Acta Linguistica Hungarica, Vol.49 (3-4), 2002. 321-345.

Model to predict Hungarian sound durations for continuous speech

Olaszy Gábor
Kempelen Farkas Speech Research Laboratory, Research Institute for Linguistics, Hungarian Academy of Sciences
H-1068 Budapest, Benczúr u. 33.
Hungary

Abstract

Direct measurements show that many factors influence the final value of sound durations in continuous speech. On segmental level mainly the articulatory movements determine important influence factors, on suprasegmental level the accent, syllabic stress, within-word position, the preceding and following syllable and finally the utterance position may have influence on final sound duration. So the problem how to predict the sound duration can be described with a multivariable function in which the effect of the variables can not be easily defined with good accuracy. It is difficult to select the effect of certain functions, i.e. it is difficult to model this function, making direct measurements on speech signal.

A model have been constructed and realized in which three well defined levels are separately working. In the first one (this is the segmental level) the separation of the effect of articulation from other factors is solved. On the second and on the third levels relate to the suprasegmental level of speech.

Introduction

During speech production the articulatory movements form the frequency and time structure of the speech signal. It is also well known that articulation has an influence on sound duration. Different methods may be used to describe this effect. The use of an articulatory model is described by Shiga et al. (1998) where four time-variable articulatory parameters represent the conditions of articulatory organs whose physical restrictions seem to significantly influence segmental duration.

Measurements showed that beside the effect of articulatory movements other factors also influence the value of the duration of a sound. Van Santen (1992) points out that at least eight factors matter in this process: accent, syllabic stress, vowel type, prevocalic and postvocalic consonants, within-word position, the preceding and following syllable and finally the utterance position.

Earlier measurements of sound durations in Hungarian concern both the inherent time structure of sounds (transient phases, structure of consonants, VOT, etc.) and also the overall duration of the sound. The latter has been examined by Magdics (1966), Kassai (1979) and latestly by Kovács (2002). The inherent time structure of every Hungarian sound has been examined by Olaszy (1991). The first synthesis controlled measurements for the examination of the structure and duration of Hungarian consonants were done by Olaszy (1985). All four authors gave the results mainly in the form of mean values and main tendencies. These data are somewhat different from what is required for the construction of a duration model. For example text-to-speech (TTS) conversion requires an adequate duration model for the given language. The construction of that is complicated by the multitude of phenomena which affect durations in speech (O'Shaughnessy 1981). For this reason researchers try to separate certain factors during the investigations and try to define controlled environments (limited number of words, using nonsense items, placing words or syllables in frame sentences etc.) in which only one changing factor is present at a time. For example, in a study of French vowel and consonant durations, O'Shaughnessy (1981) limited the investigation to stressed syllables in words. Van Santen (1992) used well created sentence pairs for the investigation of contextual effects on English vowel durations.

The model proposed in this paper gives the possibility to separate the different effects that influence the creation of the sound duration. First the influence of articulation is taken into consideration (segmental level of speech), secondly the influence of othe factors is discussed. The results of the segmental level part are expressed by the specific, articulation governed sound durations (duration of every sound in the function of adjacent sounds for continuous speech). These specific duration values are used as a basis in the further (word and sentence level) calculations. Thus by the model the prediction of speech sound durations can be performed for the sounds of any text without direct measurements.

State of the art

Modeling of sound durations became important by the growing development of speech technology (text-to-speech conversion, speech recognition) in the last decades. During the latest decades two main approaches have been borne: rule governed and statistical systems. In rule governed approaches the researchers try to characterize the whole complex process with rules (basically on the linguistic level). The sound duration here is characterised by an intrinsic value. In the calculation of the final duration various phenomena (mainly defined from syntactic information) are taken into consideration and applied on the intrinsic durations.

The statistical approach uses the results of statistical measurements to predict sound duration.

It is difficult to separate definitely the rule based and the statistical approach. For example the MITalk TTS system (Allen et al. 1987) is regarded by Zellner (1994) a statistical system, while van Santen (1998) mentions it as a purely rule based solution. The MITalk system seem to involve both, because this model is built around average duration, i.e., durations for individual phonemes which represent the result of statistical measurements. The final duration will be calculated after taking into consideration the position within a paragraph, the semantic novelty, the phrase structure etc.

In a newer approach Campbell (1992) proposed another type determination of sound durations. According to this, first the higher level syllable duration have to be calculated to reflect the rhythmic and structural organisation of the utterance and the durations of the sounds in the syllable are calculated from the syllable duration.

One common feature of all these approaches is that the duration data and rules are derived from natural speech material. The disadvantage of these methods is that the measured duration values contain the effect of more than one feature in many cases. Moreover, the generality of the results may be restricted by the influence of individual pronunciation (Van Santen 1992).

