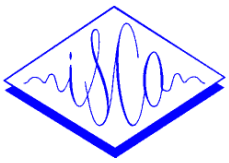


3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

EVALUATING SYNTHESISED PROSODY IN SIMULATIONS OF AN AUTOMATED TELEPHONE ENQUIRY SERVICE

Jill House(1), Catriona MacDermid(2), Scott McGlashan(2), Andrew Simpson(2), Nick Youd(3)

(1) University College London, UK (2) University of Surrey, Guildford,UK
(3) Logica Cambridge Ltd., Cambrige, UK

ABSTRACT

Dialogue material has been collected using bionic Wizard-of-Oz simulations of an automated voice-driven dialogue system; system turns, using rule-based speech synthesis, were systematically varied for both text and prosody. In a separate pilot experiment, subjects evaluated the system turns on a number of parameters. While they did not always clearly differentiate between the different conditions, they seemed to associate the "enhanced" prosody, preferred in an earlier laboratory experiment, with a more likeable personality.

Keywords: prosody, synthesis, dialogue, evaluation, simulation

1. INTRODUCTION

In the voice-operated dialogue system being developed in the SUNDIAL project, concentrating on a "flight enquiries" application, voice output for British English is provided by the Infovox text-to-speech synthesis system. The prosodic component has been adapted and expanded to optimise it for interactive dialogue as opposed to reading aloud. The selection of prosodic patterns is informed by the grammatical and contextual knowledge available to the system. In generating the output text, annotations encoding such knowledge are inserted in the text string which is passed to the synthesiser, where the knowledge encoded in the annotations is interpreted prosodically, in terms of phrasing, accent placement, degree of emphasis and pitch contour [1,2]. These interpretations are grounded in our observations of natural speech dialogues of a comparable type [3]. Our goal has been to improve the acceptability and intelligibility of the synthesised voice by making the prosody appropriate to the discourse context.

In SUNDIAL, the synthesis system may be driven by a full linguistic generator, which generates annotated text on-line, or alternatively by a template generator, in which case annotated text strings are stored within the templates. In the experiments described below, the template generator was used, partly for reasons of robustness in terms of current system capabilities, and partly because it allowed us more readily to control the different text and prosody conditions represented in the tests.

2. EVALUATION

2.1 Evaluation criteria

Continuous evaluation of different aspects of the system has been essential in the course of SUNDIAL's development. Evaluating the output messages is a non-trivial task, requiring us to define the criteria by which such an assessment might be made. We should firstly expect the majority of messages generated by a knowledge-based system to be linguistically correct and pragmatically appropriate. The success of the system's utterances may then be measured in at least two ways: *attractiveness* to the caller, and *directive potential*. Attractiveness will include, for example, whether the voice sounds pleasant and cooperative; directive potential relates to the system's success in guiding the caller towards giving the kind of information and using the kind of language that the system knows how to deal with. Prosody may be seen as making a contribution in both areas, but is difficult to assess independently of textual and other variables.

2.2 Measuring prosody in a laboratory experiment

A first attempt at an objective evaluation of the prosody, using a controlled experiment in which listeners rated the appropriateness of synthesised utterances within a dialogue context, was reported in [4]. There was a significant overall preference for utterances produced from an enriched prosodic representation over those using simple text as input, and default text-to-speech (TTS) rules. In particular, the enhanced prosody had a marked advantage when the context required the system to focus narrowly on some part of the message, or where the system needed to correct or modify the (human) caller's assumptions. In such cases, the contextual information encoded in the annotations was able to ensure an appropriate prosody. The test used exchanges abstracted from artificially constructed dialogues, allowing for a very precise control of contexts. Findings about the usefulness of dialogue-specific prosody were limited by the precise, controlled nature of the experiment; in the context of a fully working system, in which the dialogue develops unpredictably, further experiments were needed.

2.3 System simulations

In the course of its development, the output sub-system has been regularly tested using "Wizard-of-Oz" (WoZ) simulations [5,6]. Caller subjects, who interact with the system by telephone, are given a task to perform (such as establishing

the arrival-time of flight x from source city y), and assume that the system is fully automated. Caller behaviour may thus be thought to approximate to what it would be in the context of a fully working system. System responses are in fact controlled by the wizard, who selects messages from a set of template utterances, designed to reflect the language generator's capabilities. For added realism, in many dialogues the wizard simulates misrecognition of the caller's utterance, eliciting repair sequences, and sometimes simulates complete failure of the system to repair the dialogue. Most of the WoZ simulations run to date have been restricted to unenhanced (default TTS) prosody.

