Proceedings of the Institute of Acoustics

MACROPHONOLOGICAL SIGNALLING OF TEXT STRUCTURE AND AUTOMATED SPOKEN DIALOGUE

James Monaghan (1). Christine Cheepen (2)

(1)University of Hertfordshire
(2)University of Surrey

1. INTRODUCTION

There is at present a very marked growth in the number of commercially implemented automated online information systems. The development of the bulk of these systems has been technology driven, and has sought to fill market niches created by the fundamental societal and economic changes of the last few years. On the one hand, as accessibility and usability of telephones and IT in general have increased, so the applications have proliferated. However, the downside is that new systems tend to be designed on the basis of old ones and, unless you continually review underlying principles you tend to be hamstrung by elaborate fixes to technological hurdles that may be no longer relevant. For instance, automated interactive speech-based systems have been heavily influenced by the limitations of such system elements as the speech recognition components and the amount of system resources needed to generate machine responses of an acceptable level. The ways of delivering system responses whether by speech synthesis or recorded speech became the focus of attention. In doing this, the designers tended to concentrate on low level features of speech such as phonemes and diphones rather than the high level regularities which are so important to humans in spoken communication situations.

It is obvious, but easy to forget, that human beings do not process speech like computers. Computers scan an incoming speech signal for probable items and ultimately match the predictions with a vocabulary list. The grammatical, semantic and pragmatic processing of text by computer is still in its infancy. On the output side, most commercial systems attempt to give naturalness to text initially written. This is fairly good at the segmental level, but less convincing at the prosodic level, where 'tone of voice' is signalled.

MACROPHONOLOGICAL SIGNALLING OF TEXT STRUCTURE

Human-human spoken dialogue is completely different. The I/O phase is relatively unimportant compared with the generation of information. What makes language processing in general and speech processing in particular so difficult is that human beings, when speaking to each other, get as near as they ever do to sharing thoughts. While both participants monitor all aspects of each other's behaviour - non-linguistic as well as linguistic - they are internally shadowing each other in terms of the range of outcomes that are likely to come next. Linguistic 'noise' has little effect on the high level processing.

Basically, spoken interaction is a collaboration between the participants with a series of agreed goals - both implicit and explicit. Crucially, the active role of speaker and the apparently passive role of listener and constantly being exchanged dynamically in real time. They are, in fact, interleaved. As the listener is receiving the speaker's contribution he is simultaneously planning his response, changing it as his understanding of the incoming message develops. Dialogue is essentially semiogenic - the meaning of what happens is not precisely defined in advance. Even in the kind of business dialogues we are interested in, where outcomes are relatively specific, there is still room for renegotiation of goals.

For instance, everyday conversation is not 'about' any particular topic (Cheepen 1988), while dialogue designed to retrieve specific information usually starts with a strict attempt at a specification of detailed goals. If I meet you socially I might ask "How are you?", "What do you think of the play so far?", "What have you been up to since I last saw you?" etc. Conventional, vague or noncommittal answers - such as "Mustn't grumble!", "Interesting!", or "Not a lot!" - are perfectly acceptable. In contrast, if I want information, I need to specify what it is in some detail. I might ask "What's the time?", "Could you direct me to Windermere" or "What's the best computer on the market?". The responses must be relevant and specific. However, more negotiation may be needed. The above questions could reasonably, in certain circumstances, be responded to by "GMT or at JFK", "Are you driving?", or "What do you want it for?"

2. SEMIOGENIC MACROSTRUCTURES

The collaboration between participants in dialogue requires three semiogenic components - the ideational, the interpersonal and the textual (Monaghan 1979) - and the differences between the sorts of dialogue described above relate to the importance of the contribution each one has to a particular dialogue type. In business transactions the ideational predominates, although the interpersonal is an important factor in getting the right impression across. In conversation, the interpersonal dominates. A conversation is 'about' the participants. The textual is always essential to avoid incoherence.

The ideational component, then, is about facts and ideas, about what is and is not the case. Although it is the central aspect of the meaning of the kind of automated information systems we are interested in, we will not place it at the centre of our attention in the present paper, because the majority of the ideational signalling is carried by grammar and vocabulary. For more details on these aspects, see Cheepen & Monaghan (1998). We will, however, refer to it where it is relevant for phonological macrostructures.

