PALM: Psychoacoustic Language Modeling

Katherine Morton

Reproduced from: Morton, K. (1992) PALM: Psychoacoustic Language Modeling. Proceedings of the Institute of Acoustics, Vol. 16, Part 6. St Albans: Institute of Acoustics, pp. 189-197.

---

1. INTRODUCTION

Speakers can alter the way a sentence is said so that a sense of the speaker's feelings or an attitude is conveyed to the listener.

The expression of feelings or attitudes can be quite subtle. In a dialogue, such as telephone inquiry, an exchange of information occurs during which speaker and listener may change roles. For example a listener, as speaker, can respond to a request for specific information by adding emphasis or contrast to a reply; the features emphasis or contrast can convey different attitudes to the listener.

This paper discusses some physical cues useful in a human-machine dialogue system that will communicate attitudes, emotion, or beliefs without using words to describe those feelings. Among these are adding the appropriate pragmatic effects, avoiding unintended pragmatic effects, and signalling the end of information exchange.

The paper is not about classifying emotions or attitudes; I assume it is possible to do so. I also assume that emotions or attitudes are encoded in the speech waveform, and that listeners are able to recognize and report a response to the encoding.

I am concerned with modeling how a trigger for expressing emotion, whatever its origin or however it is generated, is phonetically realized as a sound wave with identifiable characteristics. The model here is a simulation of what happens during pragmatic phonetic processes in human speech [Morton 1991a]. The test of the model is to ask panels of listeners if they can identify a particular emotion or attitude within a set of utterances generated by the model.

2. THE SYNTHESIS MODEL

The method of producing a synthetic speech waveform used for developing the simulation is based on parallel formant synthesis [Holmes 1988]. In this system a waveform is assembled parametrically; time-varying values are created for each of a number of parameters judged to be essential in the waveform. These values include the center frequencies of tour or more of the lowest formants, their amplitudes, the type of excitation — periodic or aperiodic or a mix of the two — and changes of fundamental frequency. The values specified for each parameter are updated every 10ms.

When parameter values are determined from spectrograms and the results delivered to the synthesizer at a rate of 100 frames/s, the speech output is judged similar to real speech. It is intelligible, and natural sounding. Thus errors detected in the output signal most likely come from the model itself rather than the synthesizer [Holmes 1988, Morton 1991b].

3. TEXT-TO-SPEECH SYNTHESIS IN DIALOGUE

Although speech synthesis systems that replay pre- recorded or pre-analyzed human speech have their uses, they are not sufficiently flexible where pragmatic effects such as emotion and attitude might be useful. Pre-recording announcements with pragmatic effects is probably not practical, since it would require multiple recordings of the same sentence to be stored.

But pragmatic effects could certainly enhance human-machine communication in dialogues since the computer voice output would sound more natural and convey information

more precisely appropriate to the dialogue itself. In complex applications involving dialogue, flexibility is essential because

- what the machine will say, and
- how it should say it may not be predictable until the actual dialogue is taking place.

4. EMOTIVE PHONETIC VARIABILITY

A listener can detect change and variability in the human speech waveform. Speech researchers have identified two major types of variability: I call these types intended and unintended. Variability refers to changes that occur through time within the waveform as it signals the sequence of phonological units that characterize intended words. Words, in turn, encode meaning. Other types of change correlate with switching between strings of voiced and voiceless segments, and so on. Such changes are cognitively intended by the speaker and form the basis of how speech is used for communication.

Some variation also occurs in the acoustic waveform arising from constraints imposed by the mechanics and aerodynamics, etc., of the vocal tract system itself. These variations are not cognitively intended by the speaker; they are not being used directly to encode speech, and are unintended.

In addition, there are systematic variations in the signal which fall into neither category. For example, the same sequence of phonetic elements (encoding same basic meaning), and altered by mechanical and aerodynamic constraints, can produce different waveforms. Relative timing between phonetic elements and fundamental frequency contours themselves can be different. These variations are systematic; they can be repeated by the speaker, and detected and classified by the listener. This suggests they are intended and that they contribute to the dialogue.

