Natural Voice Output in Interactive Information Systems

Katherine Morton
Mark Tatham

Reproduced from In R. Lawrence (ed.) *Proceedings of the Institute of Acoustics 1996*, 43-50. St Albans: Institute of Acoustics

---

THE REQUIREMENTS FOR GOOD COMPUTER SPEECH

Computer speech for human-computer interaction is being developed for use in **information access systems**. Speech is the final stage in a speaker's encoding of his/her *thoughts* and *attitudes* into language in order to communicate them to a listener. But, in addition to good synthetic speech in an information access system it is essential to link the speech output with **a natural language processor** in order for the system to be useful interactively.

There are two aspects to good computer speech

1. the speech must be *intelligible*, and
2. the speech must sound *natural.*

Virtually all current synthesis systems are intelligible. The words can be decoded and meaning assigned by the listener. In addition, much voice output is in fact simple concatenated recordings; for example, the BT's Directory Enquiries voice consists mainly of a set of pre-recorded digits, which are simply abutted to give the caller a phone number. What is missing in the pre-recorded voice in this case is a suitable *overall* intonation contour; that is, the 'melody' of the utterance associated with English is not there. [To be fair, BT's system does have local word-based intonation provided by selecting from two or three versions of each digit, each recorded with a different contour.]

The task for synthetic speech is now to produce good intonation patterns and at the same time *appropriate* intonation. An appropriate intonation necessarily includes the right **tone of voice** which is an encoding of the attitude and feelings of the speaker.

THE SPECIAL NEEDS OF INQUIRY SYSTEMS

In an inquiry system, the computer system will give instructions, or explanations, or perhaps make a comment on the user's request. Such a system would be more usable if it produced a tone of voice acceptable to the user. The reason for this is that producing an appropriate tone of voice significantly increases user acceptance of the system.

For example, a user inquiring about bank balance over the phone, would probably want a reply that sounded helpful, rather than a series of abrupt statements of fact. Someone who was unsure about his/her inquiry, might want a sympathetic sounding voice to reply, speaking with confidence. Because these tones of voice can be readily supplied by human speakers and indeed *always are supplied*, listeners expect to hear them. If they are not present then something is perceived to be wrong.

If a human user appears confused, to repeat the same *message 'this is not correct, please try again'* becomes irritating after a few repetitions. However, if the tone of voice conveys patience, firmness, and can give the impression of sympathetic understanding to the plight of the confused user, then the user will be more inclined to persist with trying again and again.

Consider, for example, two sample dialogues, each concerned with requesting a bank statement over the phone. In the first frustration is guaranteed, with the user aborting the

exchange. In the second there is a good chance that the user will continue: the bank's computer, whilst persistent, is nevertheless polite and appears to understand the user's difficulties.

Dialogue One - phone call to computer-bank

Computer: Midgeworthy Bank. Which service do you require?
Customer: Bank statement.
Computer: Press double 0.
[pause - customer thinks he's done as requested - but he hasn't]
Computer: Press double 0.
Customer: I have.
[pause - computer waits, customer does nothing - he thinks he's already done as requested]
Computer: Press double 0.
Customer: I've done that.
Computer: Press double 0.
[Customer hangs up]

Dialogue two - phone call to computer-bank

Computer: Midgeworthy Bank . Which service do you require?
Customer: Bank statement.
Computer: Press double 0.
[pause - customer presses single 0]
Computer: Customer, press *double 0*. [change of pitch and addition of salutation]
Customer: I have.
[pause - he has not]
Computer: Please press ***double*** 0.
Customer: Oh, double 0.
[now he presses double 0]
Computer: Thank you. You're connected now.

Addition of words like *please, customer,* etc. help. But even words like *Try again* can irritate. The dialogue must be sensitive to the customer's frustration and helplessness. Adding friendly and polite words is not enough - the attitude of helpfulness can best be conveyed by tone of voice. A listener expects words conveying a basic meaning - but also *always* expects an appropriate accompanying tone of voice adding semantic nuance to those words. The synthetic utterance must, of course, use the right words (generated by the higher level language components of the system), but must in addition adopt an appropriate tone of voice if it is to sound at all convincing and maximally effective.

MODIFYING THE SPEECH WAVEFORM

The speech waveform can be modified in three ways to produce changes in tone of voice:

1. the *fundamental frequency* can be increased or decreased,
2. the *duration* of words and syllables can be increased or decreased,
3. the *amplitude* (loudness) can be increased or decreased.