Hypothesis

The hypothesis was that the surface level final durations can be built up from low level basic structures. The concept follows the theoretical separation of speech into segmental and suprasegmental levels. The segmental level durations represent the basis (speech without prosody but having the correct specific duration values of the sounds, the distribution of durations, the correct, language specific timing ratio among speech sounds). At this level only the articulation has effect on sound durations. We assume that data on this level can give the basis for the further calculations (modifications of the specific durations) which are determined on suprasegmental (surface) level.

The method used

In this paper we describe an inverse (bottom-up) method to define the final, surface level sound durations. Sound durations are determined in three steps in this model.

(1) The most important part of the whole procedure is the indirect measuring method that is applied to determine the specific durations: (t)spec**.** Their value varies only in the function of articulation. The indirect measuring method means that the duration values are not defined by measuring the sound durations in natural speech, but by using the combination of segmental level speech synthesis and perceptual evaluations. Thus the specific sound durations

characteristic for continuous speech (taking into consideration the effects of the continuous serial articulation process) will be determined in milliseconds (for a certain articulation rate).

(2) The second step is based on the results of step (1) and the modification factors defined are derived from the words as building units of speech. Word level modification rules have been formed which showed to what extent the specific duration of the sound have to be lengthened or shortened within the word (in continuous speech). The result of this step is a modification factor (M1) for every sound of the word. M1 is defined by the following variables: the length of the word and the sound map of the word (which sounds and sound combinations are in the word, and what is the sound order). All sounds of the word are supplied with M1. The series of these numbers is called: **word level duration map**.

(3) The third (suprasegmental) level of the model represents the final adjustment of the sound duration. The second modification factor (M2) is defined by **sentence level** rules (modality, phrase structure, prominence etc.).

The final sound durations (individually for every sound of the utterance in the function of the adjacent sounds) are than calculated by the following way:

(t) final= (t)spec x M1 x M2

As a result of the three steps the final sound durations of every sound in the utterance will be defined.

The experimental setup for getting specific durations was organized around a **segmental level** TTS synthesizer, a perceptual evaluation procedure and a sound duration modifier (Figure 1).

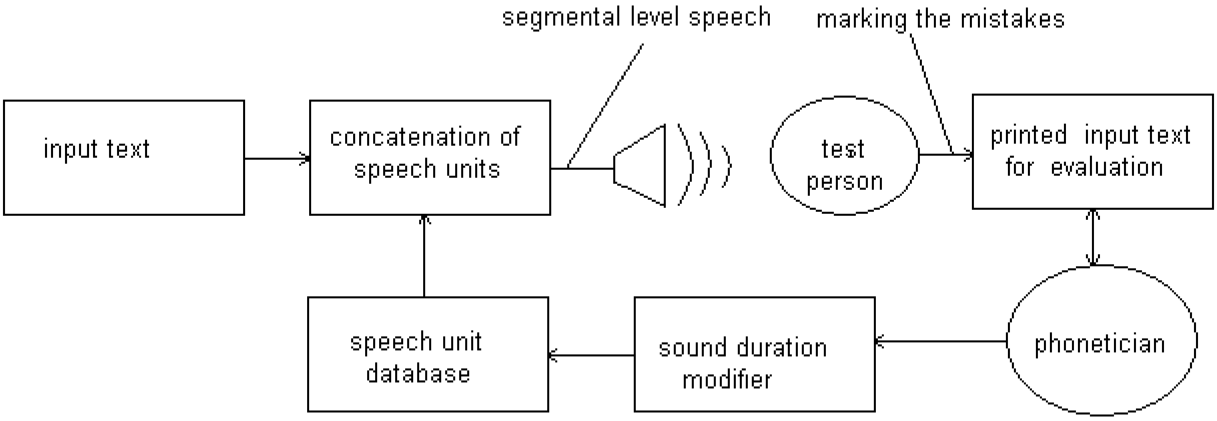

Figure 1. The test environment for the adjustment of specific sound durations

The TTS synthesizer consists of a speech unit database (waveform elements derived from human pronunciation) a concatenation module, a grapheme-sound converter and a sound duration modifier. This synthesizer produced the speech (withouth melody and accent) for the perceptual evaluation. The design and realization of this synthesizer was one of the most complicated elements by setting up the test environment. The steps of realization were as follows: a) determination of the speech sound set for the TTS conversion; b) definition and of the form of the elements of the speech unit database for concatenation; c) designing the text corpus for the creation of the elements of the speech unit database; d) the realization of the speech unit database.

Speech sound set and its representation in the experiment

The goal of the experiment was to measure the duration of the 9 basic vowels (7 short ones plus the long [a:] and [e:]) and the 23 short consonants of Hungarian (Table 1, Table 2). The symbols of the third rows of the tables represent the appropriate character for the given sound in the representation of computer programs. These characters will be used in computer generated tables and figures. In the characters of the third rows will be written between brackets like (a), (A:), (u), (U) when referring to a Hungarian speech sound. The phonetic symbols of sounds will be written as: [▱], [a:], [o] etc.

