	#df#

A definitive set of "dialogue acts" has not yet been determined, but may include acts such as 'command', 'echo question', and a subset of "discourse" labels for expressions like 'well', 'OK'.

2.3 Boundary markers

Prosodic phrasing may be required ('hard' phrasing) by the nature of the syntactic structures used: for example clause boundaries, list item boundaries, and appositives. Multi-clause utterances are not discussed here. When a sequence of entities or descriptions is required, a list structure may be used, and annotated with the symbol #li#, as in (6). The underlying conceptual structure is derived from a set of database records which might answer the question: *when are there flights to paris?*, and matched against a linguistic rule which describes conjunctions as lists, with continuation markers. In (7), the flight number requires an appositive noun phrase, which will be realised as an independent intonational phrase, and so must be demarcated at both ends, using the #a:# and #:a# labels.

Examples:
(6) There are flights to Paris at five thirty #li# six thirty #li# and six fifty
(7) The next flight #a:# AF 126 #:a# leaves Gatwick at six pm

2.4. Annotating the text string

The text-string is generated using a version of the head-driven algorithm for UCG (Reape, Calder & Zeevat, 1989). This is unification based, and uses constraint satisfaction with respect to a categorial grammar representation of the lexicon. Markers, like words, may be inserted into the text string as soon as their nature and position becomes determinate. Word-class and phrasal markers (the syntactic labels), and markers indicating special domains such as flight and flight numbers, are marked in the lexicon and inserted once their respective lexical entry is irrevocably chosen.

3. FROM INTERFACE STRING TO PROSODIC SPECIFICATION

A revised model of intonation is being implemented on the Infovox text-to-speech synthesis-by-rule system for British English (Carlson & Granstrom 1986). Intonation of the text input is specified in terms of abstract intonational categories: prosodic domains, an inventory of contour types, and relative prominence specifications. Interface markers are mapped on to appropriate prosodic markers where relevant, and the resultant text string is input to phonetic realisation rules.

3.1 Phonological form

The phonological level comprises a sequence of **intonation groups**, whose hierarchical status (major, intermediate or minor) is explicit. A complete utterance must contain at least one major group. Each group will contain one or more **accent units** (House 1990), comprising one accented (stressed, pitch-prominent) syllable together with any unaccented syllables following it; the group-final unit is the **nuclear** unit. Pragmatically determined variations in accent prominence are marked, as are syllables which receive secondary (non-accentual) stress. **Contours** are applied to accent units, corresponding to **nuclear tones** on nuclear units. Multi-accent intonation groups will receive a sequence of contour specifications. The current contour inventory (expandable) includes six feature-based 'tones': fall-to-low, fall-to-mid, high rise, low rise, fall-rise and level.

Phonetic realisation rules convert the phonological specifications into F0 specifications; accent peaks are scaled relative to other accents within the intonation group, with appropriate metrically motivated adjustments, and the groups themselves scaled according to their position in the hierarchy. There is no low-level declination component (cf Ladd 87, Johnson 90). The transform to Hz values takes place at the very end.

3.2 Prosodic processing

Processing the interface markers involves a conversion to prosodic markers and/or the assignment of features to the relevant domain. Table 2 shows how selected markers (other than part-of-speech) of both types are interpreted (markers in **bold**, features in [brackets]).

Table 2: Selected interface markers and their prosodic correlates

Interface		Prosodic	
Input	Output	Input	Output
!	emphatic; **dig**	**dig (1-9)**	booster
|: :|	[-focus]	[-focus]	[-accent]
rf	primary stress; **tg**, (	primary stress	[+accent]
np	~	~	potential boundary
		~~(~)	**tg**
		tg	minor boundary (()...
cb	**ig**	**ig**	intermediate boundary (,) (;) (:)...
ss	**sg**	**sg**	major boundary (.) (?)...
li	**ig**		
a: :a	**ig ig**		
wh	.	.	[fall-to-low]
pq	?	?	[high rise]
df	,	, or (	[fall-rise] (with/without pause)
		;	[fall-to-mid]
		:	[low rise]

A fully marked example of the interface string is shown in (8), where part-of-speech markers derived from the lexicon are placed immediately to the right of each word (#fw#=function word; #n#=noun; #v#=verb; #u#=unclassified content word). These will be used to assign primary and secondary stress, and function-words will become rhythmically enclitic to neighbouring content-words. The flight no. markers #f: :f# ensure that digits and alphabetic characters are separately accented, with boosted prominence on the first and third digit.

The word 'flight' is in negative focus #l: :l#, so its prominence will be reduced to secondary stress. Remaining primary stressed syllables are assigned the feature **accent**. The appositive markers #a: :a# are associated with obligatory prosodic phrase boundaries, and are initially replaced by #ig# (intermediate group) markers. The sentence-final #ss# marker (overriding the clause boundary, #cb#) is converted to a major boundary, #sg#. Remaining syntactic #np# boundaries mark potential prosodic boundaries, unrealised here. The prosodic boundaries are in turn substituted by punctuation symbols which act as tone marks, assigning nuclear tone features to the preceding accent unit. With no "dialogue act" markers to guide contour choice, this is made by convention: #sg# becomes #.#, [fall-to-low], while the first #ig# becomes [fall-rise], #,#. The second, appositive-final #ig# may be realised as [low rise] #:#, since this can be used to echo a fall-rise in a non-contrastive context. Pre-nuclear accent units are currently assigned [level] or [fall-to-mid] contours depending on the following nuclear tone.

In (9), there is a "dialogue act" marker #df# (disconfirmation/corrective) sentence-finally, which will ensure a fall-rise nuclear tone. The emphatic #!# marker on 'Gatwick' will place the following items in negative focus, deaccenting 'Sunday', and forcing the boosted 'Gatwick' to bear the nuclear accent.

An indication of the prosodic output is shown in (8a) and (9a): accented syllables are shown in bold, intonation group boundaries by | , and nuclear tone marks in front of the relevant syllables.

Examples

(8) The #fw# next #u# #l:# flight #n# #:l# #np# #a:# #f:# AF 126 #:f# #:a# #np# leaves #v# Gatwick #n# #np# at #fw# six #u# pm #u# #cb# #ss# .

(8a) The ᵛ **next** flight | **A F 1 2 ˏ6** | **leaves Gat**wick at **6 p ˏm** ||

(9) You can't fly from Gatwick #!# on Sunday #df# .

(9a) You **can't fly** from ᵛ **Gat**wick on Sunday ||

ACKNOWLEDGEMENT

This work was supported by ESPRIT and by Infovox AB, Sweden.

REFERENCES

Carlson, R. & Granstrom, B. (1986), 'Linguistic processing in the KTH multi-lingual text-to-speech system', *Proc. ICASSP 86*, Tokyo, 2403-2406

Gussenhoven, C. (1984), *The grammar and semantics of sentence accents*, Foris, Dordrecht

House, J. (1990), 'A revised model of intonation for synthesis by rule', *Speech, Hearing and Language* 4, University College London, 121-136

Johnson, M. (1990), 'Implementation of an intonation algorithm for synthesis-by-rule', *Speech, Hearing and Language* 4, University College London, 195-226

Ladd, D.R. (1987), 'A model of intonational phonology for use in speech synthesis by rule', *Proc. European Conference on Speech Technology* 2, 21-24

Grosz, B.J. & Sidner, C.L. (1986), 'Attention, intentions, and the structure of discourse', *Computational Linguistics* 12 (3), 175-204

Reape, M., Calder, J. & Zeevat, H. (1989), 'An algorithm for generation in unification categorial grammar', *Proc. 4th Conference of the European Chapter of the ACL*, 233-240