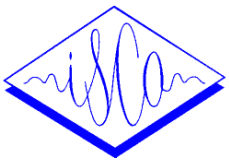

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

PRAGMATIC EFFECTS IN SPEECH SYNTHESIS

Katherine Morton
Mark Tatham

e-mail: kate@essex.ac.uk
Linguistics Department
Essex University
Colchester, UK—CO4 3SQ

ABSTRACT

Voice output for applications such as dialogue and information systems requires more than a plain or neutral tone of voice if it is to become more acceptable. These systems must be able to reproduce the acoustic characteristics of emotional and attitudinal effects driven by pragmatic procedures embedded within the dialogue controller. Possible approaches to modelling to this aspect of naturalness are discussed, and some sample data presented which illustrates some of the problems to be encountered.

1. INTRODUCTION

This paper addresses the problem of incorporating pragmatic variability into synthetic speech. Doing this is important because the overall result is an obvious and desirable increase in the perceived naturalness of the output, and for that reason an improvement in the acceptance of synthetic speech. One major use can be seen in the increasing demand for voice output systems in, for example, human–computer interface systems. The area of application addressed here is that of dialogue systems involving interaction between human beings and computers.

There are many considerations to be taken into account when attempting to make synthetic speech sound more natural; for example — ease of use and increased attentiveness on the part of the user. The effectiveness of incorporating pragmatic variability can also be judged by the formal methods being developed for the assessment of synthesis systems — a discussion of techniques and suitability can be found in the EAGLES documentation.[1]

2. VARIABILITY

If we consider the process of speech production as a whole we notice distinct types of systematic variability which have been classified by linguists and which can be modelled independently at the phonological and phonetic levels. Whether we adopt a linear approach or a non-linear approach to phonological theory, underlying abstract segments are systematically mapped onto surface variants in a one-to-many translation procedure. The object of the mapping is to account for increased detail in the specification of segments as we proceed through the levels within phonology.

Thus, for example, an underlying /L/ maps to either velar /l$_w$/ or palatal /l$_j$/ in easily identified *phonological* contexts. At the phonetic level, the mapping is from extrinsic (or intended) allophones to intrinsic (or automatic) allophones. For example, an underlying /l$_j$/ maps to both fronted and retracted variants depending this time on *phonetic* context. These examples are given in the framework of classical linear phonology and phonetics.[2, 3] Although more recent non-linear theories[4, 5, 6] treat the data differently, it is generally agreed that these are phonomena which a. need accounting for, and b. which are different in kind (that is, phonological *vs.* phonetic).

In this paper we are addressing a third kind of systematic variability observable in the articulation of spoken language and in the corresponding soundwaves. This relationship is currently under investigation in the development of the theory of speech production and perception.[7] In the acoustic speech signal the effects of this category of variability are observable as the occurrence of systematic phenomena correlating with the speaker's communication of mood or attitude — but without a special selection of words for that purpose. The acoustic effects are said to be *overlaid* on what is otherwise a neutral acoustic signal. Some of the results of these studies are already being applied in speech synthesis systems.[8]

3. PRAGMATIC PHONETICS

Spoken language can be seen as a two level model:

- production of the plain message;
- production of overlaid features on the plain message.

Taking what we can call the **overlay model** as a first approximation we are able to recognize changes or overlays to the neutral way of speaking — speech in no particularly emotive style. The pragmatic effects are expressed in the difference signal between neutral speech and speech reflecting a pragmatic effect — that is, reflecting the speaker's intention to produce utterances which are non-neutral. The difference signal expresses the speaker's emotions, attitudes and feelings.

These non-neutral effects are often described in lay terms like *tone of voice*. The fact that listeners are able to report (with a fair degree of inter-listener agreement) speaker effects such as *happy, sad, confident, etc.*, indicates reliability, systematicity or robustness in the acoustic patterning produced by the speaker. The effects succeed in providing a useful extra dimension of information not found in written text or neutrally spoken utterances.

The term **pragmatic phonetics** was introduced by Morton[7] to describe an area of speech production theory concerned with

- characterizing the acoustic strand in speech responsible for signalling emotion, etc., and
- modelling the source of the overlay.

In broad terms, pragmatic phonetics accesses a pragmatic component in linguistics, then interacts with the neutral descriptive string output from phonology to trigger phonetics to overlay the neutral pronunciation of the phonological string. Deliberate overriding of normal passive effects is provided for within the theory of **cognitive phonetics**,[9] and it is this mechanism which pragmatic phonetics employs.

4. SYNTHETIC SPEECH

We use a high level speech synthesis system in specific applications and for testing speech production and perception theory. High level speech synthesis is that part of a general synthesis system which is responsible for generating and specifying what the synthesizer is to say: it is *not* the part of the synthesizer which actually speaks. For a discussion of the differences between high and low level synthesis see the paper by Tatham and Lewis in these Proceedings.[10]

The two ways of using synthetic speech are therefore:

- in dialogue applications incorporating the pragmatic phonetic model to improve acceptability and naturalness;
- as a test bed for supporting or refuting hypotheses about acoustically signalled pragmatic effects.