2.4 The task

The pilot experiment described below represents an attempt to incorporate prosodic evaluation into the simulations. Our objectives have been to obtain some measure of the acceptability of the synthetic voice, and specifically to investigate the possible contribution of an enriched prosodic representation to the overall acceptability of the synthesised voice output in a working system. Noting that the contribution made by prosody may be more marked when other linguistic cues are not present, we also aimed to investigate how the prosody interacted with both "terse" and "verbose" text versions of messages generated in comparable contexts.

While we have attempted to collect responses in a quantifiable framework, we have been aware of the difficulties of ensuring statistical reliability, particularly in view of the small sample size available to us at this pilot stage. We have also therefore taken care to gather qualitative information, in the shape of comments and observations from system users and experimental subjects. Informal insights of this kind may be invaluable to a system developer.

3. EXPERIMENTAL METHOD

There were a number of stages in the current experiment, designed to assess both directive potential and attractiveness. Directive potential can be inferred to an extent by observing user behaviour, but the experimenter cannot always be sure what the subject thinks the system is trying to do. Questions of attractiveness need to be dealt with off-line, e.g. through some kind of post-session questionnaire, since a subject caller cannot be expected to focus on such issues while interacting with the system. Recordings of simulation dialogues have therefore been used as data to be evaluated in a separate evaluation task.

3.1 Labelling utterances

Utterances generated by the system are assigned various "dialogue act" labels (*inform, request, confirm...*) by system developers, reflecting the task which they expect them to perform in the dialogue. One measure of directive potential must be the extent to which a user's perception of an utterance's intention concurs with that of the developer. As part of our experiment, it was decided that subjects should select from a list of possibilities a suitable label for each system turn they heard. However, we could not be sure that the sometimes technical labels used by developers would be those favoured by naive users. A preliminary step was therefore to discover what labels naive subjects would assign to a range of dialogue turns. Six subjects listened to five previously recorded WoZ dialogues, and described what they thought the system was trying to do at each step in the dialogue. They were also encouraged to describe what they thought the system sounded as though it was doing (ignoring the dialogue context). A wide range of labels was collected, from which a **core** list, based on frequency of use, was chosen for the evaluation task (see Section 3.4).

3.2 Adapting the template generator

The template generator was adapted to allow us to generate both prosody and text under two conditions. Prosody in the output speech was either *standard*, based on an orthographic text input (relying on default TTS rules), or *enhanced*, using the annotated representation. The text itself was either *terse* or *verbose*. This yielded four conditions in all:

ET	enhanced prosody + terse text
EV	enhanced prosody + verbose text
ST	standard prosody + terse text
SV	standard prosody + verbose text

The template utterances included were designed to reflect the current capabilities of the system.

3.2.1 Text

The verbose text condition generated utterances which were grammatically complete sentences, fully explicit, and always including "please" when making a request. In the terse text condition these messages were reduced to include the ellipses which one might expect from a human speaker. For example, when seeking confirmation of a particular parameter in the verbose mode, the system might say:

"was that from Edinburgh"

or in the terse condition:

"from Edinburgh"

Similarly, if checking on a complete task, a verbose message might be something like:

"you want to know the arrival-time of flight BA 153 from Madrid"

while the terse message would omit the words: "you want to know". Prosody is potentially important in distinguishing between a statement of confirmation and a request for confirmation.

3.2.2 Prosody

The enhanced prosody templates were prepared by hand, using annotations consistent with the desired output of a fully rule-based linguistic generator. For example, intonational phrasing and pitch contour were specified, and focused parameters highlighted for emphasis. Realisation in synthesis was entirely rule-based. Extensive use was made of an intonation pattern involving a *high fall-rise*, which had been frequently observed in human-human dialogues, and which seemed to signal politeness and some deference, as well as encouraging a response from the hearer. It was, for example, used on all the system's acts of confirmation, as described in Section 3.2.1, in both the verbose and terse conditions. The constraints imposed by the template generator and the simulation set-up made it difficult to include certain types of utterance, notably ones involving a shift of focus, realised through accent-placement, and often involving corrections to caller misconceptions; for these types, prosody had been shown to play an important role in the earlier test. We were

therefore looking at utterance types where the advantage of enhanced prosody, though significant, had been relatively small.

In preparing the standard prosody templates, using orthographic text input, a decision had to be made about punctuation. This was particularly relevant for acts involving confirmation (see Section 3.2.1). TTS rules assign a falling intonation to stretches of text terminating in a full stop ".", and a rising intonation to those ending with "?" (except in wh-questions, not relevant here). Since both types of intonation have been observed by human speakers in such contexts, we adopted the strategy of using "?" after utterances which in their verbose form were syntactically interrogative: "(was that) from Edinburgh?", and "." in cases where the verbose syntax was declarative: "(you want to know) the arrival-time." Examples of confirmation requests in all conditions are shown below:

SV	Was that from EDinburgh?	HR
ST	From EDinburgh?	HR
EV	Was that from EDinburgh	^FR
ET	From EDinburgh	^FR
SV	You want to know the arRIVal-time.	LF
ST	The arRIVal-time.	LF
EV	You want to know the arRIVal-time	^FR
ET	The arRIVal-time	^FR

HR = High Rise; ^FR = high Fall-Rise; LF = Low Fall

Tones are realised on the capitalised syllable. Enhanced input text uses annotations as described in [1].