A combination of these changes can also be made; increasing duration and lowering fundamental frequency will give the impression of slow careful speech - useful for

instructions. Increasing fundamental frequency and shortening the duration of syllables can give the impression of excitement or encouragement.

Because *human-computer interface* (HCI) systems will be expected to deliver important information users will expect a high level of naturalness befitting the importance of the information and incorporating or overlaying the appropriate attitudinal or emotive effects. A further consideration is that the information given will probably be over an extended period of time during the exchange. A user will expect computer speech to respond as the dialogue unfolds and will not be happy with the same intonation pattern recurring for each intervention by the machine.

IDENTIFYING FUTURE DESIRABLE PROPERTIES OF SYNTHESIS FOR HCI

Synthetic speech will be fully integrated with high level language processing in future HCI development. Such HCI systems will have a sophistication exactly matching what is needed for computing the right tone of voice to use. This is because they will be able to

- recognize *continuous speech* from multiple untrained users;
- understand the *meaning* of utterances;
- make decisions about the content and the *future direction* of the conversation;
- produce natural-sounding speech *appropriate to the discourse*;
- deal with *multiple languages*;
- incorporate *many domains* with large vocabularies;
- *recover* from errors;
- *provide help* to confused users;
- ideally *interface with other modalities* such as handwriting recognition and mouth/eye tracking;
- *operate in real time*.

Although as a first approximation all these qualities will be needed, there are currently no interactive speech systems which can fully achieve these goals, although many attempt several of them.

A number of problems can be identified which are currently impeding full development of conversational spoken language systems (Peckham 1995). In addition to cost, there are problems of menu structure, identifying faster and more direct approaches to system design, and technology requirements such as real time operation, repair, indexing, and incorporating adequate language models, including suitable voice output for dialogue exchange incorporating more complex ideas than simple factual information.

HCI systems are used to deliver important information. Therefore users have an expectation of a high level of naturalness in synthetic voice output, and in some emotive situations will be happier with the appropriate emotion and attitude from the system voice. Further, in a sustained dialogue of more than a few exchanges, or where a problem or confusion arises, the user will be more likely to persist if, as the dialogue unfolds, the appropriate attitude and emotion are rendered by the system voice.

A straightforward computational method will be required in order to link the language model with the speech model. One way is to mark all syllables with respect to a range of possible acceptable changes and, in addition, perhaps marked with respect to word boundaries. The initial speech produced would be neutral, with the possibility of working within the marked ranges to produce varying tone of voice.

PHONOLOGICAL OUTPUT IS INITIALLY INSENSITIVE

Using the terminology of linguistics, the focus of attention is on *phonological* aspects of the language model. In particular we are concerned with how pragmatic and other constraints impinge on the *normative* output of a phonology which is initially insensitive to anything

other than straightforward syntactic and semantic demands. For us the reconciliation of the abstractions of the language model with the physical model of low level synthesis is the crux of the theoretical issue. Representation in the two models is particularly important since we regard a satisfactory reconciliation as largely hinging on representational compatibility.

For the future, it will be necessary to investigate the means of matching the appropriate tone of voice with the language structures involved in the interaction between human and computer *as the dialogue unfolds*. This is extremely important - it is not possible to *set* tone of voice once and for all at the start of a dialogue - changes occur as the dialogue proceeds. The *flow of changes* of tone of voice is an essential property characterising dialogue itself.

The speech model discussed here is currently being applied to several synthetic speech systems to convey politeness, encouragement, patience friendliness. The goal is to build voice output systems that not only speak the plain message with clarity, but also speak with the prosodic features that convey the appropriate attitude.

THEORETICAL PERSPECTIVES

The work of the collaborating synthesis research teams at Essex, and Essex and Bristol, is based on underlying theoretical models in speech production. Although it builds on influential work in the 70s (Fowler 1977) our current speech production uses of the idea of *gestures* in line with the **articulatory phonology** of Browman and Goldstein (1986).

Earlier speech production theory had emphasised the *verticality* of speech events by focusing on the phonetic element (deriving from the phoneme) as an entity. Although phonology had introduced the idea of distinctive features (Jakobson et al. 1961), this idea was still a device similar to earlier ones (although more formal) for classifying individual segments, while making some claims about how they were thereby enabling generalisation among cognitive phonological processes. What was needed was a feature-like element uniting cognitive and physical perspectives on speech production.

