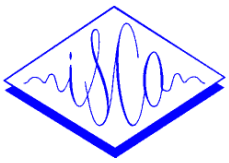

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

EXTRACTING MICROPROSODIC INFORMATION FROM DIPHONES - A SIMPLE WAY TO MODEL SEGMENTAL EFFECTS ON PROSODY FOR SYNTHETIC SPEECH

A. I. C. Monaghan

Department of Linguistics, University of Edinburgh
Adam Ferguson Building, 40 George Square, Edinburgh EH8 9LL, Scotland, UK

ABSTRACT

One of the major problems in the phonetic realisation of synthetic intonation is the modelling of local segmental effects on the course of F_0, despite the fact that in a system using concatenation synthesis these effects are already present in the units recorded from natural speech. The usual procedure in concatenation synthesis is to remove all F_0 information from the units and then reimpose a synthetic F_0 contour: however, it is possible to take advantage of the original F_0 information to model local segmental effects and thereby produce a more natural F_0 contour.

This paper proposes a method for extracting microprosodic information from the F_0 contours of diphones recorded from natural speech, and presents some preliminary results of its application in the TTS system developed at Edinburgh University's Centre for Speech Technology Research (CSTR) [3].

Problems with the work reported here are also discussed, as are directions for future research both on microprosody and on evaluation experiments [1]

I. INTRODUCTION

The segmental effects which are responsible for much of the perturbation of the fundamental frequency (F_0) contour in natural speech tend to occur at segment boundaries, i.e. where the vocal tract (and hence the waveform) undergoes rapid transitions from one relatively stable state to another. The units involved in concatenative synthesis are specifically chosen so as to start and end at stable states, with a single unit spanning each transition. In diphone units (or larger concatenative units), therefore, the segmental effects on F_0 will be contained within the relevant units. By approximating the smooth course of F_0 through such units, we can assess the location and degree of perturbation due to segmental factors: values for this perturbation can then be stored for each point in each unit, and applied to the smooth synthetic contour to recreate the natural perturbation of F_0.

The first step is to model the notional unperturbed course of F_0 through each unit. We assume that no more than one turning point is possible per diphone, given a target-and-transition model of intonational phonology such as Pierrehumbert's [13] or Ladd's [8] model. The course of F_0 in each diphone can therefore be modelled by a second-order approximation.

Once a smooth curve has been fitted to the pitch contour through a diphone, the deviation of each frame or pitch period from this curve can be measured. If this value is stored, it can be used at the time of resynthesis to perturb any synthetic contour imposed on that frame from that particular diphone: the segmental effect can thus be reproduced regardless of the particular contour shape or pitch imposed on the diphone.

This procedure was performed on a set of diphones for British English, and a pilot evaluation experiment was carried out involving pairwise comparisons by native listeners: subjects compared synthetic speech with and without the microprosodic perturbations extracted as above. Although the results of this experiment are not conclusive, they indicate that further work should be undertaken to investigate the full potential of this procedure for synthesising microprosody. Full details of the method proposed, and of the evaluation experiment, are given in the remainder of this paper.

II. RATIONALE

There has been considerable improvement over recent years in the segmental quality of synthetic speech, largely as a result of two fundamental departures from the 'traditional' phoneme-based formant synthesis methodology applied in systems such as MITalk [1] or INFOVOX [4]. The first of these departures, the concatenation of units of recorded speech, brought an obvious improvement in segmental quality within each concatenated unit but introduced unnatural transitions at the boundaries between units: the second departure provided a solution to modelling these transitions by concatenating diphone (or larger) units, thereby containing most of the transitions within a single unit. The majority of unrestricted high-quality speech synthesis is currently based on diphone concatenation (with various modifications), giving segmental quality comparable to natural speech in many cases (see for instance [16,17]).

This very increase in segmental quality has served to highlight the inadequacies of suprasegmental modelling in synthetic speech: as Terken & Lemeer [18] point out, improved segmental synthesis requires improved synthetic prosody. Although much work has been done on both the phonetic and the phonological aspects of prosody for speech synthesis (see for instance [8,9,10]), the modelling of segmental effects on the course of F_0 (known as microprosody) has been largely ignored.

