



METHODS FOR THE SIMULATION OF NATURAL INTONATION IN THE "SYRUB"
TEXT-TO-SPEECH SYSTEM FOR UNRESTRICTED GERMAN TEXT
M.Kugler-Kruse, R.Posmyk*

ABSTRACT

The SYRUB text-to-speech system had been designed to translate unrestricted German text into a sequence of parameters that can be used to drive different speech synthesizers (ref 1). The interface parameters consist of the phoneme code, fundamental frequency (f_0), sound duration, and sound intensity. For synthesizers that do not operate with phonemes as the basic units additional information, e.g., for controlling coarticulation, is available. To produce a fairly natural intonation, several steps are required: a morphemic analysis generates information for phonemization, word stress assignment, and segmentation into phonetic syllables. An end grapheme analysis supplies word classes needed to mark phrase boundaries. For f_0 assignment a declination line and stress patterns are applied. Sound duration is governed either by context-dependent rules or by isochrony, generating rhythmic speech.

PROVIDING INFORMATION FOR INTONATION

The process which generates sentence prosody can generally be divided into two parts: At a first step, structural information about the sentences has to be obtained by reconstructing the syntax. Much of this required information can be derived from the word processing level, at which probable morpheme boundaries as well as possible word classes of single words are already marked by a cluster analysis (Ruhl (ref 2)). After subsequent analyses of prefixes and suffixes words are phonemized and word stresses are marked. In a second step, algorithms can take this information to generate the prosodic parameters (i.e. fundamental frequency, duration, intensity).

One aspect of intra-word prosody is represented by the phonetic syllable. In contrast to written language, where the unit 'syllable' is clearly defined, such a definition does not exist for spoken language. Ortman (ref 3) introduced a simple algorithm based on statistics for his definition of phonetic syllables, which was implemented in the SYRUB system. The German language includes many long composite words that have to be segmented prior to the application of any analysis of smaller sub-units (e.g. syllable segmentation). Based on the morpheme boundaries, those beginnings of prefixes and of stems are determined, which are certain boundaries of phonetic syllables, whereas the beginnings of suffixes do not always coincide with syllable boundaries. The basic rule for the determination of a phonetic syllable is the requirement that it consists of exactly one vowel. The following boundaries can be given:

V_1-V_2 , $V_1-C_1V_2$ with V_1 being a long vowel, resp.

$V_1C_1-C_1V_2$ with V_1 being a short vowel (consonant C_1 doubles).

Boundaries within longer consonant clusters are generally in front of the last consonant of a cluster (apart from some inseparable sequences, where the boundary is in front of this sequence, e.g. 'Re-klame' vs. 'welt-lich'); the marking of those syllable boundaries, however, relies much on an exact determination of morpheme boundaries. Especially the recognition of the 'Fugen-s', which is difficult for a rule-based system, proves to be a source of error.

*Lehrstuhl für allgemeine Elektrotechnik und Akustik,
Ruhr-Universität Bochum, PO Box 102148, D-4630 Bochum 1, FRG

SENTENCE ANALYSIS

The main task of the sentence analysis consists of the word classification. For function words this can be done by a lookup in a limited exceptions list that was introduced due to the frequent irregularities concerning the phonemization of these words in German. By this means an unambiguous determination of the word classes is achieved. For all other words an analysis of the end graphemes is done: After cutting off the endings, which leads to a fairly unambiguous word class only in very few cases (e.g. /-er/(not verb), /-el/(noun)), a matching of the remaining word end with an end grapheme list takes place. This list consists of a variable initial context, the end graphemes themselves, and possible endings with a corresponding word class. The exactness of this classification varies, but generally it allows the decision whether the word belongs to a nominal or to a verbal group. In those cases, where a word cannot be classified, it is assigned to a verbal group.

Using the information about word classes, single words can be concatenated to form phrases. These phrases are sequences of words that cannot be segmented anymore without distorting the meaning of the sentence. Within spoken language, phrases are marked by inserting pauses of various length or by lengthening of the pre-boundary syllable. The distribution of pauses, however, depends much on the speech rate in general and, of course, on the speaker's comprehension of the meaning. Since in the SYRUB system semantics is not taken into consideration, the grouping of words must be accomplished by syntactical rules that are derived from transition probabilities for the combination of words. The classification system that applies to the exceptions list and to the end grapheme analysis supplies up to 67 different word classes (without punctuation). To build a rule system that is applicable in general, word grouping for the most frequent word classes is sufficient, since about 60 percent of the sentences are built from only 4 syntax plans (i.e. S/P/acc-0, S/P/prep-0, S/P, S/P/Apx). By including merely 12 rules for common word classes, the amount of words that are processed correctly is more than 98 percent, referring to a text corpus of 1796 sentences.

SENTENCE ACCENT

With the detailed syntactic classification and information on word stress, phrase and sentence accents are determined. These accents serve as the basis for an assignment of fundamental frequency (f_0) and duration contour within the sentence.

At first, a phrase accent is assigned to the last content word within each phrase; thereafter the last phrase in a sentence gets the sentence accent. All remaining main stresses on the word level are reduced consecutively according to the nuclear stress rule. The stronger the accent the more pronounced is the f_0 rise or fall. f_0 values vary between 500 and 1000 cent. If two equally strong stresses follow each other the value of the first one is reduced and its corresponding vowel is lengthened (ref 4). This processing leads to a labelled sentence with weighted accents according to the sentence structure. Similar deductions were introduced by *Bannert* (ref 5).

