INTONATION BY RULE IN TEXT-TO-SPEECH APPLICATIONS

René Collier* and Jacques Terken*

ABSTRACT

The high degree of standardization that can be achieved in the intonational synthesis by rule of isolated utterances may not be applicable to that of longer texts, because the listener may become aware of the unnatural recurrence of invariably the same melodic recipe. Two experiments were run in which listeners evaluated the naturalness of a prose text that had been intoned with varying degrees of melodic standardization.

STANDARDIZED PITCH CONTOURS

The synthesis of fundamental frequency contours in a text-to-speech application requires at least two sorts of generalization: (1) a standard specification of the distinct types of elementary $F_0$ changes, and (2) a statement of their combinatory possibilities at the level of the $F_0$ contour. The criterion for what counts as a distinct type of $F_0$ change is a perceptual one: any two $F_0$ changes that sound as "different" pitch movements have to be considered as representatives of distinct intonational classes. Conversely, $F_0$ changes that produce a melodically "equivalent" impression have to be characterized as tokens of a common underlying type. This common origin can be expressed by giving all these equivalent $F_0$ changes the same standard acoustic specification with regard to their direction, magnitude, rate of change and timing. For instance, a "Type 1" rise in Dutch is defined as an increase of $F_0$ over an interval of 5 semitones (ST), with a rate of change of 50 ST/s (and, consequently, a duration of 100 ms), timed in such a way that its offset is located 40 ms beyond the vowel onset of an accented syllable. It is distinct from a "Type 2" rise, the offset of which coincides with the end of voicing in a clause-final syllable.

When the inventory of all the perceptually relevant pitch movements has been established and all the rises and falls have been given their standard acoustic specification, one can formulate rules that express the sequential constraints of these pitch movements at the level of the pitch contour, a melodic unit which is co-extensive with a phrase or clause. For example, in a Dutch contour a sequence of rises may be $<$Type 1+ Type 2$>$, not $<$Type 2 + Type 1$. By combining some 10 basic pitch movements in various ways, Dutch can build hundreds of different pitch contours (ref. 1). The same is true of British English, for which a similar melodic model has been developed. A sample of the contours generated by this model has been evaluated by native listeners, who assigned the same degree of "acceptability" to these standardized pitch contours as to their natural counterparts (ref. 2, ref. 3). In these evaluation experiments listeners were presented with isolated utterances. The question remains whether highly standardized contours would be equally acceptable if used in longer stretches of speech, for instance in the synthesis of a running text. There are at least two reasons why this might not be the case.

Indeed, it may be that listeners who are exposed to a longer text with synthetic intonation readily detect the standardization of the pitch movements and dislike the unnatural recurrence of their invariable size, rate of change, etc. A second cause of dissatisfaction may be that there is not enough variation in the choice of successive contour types. Indeed, although the inventory of possible contours is rich, the choice of a suitable variant may be limited or even impossible if it depends on syntactic or pragmatic information that is not available in the input text. Therefore, in actual practice just one contour type is often used all through the text, namely one that is considered "neutrally declarative" and does not take (clause-internal) syntactic information into account.

*University of Antwerp (Belgium); **Institute for Perception Research (Eindhoven, the Netherlands)
In summary, the standardization of the acoustic properties of the pitch movements and their predictable recurrence in just one type of contour may have an adverse effect on the perceived naturalness of synthetic intonation in running speech. A synthesis scheme that works well on single utterances may be too limited a strategy for use in a text-to-speech application.

VARYING STANDARDIZATION AND PERCEIVED NATURALNESS

Experiment 1

**Experimental set-up**. The purpose of the experiment was to evaluate the naturalness of variable amounts of standardization applied to the intonation of a short text. As a starting point we had an experienced non-professional speaker read a prose passage in Dutch. The recorded speech was analysed and resynthesized using the LPC system at IPO. We will refer to this version of the text as the original one (ORIG). As a first degree of stylization we replaced the original course of $F0$ by an approximation with a minimum number of straight lines, without losing the impression of perceptual equality between this "close copy" version (henceforth: CC) and ORIG. The CC variant mainly eliminates the micro-intonational modulations from the original $F0$ curve and respects whatever variation there is in the overall shape of the individual pitch movements and in the rate of declination of the individual clauses. In the third version of the text all the pitch movements were given an identical standard size of 6 $ST$ and the declination rate was computed as function of utterance length, using a standard formula (ref. 4). This manipulation results in a version that respects all of the original variation in the choice of contour types, but implements each type according to the standard specification for the pitch movements of which it is composed and superimposes these movements on a declination line with a standardized slope (henceforth: version STAND). The fairly rich variation of contour types in STAND is possible because the speaker, whose choices have been copied here, can rely on his linguistic knowledge (regarding, for instance, the location and the depth of syntactic boundaries). If no syntactic knowledge is available and only the location of the accented syllables is given, then basically just one contour type is available: each accented syllable is given a steep rise of 6 $ST$, followed by a gradual fall (over the same interval) that extends up to the onset of the rise on the next accented syllable, etc. The last two accented syllables are given a steep rise and a steep fall, respectively. The declination slope is standardized as well. This contour variant is called "sawtooth + flat hat" (henceforth: SAW).

