Automatic Prosody Generation – a Model for Hungarian

*Gábor Olaszy, Géza Németh, Péter Olaszi

*Kempelen Farkas Speech Research Laboratory, Hungarian Academy of Sciences, Hungary
Dept. of Telecommunications & Telematics, Budapest University of Technology and Economics,
{olaszy,nemeth,olaszi}@alpha.ttt.bme.hu

Abstract

Prosody generation relates to the composition of the detailed time structure of continuous speech and the realisation of the fundamental frequency and intensity structure, embedded in it. In our model a complex function set is described for the three prosody components for read speech. Each of the three prosody components is modelled separately by a three-step procedure. It was found that the correct modelling of the time structure (sound durations and breaks) is the most important component of prosody, because the time structure gives the framework of the melody and intensity structure. A new method, based on indirect determination of specific sound durations was developed. Final duration values are calculated from the specific durations in two further steps. The procedure starts on the lowest level (1st step on segmental level) and gradually goes towards higher levels, ending at sentence level. This duration model describes the behaviour of sound durations in continuous read speech. Modelling of fundamental frequency changes is also described by three levels, starting with rules on sentence level, followed by rules for word and syllable level. Finally micro intonation on sound level is applied. Similarly a three level model serves to describe the intensity structure, i.e. rules applied on sounds, words and on the complete sentence. The models of the three prosody components have influence on each other during prosody generation. Cross effects among them are also mentioned. The model can be applied in speech research and in applications (synthesis and recognition). It was tested for Hungarian.

Keywords: prosody generation, three level model, specific sound durations, word level duration map

1. Introduction

Modelling of prosody generation is in the focus of speech research these days. If functions can be determined for modelling the prosody of speech, the results can help speech recognition and can improve the speech quality of TTS systems. Furthermore such models can be used in general speech research, for example to help the evaluation of phonological rules describing a given intonation system for a language theoretically or to link phonological representations with phonetic realisations. For these complex reasons prosody components have been investigated intensively during the latest years (e.g. [1,5,6,8]) and the results have been applied mainly in TTS systems and in some cases in ASR solutions as well. A prosody model is language dependent to certain extent. As to the description of the time structure of a language, only the general behaviour of sounds can be determined and compared with other languages. Such general feature is, for example, that the mean duration of vowels depends on the height of the tongue. Detailed modelling of the time structure needs language specific rules.

The presented duration model can be used for other languages, but the detailed rules in this paper are valid for Hungarian only. The same concerns modelling of fundamental frequency and intensity structure.

2. The time structure model

The concept of this duration model follows the theoretical separation of speech into segmental and suprasegmental level. The segmental level represents the basis which is produced independently of the intention of the speaker (speech without prosody but having the correct specific duration values of the sounds and the distribution of durations, the correct, language specific timing ratio among speech sounds). On suprasegmental level the word structure and the sentence level effects influence sound durations. The duration model contains three levels for the calculation of final durations of speech sounds for continuous speech.

Level 1. Determination of segmental level specific durations. Specific duration in this model is defined by two features: (i) the duration of the given speech sound, is influenced only by the articulation of adjacent sounds, (ii) the specific sound duration does not contain suprasegmental effects, i.e. segmental level speech represents the source for measurements. In the model specific durations give the basic value for every sound of the given utterance for continuous speech. These values are used for further calculations in level 2 and 3.

Level 2. Modification of specific durations based on the sound map function of words on suprasegmental level. The function contains the following variables: the type of the sentence, the length of the word, the phonologically short and long sounds (quantity class) in the word, the type of sounds, and the order of sounds. At the end of this step the sound durations in the utterance are very close to the final values.

Level 3. Final durations are formed at this level. Results got in level 2 are modified by the variables: word length, position of the word in the sentence, phrase boundaries. Durations are set about 98% correctly at this level. Breaks are set separately at certain phrase boundaries and between sentences.