Table 1. The basic Hungarian vowels used in the experiment

IPA symbol	a:	□	o	U	y	i	ε:	O	E
written form	á	a	o	U	ü	i	é	ö	e
symbol in this experiment	A	a	o	U	U	i	E	O	e

Table 2. The basic Hungarian consonants used in the experiment

IPA symbol	b	p	d	t	γ	k	ǀ	χ	m	n	)	j	h	v	f	z	s	ts	Z	Σ	tΣ	l	r
written form	b	p	d	t	g	k	G y	ty	m	n	ny	j	h	v	f	z	sz	c	zs	s	cs	l	r
symbol in this experiment	b	p	d	t	g	k	G	T	m	n	N	j	h	v	f	z	s	c	Z	S	C	l	r

Definition and realisation of the speech unit inventory for synthesis

The goal by the definition of the form of speech units was to produce good quality (close to natural voice timbre) speech by the synthesizer. Thus the sound quality will less influence the listeners by their duration evaluation. As the basic goal was to define the duration of a sound taking the effect of the adjacent sounds into account, theoretically C**V**C, V**C**V, C**C**V, V**C**C, V**V**C and C**V**V elements could have been used as building units. As perceptual experiments showed that listeners are more sensitive of duration failures in vowels than in consonants (Kato et al. 1998), we treated the vowel duration as the most important kind of data, especially in C**V**C combinations. Our latest measurements showed, that this combination type occurs most frequently in Hungarian (80% of the triphone units are CVC structures, measured in the corpus of 2 million different word forms). The duration of the vowel can be determined the most correctly if the vowel is treated during the synthesis as an individual element influenced only by the actual surrounding consonants, i.e. every vowel in every C**V**C combination has its own specific duration and this duration value represents the duration of the vowel only in the given C**V**C combination. If we take into account the fact that the given vowel may be preceded by any consonant and may be followed also by any one, theoretically the effect of articulation of adjacent consonants on the duration of any vowel can be defined by four cases as indicated in Table 3.

Table 3. The theoretical effect of consonants on vowel duration in CVC sequences

preceding consonant		following consonant		the final duration of the vowel
lengthening the vowel	shortening the vowel	lengthening the vowel	shortening the vowel	
+		+		lengthened (doubled lengthening influence)
+			–	equalized (not changed)
	–		–	shortened (doubled shortening influence)
	–	+		equalized (not changed)

The final decision was to use triphone C**V**C elements in the speech unit inventory to ensure the possibility of most precise adjustment for vowels in C**V**C combinations during the perceptual evaluations. This fact defined the final content of the speech unit database: **vowels in CVC combinations were generated form CVC triphones, all other sound combinations were generated using the concatenation of CV, VC, VV, and CC diphones**. One triphone element contained two half consonants and the vowel between them. One diphone element contained two half speech sounds (e.g. a CV unit has the second part of the C and the first half of the V). The speech unit database was planned to have 4761 C**V**C triphones, 207 CV, 207 VC, 81 VV and 529 CC diphones.

The criteria for the creation of the speech unit database

The elements of the speech unit database were created from human voice items. A text corpora had had designed which was read by a male announcer. Three aims were kept in mind when designing the text corpora: (1) to keep the correct formant structure in vowels (mainly in CV, VC and VV diphones; (2) to reduce the effect of suprasegmental factors (accent, rhythm, melody etc.); and (3) to have controlling possibilities for keeping sound intensity close to a constant value during the recording.

To meet these requirements three-syllable meaningless text items were defined for the announcer.

An example of the meaningless text items containing (o) vowel for CVC triphone units, where the vowel was preceded by the consonant [b] and followed by all consonants looked like: a**bob**a [□**bob**□], a**bop**a [□**bop**□], a**bod**a[□**bod**□], a**bot**a [□**bot**□], a**bog**a [□**bog**□], a**bok**a [□**bok**□], ..., a**bom**a [□**bom**□], a**bon**a [□**bon**□], etc..

The text items for the production of the CV, VC and VV diphones were designed using a well known phonetic rule. The problem in diphone representation is that vowels are cut at their middle point. When generating a vowel in the synthesis with the concatenation of two diphones, spectral discontinuities may occur in the formant structure of the vowel at the point of concatenation. This produces distortion. To reduce these distortions the formants of vowels were governed by phonetic means to reach an optimal steady state position at the concatenation point for both in CV and VC diphones. The [k] sound was used for this purpose because this sound is the most flexible as to its articulation and it does not influence the formant structure either of the preceding (in VC combination) or of the following (CV) vowel very much. In items for CV diphones the [k] sound follows the vowel, i.e., the formants of the vowel will be close to the steady state values at the second half of the vowel where the cut will be done. Examples of the meaningless text items containing CV diphone elements are: a**bo**ka [□**bo**k□], a**po**ka [□**po**k□], a**do**ka [□**do**k□], a**to**ka [□**to**k□].....; and for VC elements: ak**ob**a [□k**ob**□], ak**op**a [□k**op**□], ak**od**a [□k**od**□], ak**ot**a [□k**ot**□]. In items for VC diphones the [k] sound precedes the vowel, i.e., the formants of the vowel will be close to the steady state values in the first half of the vowel where the cut will be done. Thus it can be assumed that the discontinuity in formants will be low and by concatenating these diphones, the formant frequencies at the concatenation point will be close to each other, therefore spectral distortion will be minimal.