These systematic changes signal the emotive content of speech — a sentence can be spoken in different ways reflecting emotions, attitudes, beliefs, etc. During the course of a dialogue these emotions can be seen to shift and to form an important part of the exchange between speakers. Remove these variations, and the dialogue consists of the basic meaning of the sentences only, without any of the pragmatic subtleties.

In standard text-to-speech synthesis systems, given a repetition of the same sentence or sequence of words, the output signal will not vary with respect to timing or f0. The result is characterized as unnatural or machine-like, particularly in a dialogue. A human being interacting with the system does vary timing and f0, but computer speech cannot yet respond in the same way.

5. MODELING EMOTIVE VARIATION

Emotive variation is dealt with as an overlay on the well-defined unemotive or neutral speech the synthesis system normally produces.

Although there are several ways of classifying pragmatic effects, ordinary terminology is used; for example, a prominent word in a sentence is described as spoken with emphasis or with contrast. A sentence can be spoken to convey the feeling that the speaker is happy, gloomy or surprised, and so on. Therefore it's possible to speak of overlays of happiness, or contrast, emphasis or surprise, etc.

The file of parameter values output from the text-to-speech synthesis system can be intercepted before it reaches the formant synthesizer. At this final stage, after the values have been set to produce a neutral sentence, the file can be modified to produce the required overlay.

In the model presented here, this process inspects and changes the neutral utterance by making adjustments to timing and the f0 contour. The modification is triggered by an appropriate marker from a pragmatic module and is usually attached to the entire sentence. The process itself knows what neutral prosodics it can expect and how to modify these in

view of specific pragmatic triggers. The changes are made to individual segment durations in sentences (thus altering the sentence rhythm) and also to their f0 contours.

Fig. 1 shows a block diagram of the overall model; the neutral utterance, generated by the synthesis system, is modified by an overlay process triggered by a signal from a pragmatic level in a dialogue system.

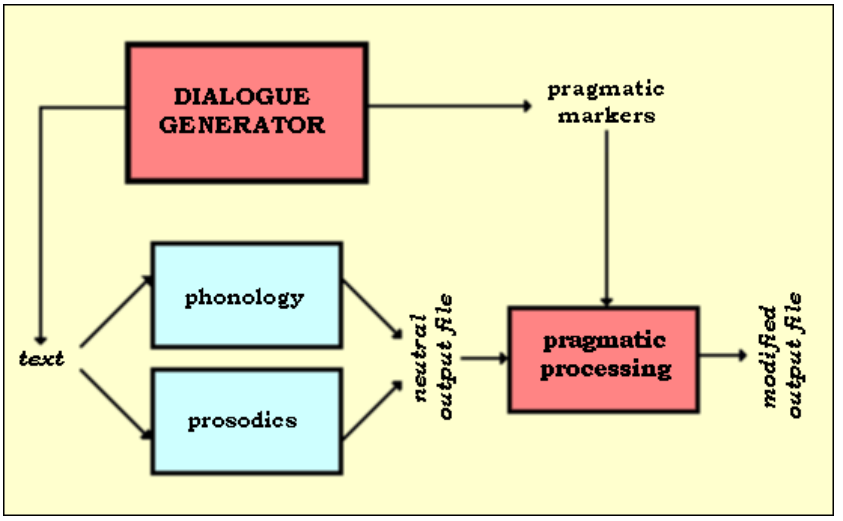

Fig. 1 The overall model for generating a pragmatically modified output file.

6. THEORY UNDERLYING THE MODEL

At the physical level phonological word stress in English can be interpreted in three different ways:

For the stressed syllable (or word in the case of a monosyllabic word):

- increase the amplitude;
- increase the duration;
- raise the value of f0.

Placing the stress within a word can differentiate syntactic category, as in pro-***ject*** [verb] *vs.* ***pro***-ject [noun]. Even when stress placement within a word is not necessary to resolve ambiguity of syntactic category, it is still important to follow the rules of English word stress; otherwise speech sounds unnatural or non-English.

Intonation

The overall phonological intonation pattern of a sentence is generated independently of the subsequent realization of word stress placement, even though at the physical level realizing word stress may involve changes to f0. Descriptive models of phonological intonation [e.g. Pierrehumbert 1977] don't take into account physical constraints involved in the later stage of generating a waveform. For example, in some dialects of English for statements and *wh*-questions declination can be seen - a progressive lowering of f0 from a focal point within a sentence due to falling sub-glottal air pressure; declination does not form part of the phonological description of the sentence since it is a phonetic phenomenon.