2.1 Determination of specific durations

An indirect method using the combination of speech synthesis and perceptual evaluation was developed to determine the specific durations for every sound in every sound combination in continuous speech. Direct measurements showed that besides the effect of articulatory movements other factors also influence the value of the final duration of a sound (accent, syllabic stress, vowel type, prevocalic and postvocalic consonants, within-word position, the preceding and following syllable and finally the utterance position [1]. The developed indirect method gives us the possibility to separate the effect of articulation from other factors and to define the specific, articulation governed sound durations. One of the most important advantage of this low level definition is that during
synthesis a well predicted basic sound duration value can be assigned to every sound of the utterance taking the effect of adjacent sounds into account. This value is modified in level 2 and 3 to get the final duration of the given sound.

The method
The indirect procedure contains two elements: synthesised speech without prosody and a perceptual evaluation process (Figure 1). The process is iterative. During the definition of the specific durations the test persons had to listen to synthesised texts (prepared for the experiment) and had to evaluate the duration of each sound in the given utterance. They were asked to mark the too long and too short sounds in the printed form of the same text. A phonetician re-evaluated the decisions and made the duration modifications in the speech unit database. This procedure finally resulted in a speech unit database that represented the specific durations. The whole procedure lasted for 8 months in 1999.

Elements for segmental level speech generation
The synthesised speech was produced by the concatenation of speech units, prepared carefully for the experiment and stored in the unit database. The following units have been used: triphones (CVC) and diphones (CV, VC, CC and VV). Other elements (CCC, VVV etc.) were generated using the former elements.

Earlier perceptual experiments showed that listeners are more sensitive to duration failures in vowels than in consonants [2]. Therefore in this procedure we treated vowel duration (especially in CVC combinations) as the most important among duration definitions. This decision resulted in the use of CVC triphones also in the speech unit database. So the duration of each vowel in CVC combinations could be handled individually, influenced only by the actual surrounding consonants.

The perceptual test. Four test subjects of normal hearing (one female and three male, ages between 30 and 50) completed the listening test, arranged always for one test person at a time. One listening session lasted for at most 30 minutes. Each person has listened to altogether 1200 sentences (basic test) and random texts (general test) during the evaluation process.

The steps of the test were as follows:
1. The test person was asked to listen to the synthesised text sentence by sentence and evaluate the duration of the sounds of the given sentence by marking in the printed text those sounds -with the predefined marker- the duration of which was heard too long (–) or too short (*). The sentence could be marked in the printed form of the same text. A phonetician re-evaluated the decisions and made the duration modifications in the speech unit database. This procedure finally resulted in a speech unit database that represented the specific durations. The steps of the test were as follows:
2. After all four test subjects had listened to the same material, a phonetician supervised the evaluation results (markers). In case of 3 or 4 corresponding opinions for the same sound he accepted the opinion and made the necessary change. In case of only 2 corresponding opinions he did not make any correction. A special, sound duration modifier program helped the phonetician to make the corrections.
3. Going ahead in tests more and more sounds had got their correct duration values characteristic for continuous speech and test subjects could mark the mistakes in durations more and more precisely. For the end of the evaluation process the sensitivity of the listeners reached the 10 ms value (see also in [3]). Going through the basic text material, all sounds in all combinations have been listened to and evaluated at least once by every test subject. (The text was redundant, i.e., many sound combinations occurred several times.)
4. Finally ordinary texts (from newspapers, articles, weather forecast, etc.) were synthesised by the system and sound duration values were tested the same way. Theoretically a statistical evaluation was thus performed as well i.e. the most frequent sounds in the most frequent sound combinations were tested once again.
5. After the evaluation procedure the segmental level speech, produced by the final speech unit database was very balanced from the point of view of correct sound duration values, the pronunciation was fluent, clearly understandable (without prosody). For example the sounds of the short sentence Jó napot kívánok. ‘Good morning.’ have the following specific durations in ms.

\[
\begin{align*}
\text{\textit{Jó}} & : [j]191 \mid [o:\](142) \mid [n][99] \mid [o]\mid (98) \mid [t][112] \mid [k][78] \\
\text{\textit{napot}} & : [n][135] \mid [v][42] \mid [k][175] \mid [n][41] \mid [o][88] \mid [k][98]
\end{align*}
\]