3.3 Collection of simulation data

The Bionic WoZ simulation incorporated the Infovox text-to-speech synthesizer, driven by an X11 R4 interface written in Prolog. The interface, run on a UNIX workstation, allowed the wizard to simulate recognition errors and manipulate the synthesiser's use of enhanced prosody and text. Ten pictorial flight-enquiry scenarios were selected, based on actual flight enquiries to British Airways. For each scenario, subjects telephoned what they believed to be the system and tried to obtain the relevant flight information. Ten subjects participated, each using the same ten scenarios but in counterbalanced order. Text and prosody enhancements were also counterbalanced across subjects.

To generate output, the wizard simply selected an utterance template from a list of keywords on the screen, using a mouse. The task parameters for the scenario were inserted automatically by the simulation software, and the utterance was then synthesised using the TTS synthesiser. In this way, 100 dialogues were recorded on audio tape using an analogue recorder and digitalized on a Macintosh computer, to provide the basis for the evaluation data.

3.4 Interactive evaluation

3.4.1 Selection of dialogues for evaluation

In selecting dialogues for the evaluation task, we had to ensure coverage of all text and prosody conditions, but without overloading subjects. Eight dialogues were selected in all, two for each condition. Some care was taken to ensure that there was a spread of scenarios and caller subjects (none was included more than twice), that the dialogues were of a manageable length (they contained an average of 21 individual turns), and were of comparable complexity.

3.4.2 The evaluation tool

The tool was designed for post-hoc evaluation of selected simulation dialogues. Using a mouse-driven menu system, subjects listen to the dialogues turn by turn. After each system utterance they were required to make a judgment of directive potential by selecting an appropriate label from the 'core' list established in the labelling task, answering the question: "Which of these categories best describes what the system said?". If none of the labels (*greeting, request, confirm...*), or more than one, was considered appropriate, the subject selected "other", and his or her own labels were noted down. For each system turn, the subject was also asked: "How natural did it sound?", and rated the utterance by selecting a point on a horizontal bar whose extremes are labelled "unnatural" and "natural".

When each turn in the dialogue has been covered in this way, the subject then heard the complete dialogue again, and gave overall ratings for *acceptability*, *voice quality* (on a scale ranging from "mechanical" to "human"), *the way it was said (accent)* (scale: "unlike English" to "like English"), *intelligibility* (scale: "not clearly spoken" to "clearly spoken"), *personality* (scale: "unlikeable" to "likeable"), and *pleasantness* (scale: "unpleasant" to "pleasant"). In each case, the subject used the mouse to mark a horizontal bar representing the relevant scale, as with the naturalness ratings.

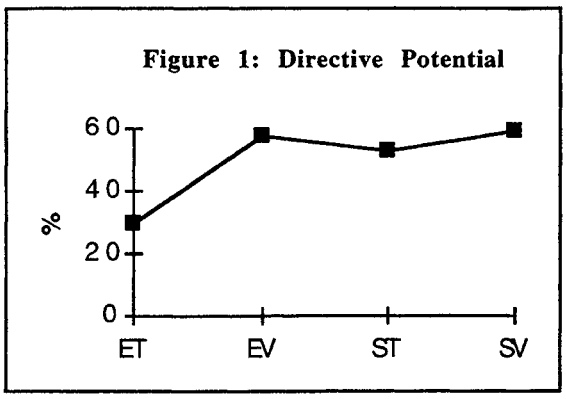
3.4.3 Running the evaluation tool

Six native English speakers participated, three male and three female. They were aged between 21 and 38 and had not participated in any WoZ simulations. They were paid for their participation. As described above, they were asked to evaluate each of the eight dialogues selected in terms of how the system utterances sounded to them, first making turn-by-turn judgements on directive potential and naturalness, then listening to the complete dialogue and making global judgements on voice quality, intelligibility, personality, etc. Each subject heard the eight dialogues in a different order, to avoid practice effects from adapting to the synthetic speech. They assigned labels and ratings using the evaluation tool, and gave additional comments verbally to the experimenter.