Microprosodic effects have been shown to make an important contribution to the perceived naturalness of speech [14], but in most synthesis systems no attempt is made to model them. There appear to be two main reasons for this. Firstly, too little is known about the detailed phonetic manifestation of microprosody: opinions differ as to its precise characteristics, and no large-scale investigations have been undertaken. Secondly, while intonation contours are increasingly specified by a

[1] The author is grateful for the support of Economic and Social Research Council (ESRC) grant no. RR10565. He also wishes to acknowledge the help of numerous colleagues at Edinburgh, in particular Richard Caley, Steve Conway, Steve Isard and John Rye.

small number of target points per utterance, based on target-and-transition approaches such as those expounded by Bruce [2], Pierrehumbert [13] and Ladd [8], microprosody requires to be modelled almost millisecond by millisecond: this fine-grained modelling does not lend itself to inclusion in many synthesis architectures, and involves considerable extra computation. These two reasons have combined to ensure that where microprosody is modelled it is done either in a rather gross fashion (see for instance [5,15]) or by a simple randomisation (see for instance [6,15]). Silverman [14] presents substantial evidence to indicate that microprosody involves fine modifications to an underlying intonation contour, and that inappropriate modifications do not aid comprehension (although they may improve perceived naturalness): he demonstrates that something better than these crude and/or random attempts is required if microprosody is to be used to improve the quality of synthetic speech.

The object of this paper is to point out that, at least for synthesis systems based on diphone concatenation, there is a much simpler and more natural way to model microprosodic effects. Diphone units from recorded speech already contain microprosodic information, and if that information can be extracted it will solve both the practical and the theoretical problems normally associated with modelling microprosody for synthetic speech. The following sections present a simple method for extracting microprosody from diphones, together with a preliminary evaluation of the effect of adding that microprosody to synthetic speech.

III. MICROPROSODY FROM DIPHONES

The diphones used for the work described in this paper are those employed in the CSTR TTS system [3]. Except for those involving schwa and/or syllabic consonants, all the diphones are taken from the stressed syllables of isolated polysyllabic nonsense words produced by an adult male speaker of the "RP" accent of British English. They have been automatically pitch-marked for pitch-synchronous LPC synthesis on the basis of laryngograph traces, with pitch interpolated through unvoiced portions.

The natural F_0 contour through each diphone is a composite of the slowly-varying intonation contour and the quickly-varying microprosodic perturbations of that contour. The first step in extracting microprosodic information is to separate out these two components, and this was done by approximating the intonation contour and subtracting it from the composite contour to leave the microprosodic component. Given the intonational characteristics of English [7,13], we have assumed that there will be a maximum of one turning point (corresponding to an intonational target) per diphone. The intonation contour may therefore be approximated by a quadratic function: this ensures that the intonation contour is modelled as closely as possible without masking the microprosodic information. Approximations can be generated for every diphone in a particular accent, using a measurement of 'goodness of fit' such as the least-square method.

The next step is to encode the microprosodic information in the diphone database, so that it can be retrieved during synthesis. A value representing the measured deviation from the smooth contour was stored for each frame in each diphone, so that whenever that frame was used in synthesis the F_0 contour would be perturbed by the appropriate amount. Silverman [14] suggests that consonantal perturbations involve a fixed number of Hertz, regardless of speaker or pitch range: however, we feel that this is unlikely, given the physiological differences between speakers, and we therefore stored the perturbations as a percentage deviation from the smooth contour (in Hz). For present purposes there is little difference between the two methods, as the pitch range used in our synthesis is generally the same as or close to that used in the speaker's original production of the diphones: however, this point merits further investigation, and is returned to below.

IV. EVALUATION

An experiment was carried out to determine whether the inclusion of microprosodic information as extracted above would improve the quality of synthetic speech. The experiment essentially involved pair-wise comparisons of synthetic stimuli, with the subjects being required to state a preference for one stimulus over the other.

4.1 Stimuli

20 sentences were selected from the 200 sentences of the ATR phonetically-balanced speech database developed at CSTR. The criterion for selecting this set of sentences was the avoidance of long voiceless stretches which would make F_0 difficult to track both for speech analysis software and for human subjects. Two examples of the sentences used are given in (1) and (2).

(1) Bob milked the cows before he gathered the chickens' eggs.
(2) He caught a glimpse of what looked like a badger.