This speech unit database was declared to be the holder of specific sound durations. About 18000 specific duration values have been defined for the sounds in the function of adjacent sounds. These duration values are valid for a 13 sounds/s rate. The defined values have been examined in a further step, and compared to earlier results. It was found that the specific durations give close to the same distributions which had been measured with earlier direct methods [4], so this indirect procedure can be nominated as an acceptable method for sound duration definition. On the basis of these results it is claimed that specific durations are characteristic of Hungarian speech production and can serve as a stable basis for further calculation of final durations on the suprasegmental level.

2.2 Duration modifications on suprasegmental level
Based on our research it was assumed that on suprasegmental level the generation of sound duration can be divided into word (level 2) and sentence (level 3) parts. We found that word level modifications are more important than those of sentence level ones. After word level modifications the final duration structure reaches the stage of 90% of the final, desired one (measured on 44 test sentences). Another conclusion was that accents do not influence the duration map of the utterance, i.e. no lengthening can be shown in most of the cases in accented vowels (accent is on the first syllable of the word in Hungarian). Strong accents (e.g. focus) are exceptions.

Sound map function for words
In the second step of duration definition the sound map function of the word is used. This function modifies (with the
The specific duration of the given sound, by taking into account the length of the word, the short and long sounds, the type, and the order of sounds within the word. Multiplication factors have been defined between 0.8 and 1.4. Twenty-five basic rules have been defined for short vowels. Rule examples are shown for the first short vowel of the word in Table 1. Other rules modify the duration of short vowels inside the word. Separate rules (altogether 48) define the modification factors for long vowels. An example rule set is shown in Table 2 for the sound [a:]. Here separate rules define the modification as a function of the number of syllables in the word.

**Table 1:** Modifying multiplication factors for short vowels in the first syllable

<table>
<thead>
<tr>
<th>Sound Sequence</th>
<th>[i]</th>
<th>[u]</th>
<th>[y]</th>
<th>[o]</th>
<th>[e]</th>
<th>[æ]</th>
</tr>
</thead>
<tbody>
<tr>
<td># C V C 1</td>
<td>1</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td># C V C 1 C</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td># C V C 2</td>
<td>1</td>
<td>0.8</td>
<td>0.8</td>
<td>0.9</td>
<td>1</td>
<td>0.8</td>
</tr>
<tr>
<td># C V C 1 C</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>1</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td># C V C 1 C</td>
<td>1</td>
<td>1</td>
<td>0.8</td>
<td>1</td>
<td>1</td>
<td>0.9</td>
</tr>
<tr>
<td># C V C 2</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>1</td>
<td>0.9</td>
<td>1</td>
</tr>
<tr>
<td># C V C 1 C</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>1</td>
<td>0.9</td>
<td>1</td>
</tr>
<tr>
<td># C V C 1 C</td>
<td>1</td>
<td>1</td>
<td>0.8</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td># C V C 1 C</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>1</td>
<td>0.9</td>
<td>1</td>
</tr>
<tr>
<td># C V C 2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td># C V C 2</td>
<td>1</td>
<td>0.8</td>
<td>0.8</td>
<td>1</td>
<td>1</td>
<td>1.3</td>
</tr>
</tbody>
</table>

V = the short vowel in question, C = any consonant, C1 = any consonant except [r], C2 = [r], u = article, # = absolute beginning position multiplication factor = for example, 0.8