For the production of CC diphones, words containing the given element were mostly given in the text list.

The structure of the text corpora described above solved other two problems too: not to have accent on the triphone and on the diphone element (in Hungarian the accent is on the first syllable of the pronounced word). With the use of [□] in the first and in the last syllable the sound intensity level (the demand was to keep it constant as far as it can be during the recording) became controllable.

Realization of the speech unit database

The text material was read by a trained male speaker in a monotonous style (keeping the fundamental frequency as constant as it was possible) but with normal speech rate. The digital representation (22kHz, 16 bit) of the wave form was labeled on sound boundaries (semi automatically) and pitch synchronous markers were placed too (semi automatically). It is obvious that the correctness of any sound duration measurement strongly depends on the definition of sound boundaries in the measurable waveform. In our case a phonetician labeled the sound boundaries manually (with visual and auditive control). Visual observations concerned the waveform of the signal and the intensity curve of it. In some special cases a spectrographic analysis was also used to define the sound boundary. The flexible "play the sound window" made the auditive control more effective, i.e. the acoustic change in the sound could be heard by adding step by step one more period to the previously selected and played part of the window. All these supports were given by the by the Hungarian Profivox Development System (PDS) software tool (Olaszy et al. 2002). For vowels in CVC combinations the onset and offset were determined mostly very correct (consonantal aspiration was not involved). In VV combinations, the auditive examination gave the most important help to determine the boundary. In the case of sonorant-vowel combinations, the analysis of the intensity curve and the auditive examination gave the result.

The speech unit database was created by a semi automatic method. The cut points for CVC elements were defined at the middle of the consonants, and for diphone elements at the middle of the sounds. This database contained individual vowel durations for every CVC combination type and created durations for all other sounds in all combinations. Created duration means that the duration of the sound will be defined by the two diphones used actually.

The perceptual evaluation

The determination of specific durations was carried out by a multi step, long lasting perceptual evaluation (Figure 1). It represented a closed circuit sound duration evaluation and correction procedure. The TTS produced the voice (without suprasegmental structure) from the input text. Two types of input text corpora were used a basic and a general text material. The basic one consisted of 1200 sentences, (5-10 words in a sentence). The general one contained texts from newspapers, books and scientific articles. The printed form of all these text materials served for marking the results of the duration evaluation.

Four test subjects of normal hearing (one female and three male, ages between 30 and 50) completed the whole test. The whole perceptual evaluation and duration correction procedure lasted for eight months. The listening was arranged always for one test subject at a time. One listening session lasted for max. 30 minutes, and about 50 sentences were evaluated. The articulation speed of the synthetic speech was 12-13 sounds/s, this referred to a medium speaking rate in Hungarian (Kovács 2002).

The steps of the perceptual evaluation were as follow:

1. The test subject was asked to listen to the synthesised text sentence by sentence. He/she had to evaluate the duration of the sounds of the given sentence, and to mark with the predefined marker on the printed text those sounds the duration of which was heard to be too long (–) or too short (*). Using a repeat function the previous sentence could be listened to several times if required. An evaluated sentence showed for example the following picture:

A tervezett   tárgyalás után   levelet   írok a   külföldi   partnernek.   *(1)*
*   _   _   *   _   _   *   _   *   _

('After the planned discussion I will write a letter to the foreign partner.')

The markers in the example show that there was one too short part at the beginning of the first word, one longer vowel was found in the second word, and so on.

2. A phonetician took part in the test too. He controlled the marked judgments of the test persons. In case of 3 or 4 same opinions for the same sound he accepted the opinion and made the lengthening or shortening according to his own decision and perceptual judgment. In case of only 2 corresponding opinions he did not make any correction. The duration change was set in the given part of the triphone or diphone in question. Thus the speech unit database contained more and more close correct durations. After making all corrections the listeners were asked (2-3 weeks later) to make the evaluation (point 1 and two) once more for the whole text. A special, sound duration modifier program (Olaszy - Olaszi 1998) helped the phonetician to make the corrections.

Going ahead in the evaluation procedure more and more sounds reached their correct, segmental level, specific duration characteristic for continuous speech. Test persons could mark the mistakes in durations more and more precisely. Already the experiments of Huggins (1972) had shown that listeners can perceive very small changes in duration. In this experiment the sensitivity of the listeners reached the 10 ms value in the final phase. The test procedure was done all together four times with the four test persons.

