AUTOMATIC SYNTHESIS OF NATURAL-SOUNDING INTONATION FOR TEXT-TO-SPEECH CONVERSION IN DUTCH

J.M.B. Terken, R. Collier
Institute for Perception Research
P.O.Box 513, 5600 MB Eindhoven, the Netherlands

Abstract

A set of rules is proposed for the automatic synthesis of natural-sounding intonation as part of speech synthesis in Dutch from unrestricted text. Results of a formal perceptual evaluation show that the synthetic intonation is judged to be as natural as human intonation for isolated utterances; for texts, additional provisions are required to model contributions of text structure. It is suggested that the regularities expressed by the rules may apply to other languages as well.

INTRODUCTION

In devising acceptable intonation for unrestricted text, such as for automatic text-to-speech conversion, we cannot suffice with copying the pitch contours of prerecorded utterances, because this would confine us to a limited set of fixed messages. Rather, we must formulate a set of rules which produce natural-sounding pitch contours for utterances that may have never been spoken before. In order to do so, we must use our knowledge of the structure and function of intonation in human speech.

The structure and function of intonation can be described in terms of three main components: tune, scaling and declination. Tune refers to the succession of pitch movements, scaling refers to the size of the pitch movements, declination concerns the onset and offset pitch of successive intonation phrases. Each component contributes to the meaning of an utterance:

- the tune contributes to the reading of the utterance; among other things, it reflects the speaker's attitude; that is, it determines whether the utterance is to be interpreted as simply declarative or as expressing a more marked emotional state such as amazement or indignation, etcetera;
- the scaling of successive pitch movements contributes to the perceived prominence relations between words or word groups; that is, it indicates the relative importance of portions of the utterance;
- the declination contributes to the perceived relations between phrases and utterances; that is, it contributes to the perception of a succession of phrases as a coherent piece of text.

Summarizing, we can say that intonation in human speech makes a distinct contribution to the meaning of the text, and that this contribution merges with the contributions from the word level and the syntactic level to express the intention of the speaker. Among other things, this implies that in the normal case there will be no conflict between meanings from different levels, because the speaker tunes his intonational choices to the choices made at other levels.

In text-to-speech conversion we are unable to determine what overall meaning each utterance should express: we have no access to the meanings of words and syntactic devices, and therefore we cannot determine which intonational choices would be appropriate for each utterance. In practice, this confines us to neutral, non-biasing intonational choices, so that the meanings contributed by the intonation will at least not conflict with the meanings contributed by the other levels of the text.

NEUTRAL INTONATION

Investigations of human intonation (1,2,3,4) have shown that speakers have available a limited set of tunes; dependent on its length, each utterance is segmented into one or more intonation phrases; for each intonation phrase, the chosen tune is associated with the last prominent word in the phrase, the nucleus; in general, the realization of prenuclear accents is independent of the realization of the tune. The size of the pitch movements on prenuclear and nuclear accents is variable, and affects the perceived prominence of the accented words. Also, the onset pitch of the whole utterance is variable, whereas the offset pitch is rather fixed for each individual speaker; it has been argued that the onset pitch can be derived by a simple formula based on the duration of the utterance.

From these investigations we may derive a number of measures for simulating neutral intonation in Dutch:

1. one specific tune, which has been dubbed the hat pattern, is the most frequent one in neutral, factual statements and questions;
2. the pitch movements on words to be rendered prominent may be given a fixed size of 6 semitones; this will make the accented words to be perceived equally prominent;

3. as mentioned above, the declination rate may be obtained by computing the onset pitch from a simple formula based on the duration of the utterance, given a fixed offset pitch;

4. sentence-internal clause boundaries are signalled by a non-prominence-lending rise on the final syllable; if the time interval between the final accented syllable and the clause boundary is short, the final accent is realized by a rising pitch movement, and the pitch remains high up to the boundary.

A simple set of rules, implementing these measures, produces acceptable intonation for short, isolated utterances, such as are produced by aids for speech-impaired people (5). We will use the label VRD hereafter to refer to this first attempt at synthetic intonation (5,6). Additional experimentation has shown that this neutral intonation is far less acceptable for coherent texts (7). Thus, in order to produce natural intonation for coherent text, we must see how we can diversify the output of our rules, within the limits set before, namely that we are confined to neutral intonation, since the system has no access to a meaning representation of the text.

NATURAL INTONATION FOR SYNTHETIC SPEECH

The results of the study on the appreciation of synthetic intonation mentioned above (7) can be summarized as follows. A spoken text was manipulated in such a way that components of the intonation were modelled after a human speaker or produced by VRD rules. The properties of VRD most strongly disapproved of in longer stretches of speech were standardization of movement size and of declination. Standardization of the choice of tune was not strongly disapproved of. Thus, if we can appropriately model the movement size and declination rate after a human speaker, we may expect to arrive at more acceptable intonation for longer stretches of speech.