**Table 2:** Modifying factors for [a:] if it is the only long vowel in the word (for 1.2, 3, 4, 5 and 6 syllable words)

<table>
<thead>
<tr>
<th>[a:] in the syllable</th>
<th>Seq. u</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st syll.</td>
<td>VC1</td>
<td>1</td>
<td>0.9</td>
<td>0.8</td>
<td>0.8</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>VC2</td>
<td>1.3</td>
<td>1.2</td>
<td>1.2</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>2nd syll.</td>
<td>VC1</td>
<td>0.9</td>
<td>0.8</td>
<td>0.8</td>
<td>0.85</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>VC2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3rd syll.</td>
<td>VC1</td>
<td>0.9</td>
<td>0.8</td>
<td>0.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>VC2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>VC1</td>
<td>0.8</td>
<td>0.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>VC2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Last</td>
<td>VC1</td>
<td>1.2</td>
<td>0.85</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>VC2</td>
<td>1.3</td>
<td>1.2</td>
<td>1.2</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
</tbody>
</table>

V = sound [a:], C1 = any consonant except [r], C2 = [r, I]

The specific duration of consonants is modified by 5 rules. The number of both rule groups show that in continuous speech the duration of vowels varies more dynamically than that of the consonants. At the end of this step every sound of the word gets a multiplication factor. For example in the word láthatatlan 'invisble' the modifying factors will be:

<table>
<thead>
<tr>
<th>[l]</th>
<th>[a:]</th>
<th>[o]</th>
<th>[h]</th>
<th>[a:]</th>
<th>[i]</th>
<th>[e]</th>
<th>[t]</th>
<th>[a:]</th>
<th>[o]</th>
<th>[i]</th>
<th>[a:]</th>
<th>[i]</th>
<th>[n]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td></td>
</tr>
</tbody>
</table>

Comparative measurements have been done between natural and synthesised durations at this level. It was found that 90% of the durations was very close to the natural one. Figure 2 shows the duration map of the first part of a declarative sentence. Duration curve of synthesised sequence show the same tendency as the natural one. This result shows that sound durations in continuous speech are defined mostly by the specific durations and their modification on word level (step 1. and 2. in the model).

**Figure 2:** Sound duration values in the same natural and synthesised part of a sentence: ‘After the planned discussion…’ The horizontal axis contains the sounds. The articulation rate was 14 sounds/s in natural speech and 13 sounds/s in the synthesised one.

### Sentence level duration modification rules.

In the third step of duration definition only slight modifications are performed, mainly concerned lengthening in the last word of the sentence, in the last syllable of the word in phrase boundaries and in questions in the syllable where the pitch peak is realised.

### 3. The model of F0 structure

F0 modelling begins on the highest (suprasegmental) level and lower level structures are superimposed gradually, by finishing the procedure on the lowest (segmental) level. The solution is similar to Fujisaki's model and to the Bell Labs superpositional approach described by van Santen.

**Step 1.** The basic melody form on sentence level is modelled. The definition factors are: the type of the sentence, the number of phrases, the position of phrases in the sentence.

**Step 2.** Superimposing the word and syllable level fundamental frequency changes (the influencing factors are: sentence type, accent distribution).

**Step 3.** Segmental level. Micro intonation is superimposed on the fundamental frequency curve, (the type of sounds and the previously determined fundamental frequency changes determine the presence of micro intonation).

#### 3.1 Building units for sentence level melody

Three types of general building units have been defined: rising, level and falling, on the basis of analysis of 452 sentences (statements, questions in dialogues). The basic melody forms for statements, questions, commands and warnings have also been defined [7]. For the general representation of different melody patterns in different types of sentences a unified description form has been constructed in which the frequency data was expressed in %, relative to a predefinded pitch value (the start pitch in a synthesiser) [7]. This form allowed us on one hand to describe -in the form of general rules- the relationship among the melody patterns of different types of sentences (e.g. in dialogues). On the other hand transformation of the whole melody curve on higher or lower carrier frequencies was also possible. Every building unit had different fixed starting and end points (e.g. rising: from 85% to 95%, or from 75% to 100%). The length of the units was an open variable, defined by the length of the phrase in which it was used. A total of 13 falling, 7 level and 12 rising melody units were defined as building elements. During modelling of the melody curve these units are concatenated. The concatenation points can be defined by statistical empirical rules, by specialised syntactic analysis [9], by phonological rules etc. The predefinition of units can be given in the text with the combination of back slash(es) and a unit number (e.g. //12).