3. After this phase ordinary texts (from newspapers, articles, weather forecast, etc.) were synthesised by the system (without prosody parameters) and sound duration values were tested the same way as in point 1 and 2. Such texts automatically contain the language

specific occurrence of segmental units. So, the duration of the most frequent sounds in the most frequent sound combinations was evaluated and corrected (if needed) once again.

4. After the whole 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 in continuous speech. The produced synthetic speech (without prosody) was fluent, and clearly understandable. **This database was then declared to be the reference database that incorporates the specific sound duration values (for all sound combinations) involving the influence of articulation on duration. These duration values are characteristic of Hungarian speech production and can serve as a stable basis for further calculation of final durations on the suprasegmental level.**

Results and criticism

The goal of the whole procedure was not only to determine the segmental level sound durations, but also proofing the correctness of this new indirect procedure and the results obtained. Therefore besides the definition of specific sound durations distribution measurements have been performed to study the data, produced by the first level of the model. The aim of these distribution measurements was to get an overview (on data level) about the behavior of specific sound durations in different sound combinations. The data have been compared with earlier results (derived from direct duration measurements by Kassai (1979) and Magdics (1967)). It was assumed, if these new results correlate with earlier results the method presented can be accepted as an objective procedure for the definition of segmental level, specific, articulation governed sound duration structure of a language.

Vowels in CVC combinations

The results contain duration values for nine vowels in 4761 different combinations. The data are presented in the form of matrices for every vowel. A sample matrix for the sound (o) is given in Table 4. The table shows the specific duration values of **(o)** in all CVC combinations. The leftmost column of the matrix represents the preceding C, the top row the following C. The target vowel **(o)** is shown at the upper left corner of the matrix. So if we want to get the specific duration of **(o)** in the sequence *boldog* [boldog] ‘happy’ we take the row of **(b)** and the column of **(l)**. The result is 84 ms for the given articulation rate. For the second **(o)** we take the row of **(d)** and the column of **(g)**. The result is 91 ms.

Table 4. The specific durations of (o) in C**V**C combinations in ms in continuous speech

o	**b**	**p**	**d**	**t**	**g**	**k**	**G**	**T**	**m**	**n**	**N**	**j**	**h**	**V**	**f**	**z**	**s**	**c**	**Z**	**S**	**C**	**l**	**r**
b	88	93	84	95	93	90	93	103	83	84	94	95	95	85	94	93	90	94	94	85	83	84	94
p	88	93	83	95	92	90	92	103	82	83	93	95	95	84	93	93	90	94	94	84	83	83	93
d	86	91	82	93	91	88	91	101	81	81	92	93	93	83	91	91	88	92	92	83	81	82	92
t	84	90	80	92	89	86	89	100	79	80	90	92	92	81	90	90	87	91	91	81	80	80	90
g	87	92	83	94	92	89	92	102	82	82	93	94	94	84	92	92	89	93	93	84	82	83	93
k	79	84	75	86	84	81	84	94	74	74	85	86	86	76	84	84	81	85	85	76	74	75	85
G	90	95	85	97	94	92	94	105	84	85	95	97	97	86	95	95	92	96	96	86	85	85	95
T	99	104	95	106	104	101	104	115	94	95	105	106	106	96	105	105	101	106	106	96	95	95	105
m	79	85	75	87	84	81	84	95	74	75	85	86	86	76	85	85	81	86	86	76	75	75	85
n	90	96	86	98	95	92	95	106	85	86	96	98	98	87	96	96	93	97	97	87	86	86	96
N	94	99	90	101	99	96	99	109	89	90	100	101	101	91	100	99	96	100	100	91	89	90	100
j	80	86	76	88	85	82	85	96	75	76	86	87	87	77	86	86	82	87	87	77	76	76	86
h	90	96	86	98	95	92	95	106	85	86	96	98	98	87	96	96	93	97	97	87	86	86	96
v	88	94	84	96	93	90	93	104	83	84	94	96	96	85	94	94	91	95	95	85	84	84	94
f	80	86	76	88	85	82	85	96	75	76	86	88	88	77	86	86	83	87	87	77	76	76	86
z	91	96	87	98	96	93	96	106	86	86	97	98	98	88	96	96	93	97	97	88	86	87	97
s	87	92	83	94	92	89	92	103	82	83	93	94	94	84	93	93	89	94	94	84	83	83	93
c	93	98	88	100	97	95	97	108	87	88	98	100	100	89	98	98	95	99	99	89	88	88	98
Z	87	92	82	94	91	89	91	102	81	82	92	94	94	83	92	92	89	93	93	83	82	82	92
S	77	83	73	85	82	79	82	93	72	73	83	85	85	74	83	83	80	84	84	74	73	73	83
C	88	94	84	96	93	90	93	104	83	84	94	95	95	85	94	94	90	95	95	85	84	84	94
l	80	85	76	87	85	82	85	95	75	75	86	87	87	77	85	85	82	86	86	77	75	76	86
r	89	95	85	97	94	91	94	105	84	85	95	97	97	86	95	95	92	96	96	86	85	85	95

The duration data in Table 4. contain the effect of articulation on the duration of (o) in CVC combinations. The mean duration calculated from these data for (o) is 90 ms. The minimal duration is 72 ms, the maximum is 115 ms. The distribution of duration values as a function of C**V**C combinations is shown in Table 5. The different duration values for (o) can be summarised into four 10ms groups i.e. C**V**C elements where the duration is between: 70 and 79 ms, 80-89, 90-99, 100-109 ms. The duration exceeds 110 ms only in the (ToT) combination. This distribution shows that the duration of (o) is the longest in the neighborhood of palatals and it is the shortest in the neighborhood of nasals and the (S).