In the case of dialogue applications we note that even the best voice output systems[11, 12] currently produce only neutral speech with no attempt at pragmatic coloring of the plain message. This is a clear shortcoming because in normal conversation we *always* expect speech to have some emotive content. We also expect that the emotive content should change over time as the dialogue unfolds. If this is the expectation of the human participant in the dialogue it is clearly the case that the synthetic side of the dialogue will sound more natural and more informative if pragmatic variation is incorporated.

Synthesis which goes beyond the simple, neutral conveying of information and moves toward incorporating a range of pragmaticly significant variation can communicate, for example, warning or instil confidence in nervous users, etc., by using the appropriate tone of voice. This is what human beings do — and so this is what users of dialogue systems expect of a synthesizer. If they do not perceive these effects they detect that 'something is wrong' and react appropriately.

A dialogue between human speakers and listeners will often call for pragmatic effects to be used in a continuously adaptive way. A neutral tone, for example, may be used in response to a simple request, but if the speaker continues to ask for information, the constant use of a neutral tone for replying evokes a sense of irritation. A tone of voice conveying an impression of firm, friendly interest is more appropriate, and the human participants in the dialogue will adapt their own styles of speech appropriately as the dialogue unfolds. It is this effect that we are trying to replicate.

5. THE APPROACH TO MODELLING

There are two things to be modelled if we are to capture in high level synthetic speech the particular pragmatic effects we have been discussing:

- how the correct emotive or attitudinal marker is to be called (handled within the pragmatic processing routines of the model), and
- how the generated marker is to be interpreted within the speech production routines.

Within the dialogue system the dialogue interpreter 'understands' what the human being is saying and tracks changes in attitude for the human side of the dialogue. This in turn is triggering responses in the dialogue manager, whose function it is to retain

control of the conversation and generate appropriate machine responses back to the human user. Thus the dialogue manager triggers pragmatic procedures to call emotive or attitudinal markers at the right time.

The generated marker is abstract and terse: it says nothing more than something like: 'The response to this question must be spoken with firmness, highlighting the key words.' We call what is to be spoken the **script**, which is then marked up with these pragmatic indications. The result is the **marked-up script**. In a basic system the marked script could be thought of as the input to a text-to-speech synthesizer, but in a more sophisticated system it could equally well be a concept representation for a concept-to-speech synthesizer.[13]

Our speech production and perception models are computational and fully explicit. They are within the general scope of non-linear Articulatory Phonology[5], Task Dynamic Theory[14], Cognitive Phonetics[9] and Pragmatic Phonetics[7] — all of which are computationally oriented theories. The models must be computational to enable them to be programmed either for testing or for use in applications such as dialogue systems. The more traditional cursory or impressionistic modelling of speech production and perception is not suitable for the rigorous environments of dialogue and similar applications.

6. SOME DATA EXAMPLES

Papers such as this are intended to give a flavour of what is being reported, rather than give long and detailed algorithms or even program examples — these presentations are reserved for another format. But it will be useful at this point to give a few examples of some data collected while studying the intonational dynamics of conversation involving exchanges of high information content — exactly the kind of situation encountered by a telephone centered information service, for example.

Results reported elsewhere[7] of studies of intonational dynamics under varying pragmatic conditions were based on the analysis of complete sentences uttered by subjects using a variety of markers. In dialogue however human speech rarely consists of full sentences; the utterances are often fragmented and much of the conversation consists of what linguistics describes as *phrases*, or other sub-sentence strings. A dialogue model for human/computer interaction which can move the linguistic focus away from the normal sentence unit might be helpful.

A system for marking phonological intonation and relative timing devised by Tatham for general modelling of sentences and sub-sentence fragments was used in describing information exchange between speakers in a map reading experiment. The experimental environment was classical, with each participant having a slightly different map, and with the group asked to collectively plan a route between given points. As expected, very few linguistically 'complete' sentences were exchanged, with the conversation consisting of phrases or other sub-sentence units. The phonological analysis of intonation and relative timing was explicitly linked to instrumental analysis giving data on fundamental frequency (f0) change and absolute timing.

Some simple, general and exemplar observations can be reported which show the kind of effect which needs accurate modelling when simulating such an exchange of information.