The interpersonal is where the mutual roles of the participants are generated, changed and confirmed. In the study of conversation, we highlight the interpersonal features where speakers establish or reinforce a long-term relationship through the speech activity. In business transactions, however, the crucial roles are to do with instructions and questions. In automated dialogues, the system is programmed to recognise the relevant kind of questions it is designed to answer, and to respond to them. In other words, after the welcome message, the system invites questions. As a fall-back position, there are a series of more generalised questions which the user is asked to select from to get the dialogue back on track.

Here the importance of the interpersonal comes to the fore. *Asking* and *giving instructions* are interpersonally charged activities because they are associated with power relationships, in this case based on the possession of information. If the system is not behaving in what seems to the user as a sensible fashion then he will hang up. Even the wrong tone of voice can irritate an enquirer who is in a hurry. It is not as obvious as it may appear that a 'natural' sounding voice is the best (Williams & Cheepen 1998), but it is

certainly an area worthy of study. Ideally, a system able to monitor and react to user tone of voice would be an asset, but we are still a long way from that.

The textual component is where the coherence of the dialogue is generated. It is only the focus of attention in poetry, some kinds of jokes etc., but without it functioning properly the whole dialogue breaks down. In any but the simplest of utterances, the information structure signalled by stress, intonation and rhythm point out the important information and what the state of knowledge of the participants is. In most written texts, and in the highly structured dialogues underlying most business transactions, the author provides the textual structure. In spoken dialogue this is normally more interactive, and in automated systems more constraints apply again.

3. MACROPHONOLOGICAL STRUCTURES

Macrophonological features are of several types. From a linguistic point of view, some are fully accepted as part of grammatical description, including stress, intonation and rhythm, and others are paralinguistic, like pauses, loudness and individual voice quality.

Intonation has several functions. At the highest level it organises the text into paratones. These allow speakers to indicate thematic units and ends of turns. The most important units for the kind of dialogue we are concerned with are the information units - each one of which represents one package of information. The information unit constitutes a tone group ending with a tonic syllable, which is the loudest syllable in the unit, and has one of a series of five tonic movements, each one with a characteristic meaning such as 'question' or 'incompleteness'.

In the context of dialogue design, another crucial macrophonological feature is the pause. Long dismissed as involuntary imperfections, like sneezes, such phenomena are now recognised as much more diverse and important. They can be divided first into intra-turn and inter-turn. Intra-turn is where a speaker is doing something or thinking of an answer, and there is no doubt that he is next to speak. Inter-turn gaps could be like that between a question and the answer. Depending on the last item of the first speaker, such as

"Hang on a minute!", or "Trying to connect you", the listener knows he is being invited to wait. Few systems allow the user to make such an input, and usually repeat the question after a timeout period, or say "I have not understood your reply".

Human pauses are often filled by some voiced sound - often /a:/ or /m/ to indicate that they are intra-turn. Some dialogue systems employ music to indicate that the channel is still open.

In the overwhelming majority of current commercial dialogue systems the recorded system prompts are designed to emulate the utterances of a human agent. Invariably, systems of any complexity at all will contain many explicit interpersonal tokens, such as "Please tell me the number of your account", "Thank you for calling XXX" and the like. In this way, dialogue designers attempt to create the illusion that it is a human agent who is on the line, rather than a programmed system. This kind of attempt to lead the caller 'up the garden path' is not, however, very successful in terms of giving callers what they want, and recent experiments have shown that callers interacting with automated dialogue systems tend to prefer a more 'machine-like' style of system prompt (Cheepen & Monaghan 1998, Williams, Cheepen & Gilbert 1998).