The summerised mean specific duration values of the 7 short and two long Hungarian vowels are given in Table 6 and Figure 2. Vowel order data obtained with this inverse method correlate with earlier results of Kassai (1979) who gave the duration order of short vowels in accented position as: [i] < [u] < [y] < [o] < [E] < [□] < [O] (< sign means shorter than). The present data give the same vowel order.

Table 5. The distribution of specific durations of (o) in C**V**C combinations

70-ms	Tom	toC	kob	kod	kom	kon	kov	KoS	koC	kol	mob	mod	mom	mon	mov	moS	moC	mol	jod	jom	
	Jon	jov	joS	joC	jol	fod	fom	Fon	fov	foS	foC	fol	Sob	Sod	Sok	Som	Son	Sov	Sos	SoS	SoC
	Sol	lob	lod	lom	lon	lov	loS	loC	lol												
80-	Bob	bod	bok	bom	bon	bov	boS	boC	bol	pob	pod	pok	pom	pon	pov	pos	poS	poC	pol	dob	
	Dod	dok	dom	don	dov	dos	doS	doC	dol	tob	top	tod	tog	tok	toG	ton	tov	tof	toz	tos	ToS
	Tol	gob	god	gok	gom	gon	gov	gos	goS	goC	gol	kop	kot	kog	kok	koG	koN	koj	koh	kof	Koz
	Kos	koc	koZ	kor	Gob	God	Gom	Gon	Gov	GoS	GoC	Gol	mop	mot	mog	mok	moG	moN	moj	moh	Mof
	Moz	mos	moc	moZ	mor	nod	nom	non	nov	noS	noC	nol	Nod	Nom	Non	NoC	Nol	job	jop	jot	Jog
	Jok	joG	joN	joj	joh	jof	joz	jos	joc	joZ	jor	hod	hom	hon	hov	hoS	hoC	hol	vob	vod	vom
	Von	vov	voS	voC	vol	fob	fop	fot	fog	fok	foG	foN	Foj	foh	fof	foz	fos	foc	foZ	for	zod
	Zom	zon	zov	zoS	zoC	zol	sob	sod	sok	som	son	sov	sos	soS	soC	sol	cod	com	con	cov	coS
	CoC	col	Zob	Zod	Zok	Zom	Zon	Zov	Zos	ZoS	ZoC	Zol	Sop	Sot	Sog	SoG	SoN	Soj	Soh	Sof	Soz
	Soc	SoZ	Sor	Cob	Cod	Com	Con	Cov	CoS	CoC	Col	lop	lot	log	lok	loG	loN	loj	loh	lof	loz
	Los	loc	loZ	lor	rob	rod	rom	ron	rov	roS	roC	rol									
90-	Bop	bot	bog	boG	boN	boj	boh	bof	boz	bos	boc	boZ	bor	pop	pot	pog	poG	poN	poj	poh	
	Pof	poz	poc	poZ	por	dop	dot	dog	doG	doN	doj	doh	dof	doz	doc	doZ	dor	tot	toT	toN	toj
	Toh	toc	toZ	tor	gop	got	gog	goG	goN	goj	goh	gof	goz	goc	goZ	gor	koT	Gop	Got	Gog	Gok
	GoG	GoN	Goj	Goh	Gof	Goz	Gos	Goc	GoZ	Gor	Tob	Tod	Tom	Ton	Tov	ToS	ToC	Tol	moT	nob	nop
	Not	nog	nok	noG	noN	noj	noh	nof	noz	nos	noc	noZ	nor	Nob	Nop	Nog	Nok	NoG	NoN	Nov	Nof
	Noz	Nos	NoS	Nor	joT	hob	hop	hot	hog	hok	hoG	hoN	hoj	hoh	hof	hoz	hos	hoc	hoZ	hor	vop
	Vot	vog	vok	voG	voN	voj	voh	vof	voz	vos	voc	voZ	vor	foT	zob	zop	zot	zog	zok	zoG	zoN
	Zoj	zoh	zof	zoz	zos	zoc	zoZ	zor	sop	sot	sog	soG	soN	soj	soh	sof	soz	soc	soZ	sor	cob
	Cop	cot	cog	cok	coG	coN	coj	coh	cof	coz	cos	coc	coZ	cor	Zop	Zot	Zog	ZoG	ZoN	Zoj	Zoh
	Zof	Zoz	Zoc	ZoZ	Zor	SoT	Cop	Cot	Cog	Cok	CoG	CoN	Coj	Coh	Cof	Coz	Cos	Coc	CoZ	Cor	loT
	Rop	rot	rog	rok	roG	roN	roj	roh	rof	roz	ros	roc	roZ	ror							
100-	BoT	poT	doT	goT	GoT	Top	Tot	Tog	Tok	ToG	ToN	Toj	Toh	Tof	Toz	Tos	Toc	ToZ	Tor	noT	
	Not	NoT	Noj	Noh	Noc	NoZ	hoT	voT	zoT	soT	coT	ZoT	CoT	roT							
110-	ToT																				
1)																					

