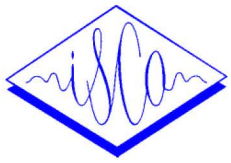

First European Conference on
Speech Communication and Technology
EUROSPEECH '89
Paris, France, September 27-29, 1989

DRIVING A SPEECH SYNTHESIZER FROM CONCEPTUAL INPUT IN THE CONTEXT OF A VOICE DIALOGUE SYSTEM

N.J.Youd
Logica Cambridge Ltd
104 Hills Road, Cambridge, UK

F.Fallside
Cambridge University Engineering Department
Trumpington Street, Cambridge, UK

ABSTRACT

This paper gives an overview of the message generation component of a voice dialogue system. This takes a conceptual representation of the message, and generates a syntactically labelled surface structure using domain independent linguistic rules. At the syllable level, features representing prosodic focus are used to constrain sentence accent placement. The output is used to drive a speech synthesizer at phoneme level.

INTRODUCTION

Voice output systems designed to deal with unrestricted text as input are necessarily limited in the prosodic decisions they can make. Beyond parts-of-speech labelling, punctuation, and flagging repeatedly used lexical items, there is not much useful knowledge that current systems are capable of extracting. The resulting prosody often represents a series of compromises. However, in situations where speech is being output from a knowledge base, or expert system, as the result of Human–Computer dialogue, more detailed semantic and pragmatic knowledge is potentially present.

The component described here is designed to form part of the Alvey VODIS (Voice Operated Database Inquiry System) system. The application domain is that of train timetable inquiries; besides database knowledge, the dialogue controller

VODIS:	where do you want to TRAVEL from
CALLER:	from London to Aberdeen
VODIS:	sorry, I didn't hear you
VODIS:	WHERE did you say you are TRAVELLING from
CALLER:	from London to Aberdeen
VODIS:	from LONDON
CALLER:	yes on Monday morning
VODIS:	are you TRAVELLING on MONDAY in the MORNING
CALLER:	yes
VODIS:	did you say you are TRAVELLING from LONDON to ABERDEEN
CALLER:	yes

Figure 1: Sample VODIS dialogue: words intended to receive prosodic prominence are capitalised

has some knowledge about the structure of conversation. Figure 1 shows a typical dialogue, with output produced by the message generator. The VODIS dialogue controller, which is designed to allow a mixture of system and user initiative, is described in [6].

This paper discusses the linguistic aspects of generation down to the segmental (phoneme) level. These include: transformation of the conceptual message into a domain-independent form, expansion of this using a unification grammar, propagation of prosodic focus and the generation of elliptical utterances. In addition, we discuss the production of output in a suitable form for driving a speech synthesizer.

LINGUISTIC GENERATION

Conceptual Representation

An example input message is shown in Figure 2. The structure can be described as a bundle of feature-labels, and values, which may themselves be atomic or complex. The 'deep case' labels relate to the knowledge representation used in the Dialogue Controller. As a first stage of transformation, the message is recursively rewritten into domain-independent *surface caseframe* form, where the feature labels relate to linguistic function. This is achieved by the use of recursive rewrite rules, similar in function to those of [5]. In general the caseframe head will correspond to the main verb of a

```
dact = yesno ,
said =[
    prot = you ,
    saffairs =[
        dest =[
            val = aberdeen ,
            emph = plus ],
        prot = you ,
        source =[
            [val=london]]]]].
```

Figure 2: Conceptual input to message generator

clause; in cases where a value for this is absent, one is inferred from the caseframe slots: currently this means that a travel verb is inferred from its arguments. For example, for the feature–value pair `[source= london]`, the translation "travel from london" is chosen; if however there is an additional feature–value pair: `[source= london, deptime= 3]`, then a different verb is chosen: "leave london at three".

Expansion into surface structure

The linguistic rules, in common with many unification-based grammar formalisms, define a mapping between (surface-caseframe) semantic structure, and surface structure, which consists of syntactically bracketted and labelled English phra-

10.21437/Eurospeech.1989-107

ses. The transformation to surface structure is defined by Definite Clause Grammar rules, which use features and feature conventions similar to those of GPSG [2]. An example grammar rule is:

```
vpadj1::
verb_phrase1([h(vp):VPd,m(adj):Ad],
             c(HD,N,FT),
             VPm^[cse-P:=Am])
     >->
     verb_phrase1(VPd,c(HD,N,FT),VPm),
     adjunct(Ad,c(_,_,-),[cse-P:=Am]).
```

This states that a **verb_phrase1** may rewrite to a **verb_phrase1** followed by an **adjunct**. Prolog variable conventions are followed. Each term representing a category has three fields: the first the syntactic subtree generated by the rule, the second syntactic features which may further instantiate the category, and the third the corresponding portion