FUNDAMENTAL FREQUENCY

Information about the type of sentence (statement, question, ...) and about a preliminary structuring (brackets, quotes, comma or colon as markers for sub-sentences, ...) is supplied by the text-preprocessing

module of the SYRUB system. This information is used to determine the overall pitch contour. Concerning the type of sentence, questions and statements mainly differ in f_0 at the last stress position. According to the findings of *Isacenko* and *Schädlich* (ref 6) a 'post-ictic' f_0 rise is assigned to questions. Statements are characterized by a final f_0 fall late in the last stressed syllable of a sentence (ref 7).

For the realization of the overall pitch contour, two different strategies are applied: One is based on the declination line or 'baseline' approach by *t'Hart* and *Cohen* (ref 7), the other on stress patterns.

The declination line approach combines a slow decrease of f_0 over the phrase with a quick f_0 change at stress positions. In the SYRUB system the declination line contour is applied to sub-sentences. Within a sub-sentence the decrease of f_0 depends on the length of the utterance: if the utterance does not last longer than 3 seconds, the decrease is set to 50 cent/100 ms; otherwise the fall becomes less steep. Alternatively, a baseline can be applied to single words with a reset at the main stress position.

Depending on the position within the text upper and lower limits for f_0 are set: At the beginning of a paragraph the range for a f_0 variation is wider than at the end.

Concerning the correlation between accents and f_0 the number of phrases in a sentence equals the number of non-final f_0 falls plus one (*Isacenko* and *Schädlich*). The phrase accent is always connected to a rapid change to a different f_0 level (i.e. an offset from the declination line). If the new level is kept up to the next accent position the result is a hat shaped pattern; if, however, f_0 returns to the previous level after the stressed syllable immediately, an accent emphasizing f_0 pattern is formed. In order to achieve a non-deterministic intonation pattern, in the SYRUB system a random distribution of non-final f_0 rises and falls is currently used. Depending on the value (n) of the accent rises and falls of f_0 are set to $n \cdot 400$ cent/100 ms (*t'Hart* & *Cohen*); in the final main stress position n equals 1, in the remaining positions stresses are reduced ($0 < n < 1$).

DURATION

The duration of each sound depends on its phonetic context, its position within the word and within the sentence, its word class, and its stress position. Two methods for dealing with the dependencies mentioned above have been implemented: context-dependent sound duration (after *Umeda*) and the principle of isochrony.

Umeda's findings on context-dependent duration in American English have been adapted to the German language. Her formula for the calculation of vowel duration (T) gives a relationship of consonantic context (C), morphosyntactic context (S), and vowel-specific duration (T_0 : minimal duration, K_1 , K_2):

$$T = T_0 + S \cdot (K_1 + C \cdot K_2) \quad (1)$$

Umeda (ref 8) groups sounds in three main classes depending on their position within the word (initial, medial, final). She distinguishes between stressed and unstressed vowels; in the SYRUB system the minimal duration is adapted to the stress values described above.

The values for the parameters, however, are language-specific and thus had to be determined by hearing experiments. In the current implementation the sound duration is computed from those parameters; these are stored in a look-up table. As the phonotactic rules for German differ from those for English, rules for consonantic contexts had to be modified accordingly. A positive or negative correction factor based on the sound context is added to determine the final consonant duration.

Another method for assigning sound duration generates weak isochrony (cf. *Witten* (ref 9)). Weak isochrony is used, since it respects minimal and maximal durations of sounds and thus avoids an intolerable overreduction. The standard-foot-duration (t) of a one-syllable foot (T) given in the formula (2) is a basis for a rhythmic speech. The speech rate can be varied externally by modifying the duration of a standard foot.

$$t = (1 + k \cdot (n - 1)) \cdot T \quad (2)$$

(n is the number of syllables in the foot, k is a constant). In experiments $k = 0.2$ yielded the most natural value. Depending on the different syllable types the foot duration is distributed to the syllables.

Sound duration is set according to the phoneme type, which is either extrinsic or intrinsic (cf. *Lawrence* (ref 10)). Extrinsic phonemes are more subject to context variations than intrinsic phonemes are. Isochrony is applied to phrases, not to whole sentences; the last syllable of a phrase is lengthened to serve as a boundary marker and thereby establish the rhythmic structure.

INTENSITY

For the German language only few data about the correlation of intensity and stress have been acquired. In the SYRUB system intensity changes coincide with f_0 changes. The resulting deterministic effect is reduced by applying a jitter (positive and negative random changes).

ACKNOWLEDGEMENT

Only little of this work could have been carried out without the work of H.-W. Rühl and W. Kulas.

1. Kulas, W.; Blauert, J.: German Text-To-Phoneme Software drives any Speech Synthesizer, *SPEECH TECH'86* (1986), pp 95-98
2. Rühl, H.-W.: Sprachsynthese nach Regeln für unbeschränkten deutschen Text, Diss., Bochum (1984)
3. Ortmann, W.D.: Sprechsilben im Deutschen, München (1980)
4. Bolinger, D.: Intonation and its Parts, London (1985)
5. Bannert, R.: Modellskizze für die deutsche Intonation, in: *Zeitschrift für Literaturwissenschaften* 49 (1983), pp 9-34
6. Isacenko, A.V.; Schädlich, H.-J.: Untersuchungen über die deutsche Satzintonation, in: *Studia Grammatica III* (1966), pp 7-67
7. t'Hart, J.; Cohen, A.: Intonation by rule: A perceptual quest, in: *Journal of Phonetics* 1 (1973), pp 309-327
8. Umeda, N.: Consonant duration in American English, in: *JASA* 58, No 2 (1977), pp 846-858
9. Witten, I.H.: A flexible scheme for assigning timing and pitch to synthetic speech, in: *Language and Speech* 20 (1977), pp 240-260
10. Lawrence, W.: The phoneme, the syllable and the parameter track, in: *Proc. Speech Comm. Seminar, Stockholm* (1974)