As we look the situation in other languages, similar results were reported by O'Shaughnessy (1981) for French vowels in closed syllables, where the shortest vowels were the high ones [i, u], the mid-vowel [e] was longer and the low vowel [a] was found to be the longest. Measured data for English (van Santen 1992) follow the same order both in stressed and in unstressed position. Thus the correlation between the duration and the height of the tongue during articulation is involved in our indirectly measured data as well.

For the examined two long vowels our results also correlate with those of Kassai, i.e. the sound [ɛ:] is shorter than [a:]. The distribution of short vowels ranges from 55 ms to195 ms according to Kassai, the present results are: 61-115 ms. The latter difference can be explained by the fact that Kassai measured the data from complex speech (with normal rhythm, accent etc.), but now we derived them from a segmental level signal where the distribution is obviously narrower.

Table 6. Specific duration values determined for Hungarian vowels in ms in continuous speech.

vowel	(i) [i]	(u) [u]	(U) [y]	(o) [o]	(a) [□]	(e) [E]	(O) [O]	(E) [e:]	(A) [a:]
Mean	80	86	86	90	91	91	92	146	164
Min.	61	69	61	72	73	64	71	124	128
Max.	99	113	103	115	113	115	109	170	196

The average duration of all vowels is 102 ms. For English Van Santen (1992) defines this value as: 106 ms. The average of all short vowels for Hungarian is 88 ms, while van Santen gives the average duration data for English /i/ and / ℘/ as: 80 and 88 ms, respectively. However at some points the present results do not correlate with Kassai’s measurements: we found that the duration of a vowel is not lengthened by the following (l), (r) sounds. Furthermore our data do not support the finding that the duration of the vowel is consequently longer before voiced consonants than before voiceless ones.