Gestures enable the focus to shift to *horizontality,* and at the same time introduce a common descriptive tool for both the phonological and phonetic (cognitive and physical) levels - something which the earlier distinctive feature theory had failed to do. At the cognitive level a gesture can be thought of as a phonological representation of part or all of a possible or intended unit which can function at a linguistics level in the language in question. It involves the intention to co-ordinate (at the physical level) a particular grouping of motor elements (e.g. muscles) to achieve the linguistic goal which the gesture represents. The focus is on the elements which co-ordinate and the way in which their co-ordination unfolds in time as a sentence progresses through sequenced gestures - they are strands in a score which is being dynamically played out (Browman et al. 1986).

In our view gestures are best modelled with *fuzzy* beginning and endings. Fuzzy in the sense that there is no intrinsic need for temporal simultaneity of onset or offset to all the elements involved in a gesture. It may be that the linguistic goal requires simultaneity of some of the elements, but not necessarily all. **Cognitive Phonetics** attempted to explain robustness hierarchy within the earlier verticality framework, but this idea holds just as well within the horizontality framework. It should be emphasised that Cognitive Phonetics enables accounting for variability. It is intended and carefully rendered variability that enables synthetic speech to trigger ‘naturalness’. I maintain that it is within the degrees of freedom of a system to threshold the performance of certain elements within a gesture and thereby establish a robustness hierarchy for holding onto linguistic goals despite external pressures. Fowler (1993) refers here to ‘protection’ of phonetic gestures against coarticulatory influences: we seek to establish an explanation at a linguistic level as to when and how such protection might occur.

PRAGMATIC PHONETICS

Pragmatic phonetics (Morton 1992) is the general theory characterising effects such as tone of voice and is concerned with identifying and modelling those systematic features of the

acoustic signal which reflect the mood, intentions and beliefs of the speaker - and which are able to trigger in the listener an appropriate decoding. Some of these acoustic features have been determined (Morton 1992) by data reduction using neural networks to identify the relevant overlays to the fundamental frequency and timing parameters in speech production. The overall pragmatic phonetic model has been tested by generating synthetic speech and running listening tests to determine whether the overlays are adequate in triggering the appropriate perceptual responses; results are encouraging.

CONCLUSION

Understandably early synthetic speech researchers were principally concerned with intelligibility. Now that almost all synthetic speech has achieved that goal attention has shifted toward naturalness. We are interested in interactive dialogue systems in the HCI context. For this reason naturalness is pivotal to our work - it is one of the keys to user acceptability of speech HCI systems. For the reasons we have cited naturalness is as essential in our systems as intelligibility was in earlier non-interactive systems.

We use three important developments in the theory of speech production to underpin the development of a working synthesis system for HCI:

1. the introduction of the notion of *gestures* by Browman and Goldstein - leaping ahead of the limited allophonic model;
2. the introduction of *Cognitive Phonetics* providing the theoretical basis for fully integrating cognitive and physical aspects of speech - and thereby providing the channel through the model to enable low-level modification of intonation, duration, etc. to conform with high level considerations of attitude and emotion;
3. the introduction of *Pragmatic Phonetics* with provides the theoretical characterisation of attitude and emotion in terms of speech production itself.

Initial relatively informal testing on an interactive basis indicates that we seem to be proceeding along productive lines in our research. The synthetic speech we can produce automatically now has recognisable naturalness based on listeners being able to detect the computer's attitude or mood.

REFERENCES

Browman, C.P. and Goldstein, L. (1986) Towards an articulatory phonology. In C. Ewan and J. Anderson (eds.) *Modularity and the Motor Theory of Speech Perception*. Hillsdale N.J.: Lawrence Erlbaum Associates

Fowler, C.A. (1977) *Timing Control in Speech Production*. Bloomington: Indiana University Linguistics Club

Fowler, C.A. (1993) Coordination and coarticulation in speech production. *Language and Speech*, Vol.36, pp.171-195

Jakobson, R., Fant. G. and Halle, M. (1961*) Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates.* Cambridge MA: MIT Press

Morton, Katherine (1992) Pragmatic phonetics. In W.A. Ainsworth (ed.*) Advances in Speech, Hearing and Language Processing*, Vol.2, pp.17-53. London: JAI Press

Peckham, J. (1995) Conversational interaction; breaking the usability barrier. In *Proc. ESCA Workshop on Spoken Dialogue Systems*, (ed) Paul Dalsgaard. Aarlborg

Tatham, M.A.A. (1990) Cognitive phonetics. In W.A. Ainsworth (ed.) *Advances in Speech, Hearing and Language Processing*, Vol.1, pp.193-218, JAI Press, London