Additional research has suggested the following improvements:

1. longer utterances should be divided into smaller parts, which we label intonation phrases; these phrases correspond with clauses; for longer clauses, one or more additional phrase boundaries coincide with within-clause major constituent boundaries (NP, PP, AP, VP);

2. the offset pitch of each phrase depends on whether the phrase is sentence-final or not: sentence-final phrases have a lower offset pitch than non-final phrases; the onset pitch for each phrase is always higher than the offset pitch and is computed on the basis of the offset pitch and the phrase duration;

3. for each phrase, the size of the pitch movements decreases in a simple left-to-right manner, so that movements at the end of the phrase are smaller than movements in the beginning of the phrase.

We have added a fourth measure which has not been inspired by analyses of human intonation but by informal listening to the synthetic intonation resulting from these three improvements:

4. in phrases with at least two words to be accented, two different realisations of the same tune are applied in turn; the two realisations are perceptually clearly distinct (8); the different realisations have different pragmatic values, but it is not yet clear exactly what determines the choice between them.

We will use the label TEST for the implementation of the resulting, experimental set of rules.

PERCEPTUAL EVALUATION

The two sets of rules for synthetic intonation, VRD and TEST, have been incorporated into an experimental system for synthesizing speech from orthographic input, based on diphone concatenation (6,9). With this system, two different intonational versions of a text were synthesized. The text consisted mainly of long utterances (with a mean number of 17 words per sentence, ranging from 9 to 30). In addition, we used the system to produce a set of LPC-parameters for each sentence, rather than sampled waveforms, so that we could manipulate the pitch parameter. In this way, a third version of the text was produced by modelling the pitch after a rendering of the text by a professional reader and then synthesizing the speech files; we will label this version HUMAN. The pitch register in the HUMAN version was scaled down so that the mean offset pitch of all utterances matched the offset pitch of the versions with synthetic intonation. Except for pitch, the three versions were identical in all respects; this includes the location of accents and prosodic boundaries, which were copied from the human speaker.

The three versions of the text were presented to 13 experienced phoneticians, who were asked to judge the naturalness of the intonation in each version on a 10-point scale, 1 corresponding to "very unnatural", 10 corresponding to "very natural". The judgment was to be performed in two stages. First, all sentences constituting the three versions were scrambled and presented in isolation; next, the different versions of the text were presented as a whole. The sentences and the texts were presented in three different orders so as to counterbalance order effects. The results are shown in table I.
Table I. Mean “naturalness” judgments and standard deviations for three different intonations in diphone speech, for isolated utterances (k=19) and integral texts, by 13 experienced phoneticians. The judgments are expressed on a 10-point scale. “1” stands for “very unnatural”, “10” for “very natural”. Standard deviations of means for isolated utterances are standard deviations of the means of individual subjects’ judgments of 19 utterances.

<table>
<thead>
<tr>
<th></th>
<th>HUMAN</th>
<th>TEST</th>
<th>VRD</th>
</tr>
</thead>
<tbody>
<tr>
<td>isolated utterances</td>
<td>6.1</td>
<td>6.0</td>
<td>4.9</td>
</tr>
<tr>
<td>s.d.</td>
<td>0.9</td>
<td>0.8</td>
<td>1.2</td>
</tr>
<tr>
<td>texts</td>
<td>7.5</td>
<td>5.5</td>
<td>4.2</td>
</tr>
<tr>
<td>s.d.</td>
<td>1.2</td>
<td>1.4</td>
<td>1.8</td>
</tr>
</tbody>
</table>

For isolated utterances, TEST performs as well as HUMAN, and is judged significantly better than VRD. For integral texts, HUMAN is judged better than in the “isolated utterances” condition, but TEST and VRD receive lower judgments in integral texts than in the “isolated utterances” condition. Again, TEST performs better than VRD. These findings suggest that, at the sentence level, the improved set of rules for synthetic intonation matches the performance of a human reader; at the text level, the judges clearly favour the HUMAN reader’s intonational realisation of intersentential relations, i.e. of text intonation, and censure the absence of text intonation in TEST and VRD.

CONCLUDING REMARKS

From the results of the perceptual evaluation we may conclude that the improvements in the revised set of rules for synthetic intonation are appreciated by the judges. The improvements mainly concern the level of isolated utterances. The proposed measures may apply to other languages as well. In fact, some of the measures proposed show a strong resemblance with characteristics of intonation contours of Standard Danish (10). Additional improvements seem to be possible at the level of text intonation, so that a text will indeed sound as a text and not as a succession of isolated utterances.

Acknowledgements

This research was supported by the Foundation for Speech Technology, which is sponsored by the Dutch National Programme for the Advancement of Information Technology (SPIN), as part of the research programme “Analysis and Synthesis of Speech”.

References