- When the concentration of meaningful words with no redundancy or repetition was high (e.g. '... there's a river and no bridge ...') the range of variation in fundamental frequency within the phrase is regularly narrow compared with that normally expected in a medium content neutrally uttered sentence. Although a restricted range of variation of perceived intonation is an interpreted characteristic of *gloomy* speech the utterances in this experiment did not sound gloomy at all, but rather as if the speaker were giving information that was important. Note then that the interpretation of a given acoustic event may well be context sensitive, and that this has to be taken into account in generating an f0 appropriate for a particular script marker.
- When thinking aloud speakers followed the standard phonological pattern of intonation. This occurred, for example, in an utterance such as '... I wonder if we should go along this road for a bit, and then maybe turn here or maybe over there ...' In these cases an expected intonation pattern was interpreted with a wider than expected range of fundamental frequency. This triggered a perceptual effect of 'musing'.
- If a phrase ends in a lower than expected f0 the following phrase tends to begin with a low f0. If a phrase ends in a higher f0 the following phrase may begin with a high or a low f0. Furthermore if a phrase ends with a high f0 this can serve as a cue for another speaker to speak *for the purpose of* adding information to or comment on the previous utterance.
- It is generally agreed that in English a rising f0 at the end of an utterance signals a question. How-

ever in this study speakers frequently ended a phrase with a high f0 indicating they were unsure but not that they were asking a question. For example, '... not on the same road here ...' — mistakenly interpreted by another speaker as a query marker rather than as an uncertainty marker. This example is included to illustrate the fact that pragmatic markers will often need to be interpreted in production on a probability basis and that the the production will need to be sensitive to a prediction of the likely perceptual effects.[9]

- There are some difficulties in consistently relating f0 change with the perception of intonation. Researchers have noted that f0 contours are interrupted by unvoiced segments and that when a series of voiced segments occurs it can be difficult to isolate the words being studied, and difficult therefore to interpret the emphasis the speaker has placed on those words.
- It is agreed by researchers that in English a low f0 signals the end of an utterance or the end of a topic. However in spontaneous speech this is not always the case. Frequently, a pause in the timing of speech can signal the end of an utterance whether the f0 is high or low. And some speakers use a low f0 to signal that they wish to add more information, e.g. '... there's a church here ... and a woods ... and a hill ... and we're going north now ...' This emphasizes the phenomenon of perceptual thresholding of effects and the need for this to be predicted in the production model for generating an accurate acoustic signal to interpret the pragmatic markers.

There are many implications arising from such data that good natural sounding synthesis needs to take into account. We might assume that human interaction in conversation is rule governed. But the linguistic descriptions of what happens acoustically may not be accurate — they work satisfactorily perhaps at an abstract phonological level, but they are quite inexplicit as to how they relate to the acoustic signal. Perhaps more importantly phonological descriptions are rarely probabilistic or able to accommodate the notion of thresholding.

7. CONCLUSION

Voice output systems require precise acoustic representations of the intended meaning. For the dialogue manager to retain control token handing to the user, for example, must be signalled according to precise, *yet varying*, conventions. Replying to a user with the appropriate emotional and attitudinal responses will determine to some extent the successful use and acceptability of an interactive application.

REFERENCES

[1] N. Calzolari and J. McNaught (eds.). *EAGLES Interim Report* — Document No. EAG-EB-IR-2. The European Commission. Brussels 1994

[2] N. Chomsky and M. Halle. *The Sound Pattern of English.* Harper and Row. New York 1968

[3] P. Ladefoged. *A Course in Phonetics.* Harcourt Brace Jovanovich. London 1982 (second edition)

[4] J.A. Goldsmith. *Autosegmental and Metrical Phonology: A New Synthesis.* Blackwell. Oxford 1989

[5] C.P. Browman and L. Goldstein. 'Articulatory phonology: an overview.' *Phonetica* 49. 1992, pp. 155-180

[6] C.A. Fowler, P. Rubin, R.E. Remez and M.T. Turvey. 'Implications for speech production of a general theory of action.' In B. Butterworth (ed.) *Language Production.* Academic Press. New York 1980, pp. 373-420

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

[8] B. Granstrom. 'The use of speech synthesis in exploring different speaking styles.' *Speech Communication* 11, pp. 347-355

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

[10] M.A.A. Tatham and E. Lewis. 'Naturalness in a high level synthetic speech system.' *Proceedings of Eurospeech '95.* Madrid 1995, this set of volumes

[11] J. Allen, M.S. Hunnicutt and D. Klatt. *From Text to Speech: The MITalk System.* Cambridge University Press. Cambridge 1987

[12] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text to speech synthesis using diphones. *Speech Communication* 8. 1990 pp. 453-467

[13] S.J. Young and F. Fallside. 'Speech synthesis from concept: a method for speech output from information systems.' *Journal of the Acoustical Society of America* 66. 1979, pp. 685-695

[14] E. Saltzman. 'Task dynamic coordination of the speech articulators: a preliminary model.' In H. Heuer and C. From (eds.) Generation and Modulation of Action Patterns. Springer-Verlag. Berlin 1986, pp. 129-144