#### 3.2 Word and syllable level F0 structure

In the second step the word and syllable level F0 structures were defined (marked by intonation markers, e.g. [N]). These
structures are applied dynamically on the sentence level units, i.e. the actual change in pitch is defined by the carrier sentence level unit [7]. For word level two types of structures have been defined: neutral word [N], in which no further pitch change is performed, and word with negative accent [-], which means that the pitch is set on lower frequency as it was in the sentence level unit. For syllable level changes three basic types of structures were defined: rising, falling and rising-falling, based on the analysis of the former 44 sentences. These structures were marked with [Ws] in the text, where s is the number of the rule. For Hungarian 12 syllable level rules were formed from the three basic structures. These rules can be applied for the generation of accents, and other syllable level changes (like in questions, in commands etc.). The syllable level rules define the pitch variation within one but maximum two syllables (e.g. rising and falling in the same syllable or rising in the first and falling in the second one). In our model every word has to get its intonation marker. The definition of the F0 structure of the sample sentence 'After the planned discussion I will write a letter to the foreign partner.' looks like:

```plaintext
"W1"tervezet [W1]térülési /44[N]utan /21 [W1]level
[N]irok /44]-ja [W1]kiővés [N]parnernek The // marker means the place of breaks, / indicates a new phrase without break. Figure 3 shows the synthesised F0 structure of the sentence (upper) and that of the natural one (lower).
```

![Figure 3: Comparison of the F0 structures of the synthesised sentence (composed according to the model) and that of the natural one](image)

4. The intensity structure model

Modeling begins on the lowest (segmental) level and is finished on the highest (suprasegmental) one.

**Level 1.** Corrections are defined on sound level (within words) taking into consideration the sonority order of vowels (the type of the vowel and the position of it within the word determines the rules)

**Level 2.** Corrections on word level (the type of the word, the length of the word and the position of it in the sentence determines the correction).

**Level 3.** Corrections on sentence level. (type of the previous sentence, type of the actual sentence, the complexity of the sentence determines the correction).

4.1 Sound level intensity rules

In natural speech the modification of the sound intensity levels -defined by the sonority order of vowels- is automatically done during the construction of the intensity structure of a word. In Hungarian the default intensity structure in words is mainly falling one. This structure has to be generated during synthesis as well. If the sonority order of the vowels in the word meets the general sonority order of vowels the falling intensity structure will be generated automatically (e.g. in bácsi ‘old man’). If not, correction rule have to be applied to rise the intensity of the vowel with less sonority [i,o,u] and to reduce the intensity of the vowel having strong sonority [a, e] (e.g. kiált ‘shouts’).

4.2 Word and syllable level intensity rules

On word level two types of rules were defined, the first concerns the whole word and the second a syllable size unit. Words with negative accent (articles etc.) and also the last word of the sentence get less intensity. On syllable level the impression of accent can be generated by increasing the intensity (there are cases in which the accent can not be realised by increasing F0, because phrase or sentence level melody structures are stronger than the accent and override it). In questions the syllable, carrying the F0 peak may also get higher intensity.

4.3 Sentence level intensity rules

On this level the final intensity structure of the sentence is determined. The function takes into consideration the previous sentence, the type of the actual sentence, and the complexity of the actual sentence. The intensity structure of different sentence types was determined and the rules adjust main characteristics.

5. Conclusion

The presented model describes the three basic prosody elements on sentence and on text level. The introduced indirect method for the definition of specific durations gives the basis for further duration rules. The advantages of this method are: person independent, the data are reproducible, all durations are given for continuous read speech, statistical measurements can be performed, the results can be listened to, good for rule definition. Using the sound map function the duration of the sounds of any word can be predicted with high accuracy. The model for F0 and intensity structure describes the most frequent forms that occur in continuous read speech.

The model was successfully tested in the context of generating Hungarian synthetic speech. It was used in a study to determine the phonetic representation of phonological intonation rules, too. Further research is planned to verify it in other languages and other application domains (e.g. recognition).

6. References