Similarly one can consider inclination — a progressive raising of to from the start of the sentence to a focal point, though this constraint occurs less often than declination. Phonological requirements of local perturbations in intonation can be regarded as being superimposed on, or modulating the general inclination/declination pattern determined at the physical level.

Inclination and declination do not occur all the time in all languages; this suggests that as physical constraints they can be overridden. In other words, they can come under cognitive

control. Although intonation now seems well understood at the phonological level, how it is realized at the physical level as changes in to is less well modeled for human speech.

Rhythm

Word stress also plays an essential role in sentence rhythm in English. Stressed syllables tend to have longer duration than unstressed syllables. In particular, the vowel nuclei of stressed syllables show an increase in duration at the physical level, compared with the same vowels in unstressed syllables. This is complicated by the fact that phonological vowel reduction also occurs in English, where certain vowels occurring in stressed positions do not occur in unstressed positions: they are replaced by 'weaker' vowels. In addition, segments and syllables can be completely deleted in fast speech.

Researchers haven't been able to develop a good applications model at the phonetic level in the same way that intonation has been outlined at the phonological level. Research in phonetics has not yet produced a good model of rhythm for human speech based on studies of speech waveforms correlated with an appropriate higher level rhythm generator.

In order to model the phonetic realization of a pragmatic overlay, it's necessary to determine

- how the durations of individual phonetic segments or syllables relate to overall sentence rhythm physically, and
- how the speech waveform contributes to the perception of rhythm this is assumed to relate to phonology.

Despite incomplete applications models of intonation, rhythm and their interaction designers of speech synthesis systems have tried to develop independent strategies for producing outputs which sound as natural as possible. However, intonation and rhythm are still poorly realized in machine voice output.

Work has begun on modeling how phonological markers might be realized at the phonetic level. The speech synthesis system I have been using incorporates an algorithm for determining sentence intonation based on Pierrehumbert, and generates physical f0 contours [Morton 1992]. The rhythm generator has been done under SPRUCE [Tatham and Lewis 1992], and realizing segment and syllable durations based on current work at Essex.

To simulate emotion, the task is to realize the symbolic phonological and pragmatic representations at the acoustic phonetic level (Morton 1992). The relationship between these levels can be described as follows (Fig. 2):

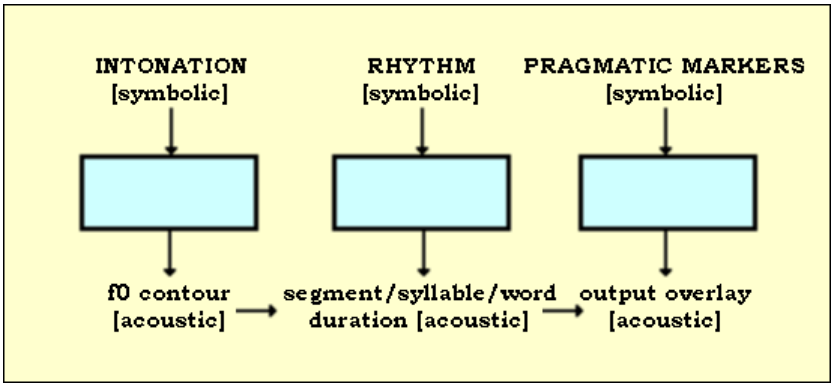

Fig. 2 The relationship between symbolic and acoustic levels in the model.

7. THE PRAGMATIC OVERLAY AND DECLINATION

As described earlier, pragmatic requirements trigger a modification process to the neutral durational and pitch (f0) assignments contained in the output files of the text-to-speech system. The modification process knows about and can implement:

a. declination — e.g.

Sentence type (1) High water's at 3:30.

Sentence stress locus can be assigned to *high*, or to *three.*

Sentence type (2) The plane, arriving at 5 o'clock, is on time.

Sentence stress focus assigned to *plane* or to *five,* the embedded phrase interrupts a decline from *plane* in the otherwise simple sentence *The plane is on time.*

b. how to modify declination.