In contrast to this strategy of trying to make system prompts as 'human-sounding' as possible (in terms of lexicogrammatical content), dialogue designers approach the job of building the prosodies of system prompts by adopting a very extreme form of signalling, which is far removed from ordinary, human-human dialogue. This is particularly noticeable in the case of systems which are employed within application domains which are concerned with leisure facilities. A leisure centre dialogue which is in current commercial operation, for example, contains the following sequence:

there's a great range of activities available at the Jarman Park leisure world <music> hot shot . it's a new sports concept . with something for everyone - twenty . ten pin bowling lanes - interactive golf ranges - American pool tables - satellite tv . and much more - it's a world of sport . game on <music> disco on ice . whether you're a novice or a skilful ice skater . you'll find plenty of fun in store at silver blades . skates one <music> now you can play bowls all year round . in our indoor bowls centre

Each fragment of music is a different type, to iconically signal the particular kind of leisure activity being announced. This is clearly an attempt to signal that the leisure provision is 'fun, fun, fun', and to underline the different kinds of fun which are available.

It is interesting to note that the enormous effort which goes into designing this kind of sequence is very much at odds with the approach to other, more directly interactional sections of the dialogue. In complex systems, where there is the potential for (perhaps numerous) exchanges between caller and system - as, for example, in some telephone banking applications where the system may be carrying out database lookup of various kinds, and the caller may be required to input different kinds of information - there will inevitably be quite substantial pauses in the dialogue, which are not, strictly speaking, part of the overall prosodic design. Pauses occur when the system is waiting for input from the caller, and when it is processing the caller's input (e.g. accessing the underlying database in order to retrieve the desired information). These 'gaps' in the talk do not fit well with the polished prosodies of the system prompts, and the result is a considerable tension between flowing, pseudo-natural system talk, and an often lengthy silence.

From the user's point of view, this is maximally confusing. Often the assumption is that the system has disconnected, or has not 'heard' the last caller input, and the caller will frequently try to repair what they interpret as some kind of breakdown in communication by repeating - or, even more disastrously, rephrasing - their last input. This can lead to very serious problems with the underlying recognition system, and the end result is too often a real breakdown in communication.

4. CURRENT RESEARCH ON SYSTEM PAUSES AND ITS RELEVANCE TO THE DESIGN OF PROSODIES

Recent work has attempted to tackle the problems users encounter when faced with system silence, and there has been some experimentation with auditory icons (Brewster et al 1995, Cheepen, Monaghan & Williams 1998). It has been shown that greater usability can be achieved (certainly in terms of few user errors) when auditory icons are used to signal the machine processing state.

This kind of investigation is, however, in the early stages of development, and more research effort must be expended in order to discover the best ways to signal the meaning of particular instances of silence in automated dialogue utterances.

For the present, callers to automated systems must continue to struggle with the prosodies of dialogues as they are - on the one hand exaggerated - and, in the caller's view, often patronising - and on the other hand so minimal as to provoke the caller into believing that communication has broken down and to attempt a repair. It is perhaps the contrast between these two extremes which is the most important factor in the dissatisfaction of users with automated systems.

The perfect solution to the problems of dialogue design is not immediately in view, but in terms of improving the usability of currently available systems, some short to medium term practical improvements are certainly possible. If designers can achieve a compromise between the two problematic positions - by toning down some of the over-emphasised prosodic and non-linguistic signalling while simultaneously making the meaning of system silences more transparent to the user - then user satisfaction will certainly be enhanced, and this will mean a substantially improved level of usability in the dialogue systems which are at present proliferating throughout the commercial sector.

5. REFERENCES

Brewster, S A., P.C.Wright, A.J.Dix & A.D.N.Edwards, 1995, The Sonic Enhancement of Graphical Buttons, Proc. of Interact'95

Cheepen, C., 1988, The predictability of informal conversation, Pinter Publishers

Cheepen, C. & J.Monaghan, 1998 (in press), Designing for naturalness in automated dialogues: some problems and solutions, in Y.Wilks (ed) 1998, Machine conversations, Kluwer

Cheepen, C., J.Monaghan & D.Williams, 1998, Is it a bird? Is it a plane? No it's a dialogue system, this volume

Monaghan, J., 1979, The Neo-Firthian tradition and its contribution to general linguistics, Niemeyer, Tubingen

Williams, D. & C.Cheepen, 1998, "Just speak naturally": designing for naturalness in automated spoken dialogues, Proc of CHI '98

Williams, D., C.Cheepen & N.Gilbert, 1998, Experiments in how automated systems should talk to users, Proc of HCI 98