Each of the 20 selected sentences was synthesised in two different versions. Both versions were identical except for the details of their F_0 contours. A phonemic transcription of each sentence, based on a natural reading and including intonational markers, was synthesised first with an F_0 contour composed entirely of straight lines (as described by Ladd [8], Campbell et al. [3] and Monaghan & Ladd [12]) and then with the same straight-line contour perturbed by the microprosodic information in the diphones. These two different versions will be referred to as "FLAT" and "BUMPY" respectively.

4.2 Presentation

In a pairwise comparison test, subjects were presented with twenty pairs. Each pair comprised the two versions of a single sentence. In half the pairs the FLAT version came first, and in the other half the BUMPY version came first: the different presentation orders were randomised. The stimuli were played through high-quality speakers in a quiet room, with a very short (1 second) pause between the two members of the same pair and a much longer (5 second) pause between each pair.

For each pair, subjects were asked to tick one of two boxes to indicate their preference. The instructions were printed on the answer papers, and the exact wording of the instructions which the subjects received was as follows:

> *You will hear twenty pairs of utterances. Each pair consists of two versions of the same sentence. Your mission, should you choose to accept it, is to decide which version you prefer. For each pair, you must decide on your preference and indicate it by ticking a box.*
> *If you prefer the first member of a pair, tick the left-hand box.*
> *If you prefer the second member of a pair, tick the right-hand box.*
> *You must tick a box for each pair, even if you are not sure of your preference!*

These instructions were also read out to the subjects by the author before the experiment, and subjects were asked if they understood the task. There were 14 subjects, seven male and seven female, all adult normal-hearing native speakers of British English and all aged between 20 and 50. No subjects failed to fulfil the experimental task, although one or two did complain of difficulty in deciding on a preference. Subjects were also asked to mark their answer sheets with an 'M' or an 'F' to indicate their sex.

V. RESULTS

There was a total of 280 responses, 140 for each order. Table 1 gives the number of "correct" responses by subject, i.e. responses preferring the BUMPY stimulus: it also indicates the sex of each subject. Table 2 gives the proportion of correct responses for each order, and the total number of correct responses.

Table 1: "Correct" Responses by Subject

14	13	13	12	11	11	10
M	M	M	M	M	M	M
13	12	11	10	10	9	8
F	F	F	F	F	F	F

Table 2: "Correct" Responses by Order

BUMPY-FLAT	correct =	72/127	(0.56693)
FLAT-BUMPY	correct =	86/153	(0.56209)
Total	correct =	158/280	(0.56428)

Table 1 shows a marked effect of gender on subjects' performance: the female scores are noticeably lower (mean of 10.43) than the male scores (mean of 11.86). This is probably attributable to the sex of the speaker whose diphones and pitch range were used in the synthesis, but in any case this factor does not interest us further here.

It is clear from Table 2 that subjects tended to prefer the second stimulus regardless of the order of presemtation. Again, this is of no further interest to us here since there were equal numbers of stimulus pairs in each order: however, order effects are not unusual in this type of experiment and in this case the effect is probably due to the increased intelligibility of the second stimulus since its phonemic content is more predictable.

A t-test on the scores in Table 1 gives a probability of more than 80% (t=1.07, 13 degrees of freedom) that subjects consistently preferred the BUMPY stimuli. This result is not conclusive, but it is a strong indication that this pilot experiment is worth following up. Unfortunately, it has not been possible so far to conduct a larger-scale experiment, although it is expected that larger numbers of subjects and stimuli would produce a clearer result.

VI. DISCUSSION

The experiment above confirms the potential of this approach to extracting microprosody from diphones. It seems likely that the addition of such a component to synthetic speech will significantly improve its perceived quality. However, more work is required, both on applying the microprosodic information and on evaluating the results. Some points for future investigation are outlined here.

The present work deals with consonantal perturbations only, and leaves the question of vowel intrinsic pitch unaddressed. Various studies have shown that vowel intrinsic pitch has a significant influence on speakers' perception of F_0 contours, and as such is an important factor in speech synthesis. In principle, intrinsic pitch information should also be deducible from diphones by examination of the peak pitch of the various vowel diphones: however, this remains to be verified by future work.

As mentioned above, Silverman [14] claims that consonantal perturbations are linear in nature, with the absolute amount of perturbation remaining constant across changes in speaker and in pitch range. On the other hand, Silverman claims that vowel intrinsic pitch is closely dependent on the speaker's range. It seems counterintuitive that these effects should be so different, and for the time being we have modelled consonantal perturbations as being dependent on pitch range, but this area clearly requires further investigation before an optimum combination of intonation and microprosody can be achieved in synthetic speech.