The parser has identified word stress and phrasing (syntactic groupings of words) such that the overlay process can alter the computed to and timing of phonetic elements.

In a dialogue the ability to make pragmatically determined overlays facilitates easier interaction because of the addition of extra information. The question is: what is that information? and *how* is it to be physically realized? Thus, in the sentence

BA. flight 546, arriving at 5 o'clock, will be on time.

depending on the question asked of the system, different emphasis can be applied on particular words or phrases. e.g.

B. A. flight ... emphasized,

... *546,* ... emphasized,

... *arriving* ... as opposed to *departing*,

... *5 o'clock* ... rather than another time,

... *will* ... as opposed to *will not*.

All these overlays on the neutral rhythm and intonation of the sentence add relevant information in response to questions like:

Which airline? Which flight number? Arriving or departing? What time? Is it on time?

In conversation, human beings anticipate more than just words. A carefully designed dialogue system could, in fact, anticipate that the human user of the system will want answers to some of these questions — and include them by using the appropriate overlays on a response to a simple question which itself may not have requested this additional information.

By including this type of information, the dialogue becomes less labored for the system user. A dialogue system that does not have this capability will probably be less acceptable, since the user must ask a series of questions each having to elicit a small piece of information.

Declination is reported to be common only in read sentences [Ladd 1984], although other researchers feel it occurs in ordinary conversation in most dialects of English, is perceptually important [‘t Hart and Cohen 1973] and is often included in the computation of to contours in text-to-speech systems [O’Shaughnessy 1977].

Interruptions to inclination/declination form a major part of the way the pragmatic overlay works in the model being described here. Declination, even in a neutral sentence, will be reset at phrase boundaries, but how it is reset can trigger the perception of a particular pragmatic effect. Where resetting is not expected, some effects can be overlaid by a reset.

Because description of the detail of declination is not fully described in the literature, particularly for interruption and resetting to create particular effects, listening tests were carried out to provide evidence as to whether declination itself contributed to more acceptable voice output in synthesis systems. These tests are not being reported in full here — an example illustrates the point:

Abstract phonological intonation patterns were assigned to some sentences. The mapping rules forming the Essex to assignment algorithm (which links the abstract phonological pattern to f0 change) were applied. For example, the following sentence was processed:

B. A. flight 546, arriving at 5 o'clock, will be on time.

Processing involved splitting sentences automatically into syllables:

| ***B.*** | ***A.*** | ***flight*** | ***5*** | ***4*** | ***6,*** | ***ar*** | ***ri*** | ***ving*** | ***at*** | ***5*** | ***o'*** | ***clock,*** | ***will*** | ***be*** | ***on*** | ***time.*** |

in which | means syllable boundary

and obtaining a neutral intonation pattern:

H[| **B**. H** | **A**. H* | **flight** H* | **5** H* | **4** H* | **6**, H*L- | H]

L[| **ar** | **ri** H* | **ving** | **at** | **5** H* | **o'** | **clock**, H*L- | H]

L[| **will** | **be** | **on** | **time**. H*L- | L]

in which

- H[means high start to *sentence* or *phrase*
- H] means high finish to *sentence* or *phrase*
- L[means low start to *sentence* or *phrase*
- L] means low finish to *sentence* or *phrase*
- H* means *stressed syllable, high*
- H*L- means *stressed syllable, high with fall*
- H** means *sentence focus*.

In the test resynthesized sentences were used [Morton 1991b], rather than the output from a text-to-speech synthesis system. The reason for this was to ensure that listeners were not distracted by errors arising in the system, and to ensure the highest possible naturalness.

An f0 contour was automatically assigned to the sentences in accordance with the generated intonation markers with declination assigned, and without declination assigned.

Listening tests were conducted on the sentence test set to determine which set of sentences (those without declination *vs*. those with declination) were more acceptable and sounded consistently more natural. All were intelligible since they were based on resynthesized utterances.

Further listening tests were carried out using the test set with the automatic addition of pragmatic markers indicating emphasis to two aspects of the sentences. Applied to our example sentence:

1. ... *546*, ... emphasized in answer to the possible question *Which flight is arriving at 5 o'clock?*,

2. ... *will* ... emphasized in answer to the possible question *Is the flight on time?*.