4. RESULTS

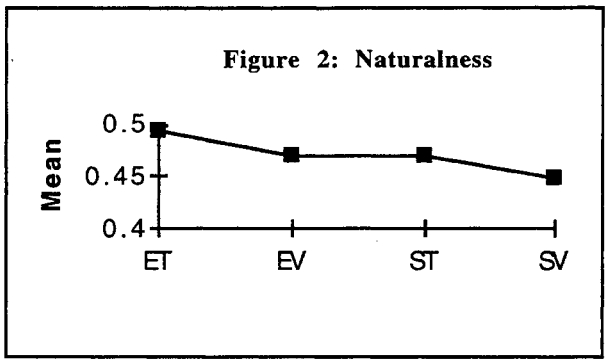
The evaluation tool automatically logged the labels assigned by subjects and their ratings. This data was collated and analysed in terms of the 4 conditions described in Section 3.2.

For each turn, the labels assigned by subjects were compared with one another to determine consistency of labelling. If more than 50% of subjects assigned the same label, then the directive potential was rated as "high". For each condition, the percentage of dialogue turns with high directive potential was calculated. The results are shown in Figure 1.



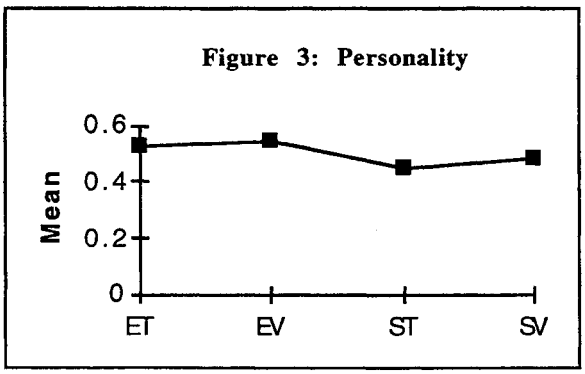
The SV, EV and ST conditions correlate with high directive potential (59.3%, 57.1% and 52.7% respectively), while the ET condition is significantly lower (29.1%).

The mean subject ratings for naturalness across the conditions are shown in Figure 2.



Naturalness ratings contrast with the directive potential results: the ET condition has the highest rating (0.49) and SV the lowest (0.44).

The mean ratings for the global metrics were also calculated for the 4 conditions. The ratings for acceptability, voice quality and accent were very similar for each of the conditions. For example, accent ratings ranged from 0.58 to 0.6. However, pleasantness and personality ratings marginally favoured the enhanced prosody conditions as illustrated with the results for the personality metric in Figure 3.



5. DISCUSSION

The results of our experiment fail to support the hypothesis that enhanced prosody gives a significant advantage to speech synthesis in this spoken dialogue system. If we regard consistency of labelling as an accurate reflection of directive potential, then the results suggest that the verbosity of the system utterances plays a more central role in allowing callers to determine the intentions of the system. The naturalness and pleasantness results, on the other hand, suggest that enhanced prosody is at least conveying the impression that the system has a more likeable personality.

Earlier WoZ simulations have not tested the degree of text verbosity, though both types of prosody have been used in more recent simulations. Caller subjects have not always explicitly noticed this prosodic difference, but where they have, they have rated the enhanced condition more highly on scales of clarity, naturalness and pleasantness. Some have described the enhanced condition as "less robotic".

There are a number of reasons why the results of this experiment do not suggest the clear preference for enhanced prosody found in the laboratory experiment. In the first place, the behaviour of the simulation software closely mirrored the behaviour of the current SUNDIAL system so restricting the type of utterances realized. In the second place, the acoustic quality of the synthesised speech was restricted to the telephone bandwidth, and subject to the uncontrollable variability of telephone line quality. Both factors made it more difficult for subjects to differentiate between the conditions. We must also recognise that the range of dialogues tested was rather limited, and that the unpredictability of true simulations makes it difficult to ensure that we are comparing like with like for all conditions.

Acknowledgements

The research was supported by CEC ESPRIT project 2218, Speech UNderstanding and DIALogue (SUNDIAL), and by Infovox AB, Solna, Sweden.

REFERENCES

[1] House, J. & Youd, N. J. "Synthesising intonation in a dialogue context", *Speech, Hearing & Language* 5, University College London, 77-89, 1991

[2] Youd, N. J. & House, J. "Generating intonation in a voice dialogue system", *Proc. Eurospeech 91*, Genova, 1287-1290, 1991

[3] Youd, N. J. "The production of prosodic focus and contour in dialogue", PhD dissertation, Open University, 1992

[4] House, J. & Youd, N. J. "Evaluating the prosody of synthesised utterances within a dialogue system", *Proc. ICSLP*, Banff, Alberta, 1175-1178, Oct 1992

[5] MacDermid, C. "Features of Naive Callers' Dialogues with a Simulated Speech Understanding and Dialogue System", *Proc. European Conf. on Speech Technology*, Berlin, Germany, 1993 (this issue).

[6] Fraser, N. M. & Gilbert, G. N. "Simulating Speech Systems", *Computer Speech & Language* 5, 81-99, 1991.