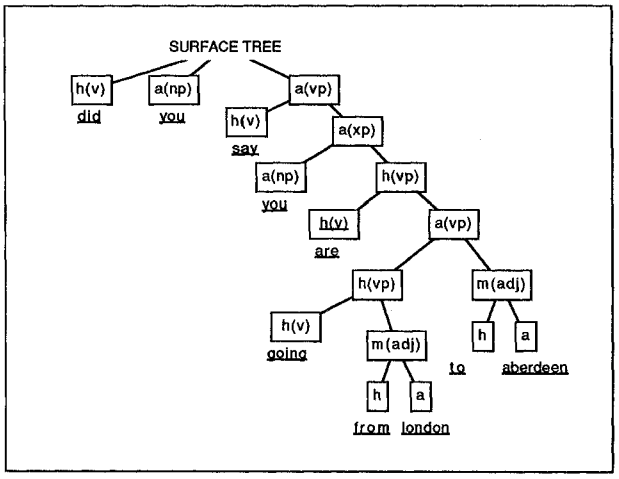

Figure 3: Surface structure corresponding to Figure 2

of semantic structure. The rules may be used in parsing as well as generation; a depth-first parser (bounded to curb left recursion) has proved useful in debugging and extending the grammar. The structure produced reflects derivation, but is somewhat flatter. Dependency labels are used to mark the nodes as *head* (h), *argument* (a), and *modifier* (m). An example of the surface structure is shown in Figure 3.

The algorithm for generating surface structure consists of simple topdown expansion: a rule is found whose category, features and semantics match with the structure to be expanded; as a result the daughters of the rule become instantiated and can be used in the next (recursive) stage of expansion. Figure 4 gives an example of two successive applications of the rule **vpadj1**.

Lexical insertion is delayed until all preterminal nodes are present, so that agreement features, which constrain lexical choice, can propagate freely in the structure. Dictionary entries are indexed by syntactic and semantic features; in addition to orthographic forms (suitable for visual output), they contain phonemic citation forms, together with syllable boundaries and primary stress.

The expansion process described above is also capable of operating in *ellipsis mode*, when the dialogue requires that

```
***EXPAND: rule vpadj1::
verb_phrase1([pred:=go,cse-to:=dundee,cse-from:=
-->                                          london])
verb_phrase1([pred:=go,cse-to:=dundee])
adjunct([cse-from:=london])

***EXPAND: rule vpadj1::
verb_phrase1([pred:=go,cse-to:=dundee])
-->
verb_phrase1([pred:=go])
adjunct([cse-to:=dundee])
```

Figure 4: Expansion of $[pred := go, cse-to := dundee, cse-from := london]$ prior to lexical insertion

a previously-said utterance is repeated (eg for purposes of confirmation). In such cases, the message consists of the semantically relevant part, but structure is also given corresponding to the previous utterance. This latter is expanded, but only from the lowest node whose semantics covers the intended message. Side branches of this which do not contribute to the message may then be discarded.

A selection of orthographic forms corresponding to utterances generated with the grammar rules, is shown in Figure 5.

GENERATION OF PROSODY

This has three aspects: boundary placement, accent placement, and selection of accent types. The placement of phrase-internal prosodic boundaries, according to criteria of phrase length (in syllables), and structurally licensed boundary positions, is described in [10]. Accent placement takes place at two stages: at the word and syllable level.

Word level accent placement

Nodes in the input message are optionally marked for prosodic focus, which can take values *emphatic*, and *given*. Any such markings are carried down to corresponding nodes on the surface structure. Terminals of the surface structure (words) are assigned one of the following values for the feature *focus*:

- where do you want to go from
- from london
- in the morning on monday
- going to dundee from london in the morning on sunday
- when are you leaving on sunday
- to where
- did you say to aberdeen
- what day did you say you want to go to aberdeen on
- the only train on a monday is in the afternoon at three
- did you say going on monday or friday
- is the only train on a monday in the afternoon at three or is there another train from london in the evening at five

Figure 5: Sample outputs (orthographic form) produced by the system

++	emphatic sentence accent
+	normal sentence accent
-	no sentence accent
?	sentence accent unassigned

Word accent assignment takes place by propagating values for *focus* down through the surface structure. The propagation is guided by dependency relations between sister nodes, based on the notion that the arguments, not the head, are significant in determining broad focus on a constituent [7]. Focus values '+' or '++' only appear on function word heads by 'default accenting', in cases where neighbouring argument constituents are marked as *given*. In this way 'default accent' [4] can appear on function words, as in the phrase: "you want to go TO london". A more detailed account of focus propagation is given in [10].

Syllable accent placement

For a word assigned focus values '++', '+', or '−', the focus marking is carried onto the syllable marked for primary stress in its citation form. Other syllables, including the primary stress of words marked '?', receive accents as following:

> assign the value '+' if the previous n syllables were $[-focus]$, and the next syllable is $[-focus]$ or $[?focus]$, where $n = 2$ in the context of content words, and $n = 4$ for function words.

In addition, *emphasis levels*, in the scale $0.1 \ldots 1.0$, are assigned to syllables marked $[+focus]$ or $[++focus]$ in a manner which gives overall prominence to nuclear (phrase-final) accents, and alternating prominence to prenuclear accents.