In these tests the aim was to determine which sentences were more appropriate to the questions put. The results of the listening tests indicated that in general the presence of declination is preferred, though no declination may be preferred in response to an implied question. Declination with added pragmatic markers in answer to a question was more effective and to be preferred in a dialogue situation as opposed to plain neutral responses which include neither declination nor pragmatic modifications.

In the tests, lack of declination in the utterance sounded like a response to an implied question although the question may not have been asked. When the sentence is long, emphasis heard on words near the end of sentence may have unwanted effects.

e.g. *Air U.K flight 726 will arrive at Heathrow, Terminal 3, at 9:30.*

- in answer to *What time does Air U.K 726 arrive?*

spoken with declination (underline shows point of perceived emphasis):

1. Air U.K. flight 726 will arrive at Heathrow, Terminal 3, at 9:30.

spoken without declination:

2. Air U.K. flight 726 will arrive at Heathrow, Terminal 3, at 9:30.

Without declination three effects can occur:

- a sing-song impression,
- cognitive processing overload because of too much information in a short space of time and
- there are implied questions answered which may not have been asked — such as *Where does the flight arrive?*

8. USE IN HUMAN-MACHINE DIALOGUE SYSTEMS

Speech output within human-machine dialogue may be better if it is different from speech produced by a human reading or from speech during conversation between humans, since, certainly at the moment, automated dialogue systems are confined to information providing services.

If machine communication at the moment is limited to providing information, certain cues may be useful: for example, a falling intonation pattern incorporating declination is expected at sentence end, and signals the end of the requested piece of information. It signals to the human that he/she can now be the speaker and perhaps proceed with a further request.

Voice inquiry systems are asking for information: a firm friendly voice can signal that the information is accurate.

In addition to the categories read speech and spoken conversation, I suggest another category for human-machine dialogue systems called Information exchange. The category incorporates declination within information bearing sentences and includes a pragmatic component for special cases:

e.g. repeating a phrase or sentence if the listener has not understood the information:

Did you say 4:30? — No, I said 5:30.

(5:30 with pause and f0 change [Morton 1992]).

9. CONCLUSION

Potential users of synthetic speech are not satisfied with the quality of current synthetic speech output, in particular with respect to its lack of naturalness in interactive dialogue environments. It seems that to a great extent naturalness is marked by including some emotive content overlaid on the neutral meaning communicated by the syntax and semantics of a sentence given by the system in response to the human user.

The model outlined here assumes that naturalness can be characterized in part by pragmatically driven modifications to the fundamental frequency contours and timing of segments, syllables, and words in a sentence judged to be neutral.

Fundamental frequency contours and timing were generated by a text-to-speech system: the output file was intercepted and changes were made to the file, triggered by a pragmatic marker. In listening tests it was found that the overlays improved the acceptability of machine responses in the dialogue environment and that it was possible reliably to perceive the added emotive content of sentences.

REFERENCES

't Hart, J. and Cohen, A. (1973) Intonation by rule: a perceptual quest. *Journal of Phonetics*, Vol.1, pp. 309-327

Holmes, J.N. (1988) *Speech Synthesis and Recognition*. Wokingham, Van Nostrand Reinhold

Ladd, D.R. (1984) Declination: a review and some hypotheses. *Phonology Yearbook* 1, pp. 53-74

Morton, K. (1991a) Pragmatic phonetics, *in Advances in Speech, Hearing and Language Processing* (W.A. Ainsworth, ed.). London: JAI Press, pp. 17-55

Morton, K. (1991b) Expectations for assessment techniques applied to speech synthesis. *Proc. of the Institute of Acoustics, Acoustics '91*, Vol. 13, pp. 601-609

Morton, K. (1992) Adding emotion to synthetic speech dialogue systems. *Proc. Inl. Conf. on Spoken Language Processing*, Banff, Canada

O'Shaughnessy, D. (1977) Fundamental frequency by rule for a text-to-speech system. *Proc. IEEE Int. Conf. ASSP*, pp. 571-574

Tatham, M. and Lewis, E. (1992) Prosodic assignment in SPRUCE text-to-speech synthesis. *Proc. of the Institute of Acoustics* (this volume)