Finally, the question of an appropriate evaluation metric requires closer examination. The comparison made above between a straight-line contour and a perturbed contour is a simple and effective one, but there are potential problems with its interpretation. Firstly, it has been shown that the introduction of random perturbation may improve the perceived naturalness of synthetic speech: it would therefore be interesting to compare our "principled" microprosody with random perturbations of the same order. Secondly, it is possible that the speech analysis techniques used above have added unnatural perturbations: it is not clear how this may be overcome, but perhaps a comparison with natural F_0 contours might be made as in [11]. Another possible refinement might be the imposition of a maximum deviation value or of some smoothing metric on the perturbed contour, to eliminate unnaturally large "bumps". It is hoped that investigations along these lines will be carried out in the near future.

VII. REFERENCES

[1] J. M. Allen, S. Hunnicutt & D. H. Klatt. *From Text to Speech: The MITalk System.* Cambridge: CUP. 1987.

[2] G. Bruce. *Swedish Word Accents in Sentence Perspective.* Lund: Gleerup. 1977.

[3] W. N. Campbell, S. D. Isard, A. I. C. Monaghan & J. Verhoeven. "Duration, Pitch and Diphones in the CSTR TTS System." *ICSLP*, pp. 825-828, 1990.

[4] R. Carlson, B. Granström & S. Hunnicutt. "Multi-Lingual Text-to-Speech Development and Applications." In W. A. Ainsworth (ed.), *Advances in Speech, Hearing & Language Processing* vol. 1, pp. 269-296. London: JAI Press. 1990.

[5] C. Choppy & J. S. Liénard. "Prosodie Automatique pour la Synthèse par Diphonèmes." *8èmes Journées d'Etude sur la Parole*, Aix-en-Provence, pp. 211-217, 1977.

[6] J. House. "Enlivening the Intonation in Text-to-Speech Synthesis: An 'Accent-Unit' Model." *ICPhS*, Tallinn, vol. 1 pp. 134-137, 1987.

[7] D. R. Ladd. *The Structure of Intonational Meaning: Evidence from English.* Bloomington: Indiana University Press. 1980.

[8] D. R. Ladd. "A Phonological Model of Intonation for Use in Speech Synthesis by Rule." *Eurospeech*, Edinburgh, vol. 2 pp. 21-24, 1987.

[9] A. I. C. Monaghan. "Rhythm & Stress Shift in Speech Synthesis." *Computer Speech & Language*, vol. 4, pp. 71-78, 1990.

[10] A. I. C. Monaghan. *Intonation in a Text-to-Speech Conversion System.* Ph.D. thesis, University of Edinburgh. 1991.

[11] A. I. C. Monaghan. "Evaluation of the Naturalness of Prosody Generated by the CSTR TTS System." *Eurospeech*, Genoa, pp. 883-886, 1991.

[12] A. I. C. Monaghan & D. R. Ladd. "Manipulating Synthetic Intonation for Speaker Characterisation." *ICASSP*, vol. 1 pp. 453-456, 1991.

[13] J. B. Pierrehumbert. *The Phonology and Phonetics of English Intonation.* Doctoral dissertation, Massachusetts Institute of Technology. 1980.

[14] K. E. A. Silverman. *The Structure and Processing of F_0 Contours.* Ph.D. thesis, Cambridge University. 1987.

[15] C. Sorin, D. Larreur & R. Llorca. "A Rhythm-based Prosodic Parser for Text-to-Speech Systems in French." *ICPhS*, Tallinn, vol. 1 pp. 125-128, 1987.

[16] H. A. Sydeserff, R. J. Caley, S. D. Isard, M. A. Jack, A. I. C. Monaghan & J. Verhoeven. "Evaluation of Speech Synthesis Techniques in a Comprehension Task." *Eurospeech*, Genoa, pp. 277-280, 1991.

[17] H. A. Sydeserff, R. J. Caley, S. D. Isard, M. A. Jack, A. I. C. Monaghan & J. Verhoeven. "Evaluation of Speech Synthesis Techniques in a Comprehension Task." To appear in *Speech Communication*, vol. 11, 1992.

[18] J. M. B. Terken & G. Lemeer. "Effects of Segmental Quality and Intonation on Quality Judgements for Texts and Utterances." *Journal of Phonetics*, vol. 16, pp. 453-457, 1988.