GENERATING SEGMENTAL OUTPUT

It is necessary, in order to be able to manipulate the $F0$ and timing contour, to write to the speech synthesizer at the phoneme level. The words are linearised into an array of syllables, at which stage accent assignment takes place (see above). This then is further expanded into an array of phoneme symbols, based on the citation forms. Context-sensitive rewrite rules operate at word boundaries, to handle phonological processes such as place and voicing assimilation. In addition, reducible vowels in unstressed syllables may be rewritten to *schwa*. A duration contour is then assigned, based on the rules of Klatt [1].

Production of the pitch contour follows closely that described in [8]. An abstract contour is viewed as a chain of pitch-accent (PA) templates, time-aligned with the vowels of the corresponding syllables, and scaled according to emphasis levels. A log transformation converts the abstract contour to a Hertz scale; PA templates are joined by linear interpolation, and smoothing is applied to the result. A graphical example of output to the synthesizer is shown in Figure 6.

DISCUSSION

The VODIS message generator described here is concerned with the transition from a conceptual message to a phonemic encoding of the English utterance. Particular attention has been given to issues regarding the generation of prosody. task. In addition, topic boundaries may be signalled by pitch range [3,8]. As regards contour, an informal investigation of recorded dialogues confirms the view that there is at best an indirect correlation between the final nuclear tone in a move, and that move's communicative function. In extended interactions, for example, information providers do not always terminate responses with a falling tone, presumably because this could be interpreted as an unwillingness to continue the conversation. It is hoped that further analysis of human-human conversations will throw more light on this.

ACKNOWLEDGEMENTS

This research was carried out while the first author was at Cambridge University Engineering Department. The authors would like to acknowledge the Alvey Directorate, who funded this research; and British Telecom Research Laboratories and Logica Cambridge, for their collaboration and help.

References

[1] Allen, J., Hunnicutt, M.S., & Klatt, D. (1987) *From text to speech: the MITalk system* Cambridge University Press.

[2] Gazdar, G., Klein, E., Pullum, G. & Sag, I. (1985) *Generalised Phrase Structure Grammar*, Basil Blackwell.

[3] Hirschberg, J. & Pierrehumbert, J.B. (1986) "The intonational structuring of discourse", in *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics.*

[4] Ladd, D.R. (1980) *The structure of intonational meaning: evidence from English*, Indiana University Press.

[5] Mellish, C. (1988) "Natural language generation from plans" in *Advances in Natural Language Generation*, (Zock & Sabah, eds). Pinter:London.

[6] Proctor, C.E. & Young, S.J. (*forthcoming*) "Dialogue control in conversational speech interfaces".

[7] Selkirk, E.O. (1984) *Phonology and Syntax*, MIT Press.

[8] Silverman, K. (1987) "The structure and processing of fundamental frequency contours", PhD thesis, University of Cambridge.

[9] Terken, J.M. (1985) "The use and function of accentuation: some experiments", PhD thesis, University of Leiden.

[10] Youd, N.J. and Fallside, F. (1987) "Generating words and prosody for use in speech synthesis", *Proceedings of the First European Conference on Speech Technology*, pp. 17–20.

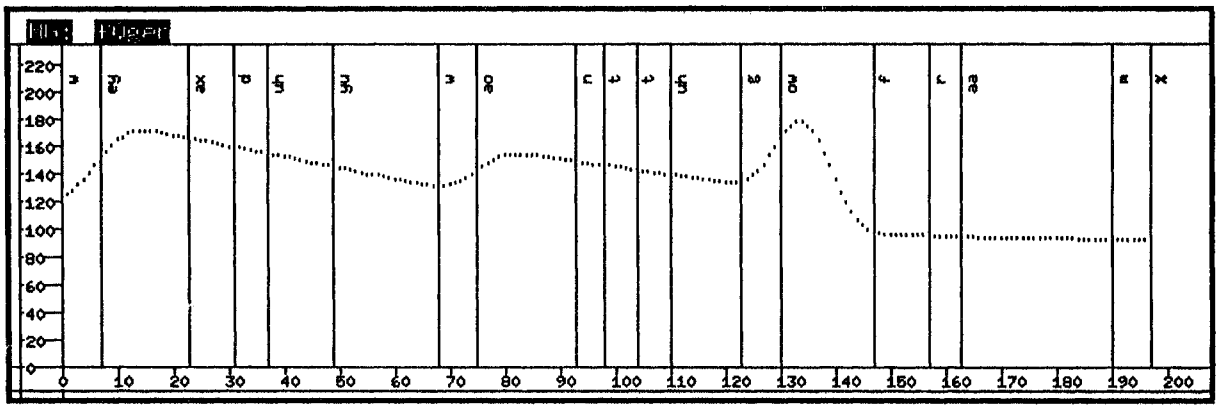

Figure 6: Segments (Klatt symbols) and F_0 contour for the utterance: "where do you want to go from"