The four versions of the text were presented to 11 listeners, first as wholes and second as single utterances. The order of presentation of the different versions was varied over 4 subgroups of subjects. The listeners were asked to evaluate the "naturalness" of each text and each individual utterance and to express their ratings on a 10-point scale.

**Results.** The outcome of the experiment is reproduced in Table I below, which lists the average scores for each condition.

<table>
<thead>
<tr>
<th>TEXT</th>
<th>ORIG</th>
<th>CC</th>
<th>STAND</th>
<th>SAW</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTT'S</td>
<td>7.72</td>
<td>7.25</td>
<td>5.97</td>
<td>5.80</td>
</tr>
</tbody>
</table>

Table I. Average scores (max.=10) in each condition for the text presented as a whole and in utterance-by-utterance fashion.

The two modes of presentation produce similar results, viz. an overall difference between conditions ($p<.001$). Post hoc analysis reveals that the difference between (ORIG,CC) on the one hand and (STAND,SAW) on the other is significant. Apparently, the two standardized versions (STAND and SAW) find little favour with the listeners. Texts that simultaneously lack variation in the excursion size of the pitch movements and in the rate of declination sound far less natural than the "human" version. On the other hand, the listeners
do not make a distinction between the STAND version, with a rich variety of contour types, and the SAW version, which makes use of just one contour type throughout the text.

**Experiment 2**

*Experimental set-up.* In Experiment 1 the standardization was double: it affected both the size of the pitch movements and the rate of declination. In order to evaluate the effect of these two variables separately, we ran a second experiment, with a different speaker and a different text. The following four versions of the text were produced:

1. STYL is a stylized approximation of the input text in which the original size of the pitch movements has been quantized in four steps (3, 6, 9 or 12 ST) and the onset and offset values of the declination lines have been copied from the original;
2. STE has the same variable declination rate as STYL and the same types of pitch movement, but their excursion size has now been fixed at 6 ST;
3. STD has the same (quantized) pitch movement sizes as STYL, but the declination rate has been standardized;
4. STED has both a fixed size for the pitch movements (6 ST) and a fixed declination rate. This version is comparable to the STAND version of Experiment 1. In particular, the alternation of different contour types is still present (as opposed to the SAW version of Experiment 1).

A panel of 12 listeners rated the "naturalness" of the four versions of the text on a 10-point scale in the same two modes of presentation as in Experiment 1.

**Results.** Table II gives an overview of the results in terms of the average scores for each condition.

<table>
<thead>
<tr>
<th></th>
<th>STYL</th>
<th>STE</th>
<th>STD</th>
<th>STED</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEXT</td>
<td>7.25</td>
<td>7.00</td>
<td>7.25</td>
<td>6.58</td>
</tr>
<tr>
<td>UTT'S</td>
<td>7.02</td>
<td>7.04</td>
<td>7.16</td>
<td>6.73</td>
</tr>
</tbody>
</table>

Table II. Average scores (max. = 10) in each condition for the text presented as a whole and in utterance-by-utterance fashion.

It appears that the listeners make no significant distinction between the four versions of the text. Post hoc analysis indicates a significant interaction between conditions and subjects: some listeners prefer the STE version, others like the STD version best. There is some indication that the double standardization of STED is less satisfactory than the single one in STE and STD, but the difference is not significant. Also, the fact that STED is not really worse than STYL is at variance with the significant preference of ORIG over STAND in Experiment 1. However, it is not excluded that, despite similar manipulations, the resulting differences between conditions are smaller in this experiment than in the first one. Detailed pairwise comparisons of the stimuli are presently being made.

**GENERAL DISCUSSION**

Our preliminary investigation seems to indicate that listeners object to an extensive standardization of pitch contours in synthesized texts. On the other hand, they appear not to mind the repetition of the same contour type all through the text, as long as they hear some variation in either the size of the pitch movements or in the rate of declination. Which sort of variation is to be incorporated in future intonation synthesis schemes depends on the following considerations. Important variations in the size of the pitch movements are likely to produce corresponding differences in the perceived degree of accentuation on the syllables that carry these pitch changes. Consequently, these variations cannot be random but have to
take into account the amount of emphasis that a particular word needs to receive. As long as it remains difficult to correctly predict the location of accented syllables in an absolute way, a rule-governed assignment of degrees of accentuation is an even more challenging task. It has to be based, among other things, on contextual semantic information. As far as the variation of the declination rate is concerned, the most conspicuous way to achieve it appears to be to "reset" the baseline at utterance-internal syntactic boundaries. The nature and the depth of these boundaries appears to play a crucial role in the decision whether a reset is optional, obligatory or forbidden. Consequently, the correct control of declination resetting depends on the availability of syntactic information concerning the input text, which then has to be parsed before the intonational rules can be applied.

References

2. J R de Pijper, Modelling British English intonation (Foris, Dordrecht, 1983)