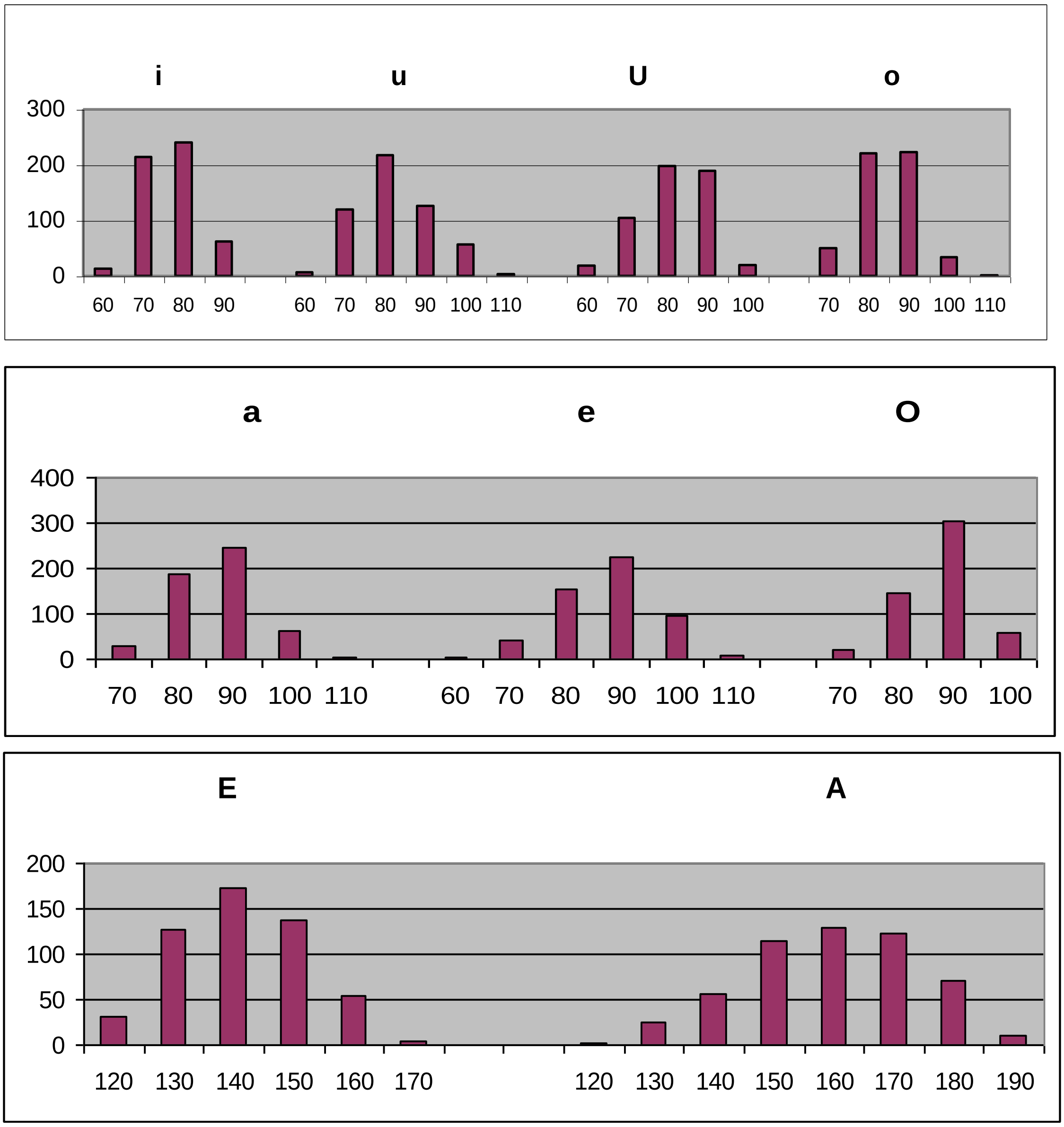

Figure 2. The distribution of specific durations of Hungarian vowels in C**V**C combinations in continuous speech. The horisontal axis shows the duration data (in ms) groups, the vertical axis shows the number of V**C**V items in which the given duration of the vowel occurs

Consonants in V**C**V combinations

For all consonants in all V**C**V combinations 1863 specific duration values were defined in 23 matrices. A sample matrix for the sound (b) is shown in Table 7 where the duration values of (b) are given in milliseconds in all V**C**V combinations. The leftmost column of the matrix represents the preceding V, the top row the following one. The target consonant (b) is shown at the upper left corner of the matrix.

Table 7. The specific durations for (b) in V**C**V combinations in ms in continuous speech

b	**A**	**a**	**o**	**u**	**U**	**i**	**E**	**O**	**e**
A	61	63	63	56	66	68	66	57	61
a	62	64	64	57	67	70	67	58	62
o	67	69	69	62	72	75	72	63	67
u	70	72	72	65	75	78	75	66	70
U	61	63	63	56	66	69	66	57	61
i	67	69	69	62	72	75	72	63	67
E	60	62	62	55	65	68	65	56	60
O	60	62	62	55	66	68	65	56	60
e	69	71	71	64	74	76	74	65	69

For example the specific duration of (b) in the sequence *abe* is shown at the cross-point of the row of **(a)** and the column of **(e)**. The result is 62 ms for the given articulation rate. The minimum duration for (b) is 55 ms, the maximum is 78 ms. The duration distribution for (b) can be arranged into three 10 ms groups: 50-59, 60-69 and 70-79 ms. The majority of cases (55) are in the 60-69 ms area. The overall distribution for all stop consonants is shown in Figure 3. The horizontal axis shows the duration groups in milliseconds, the vertical axis shows the number of VCV items in which the duration of the consonant occurs. The data show that voiceless stops (white) are longer than voiced ones (dark).

The distribution of voiceless stops shows a wider range than that of voiced ones. Comparing these data with the duration values of vowels in Figure 2, they show a wider distribution. Summarised values for all consonants are given in Table 8 in ms.

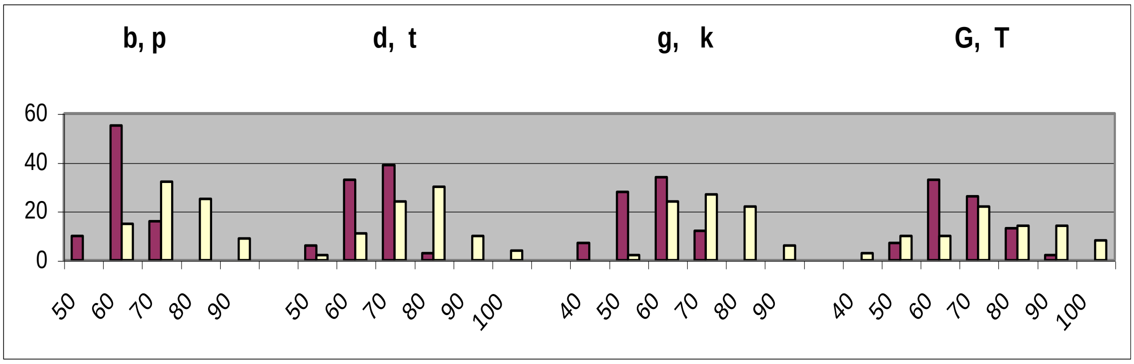

Figure 3. The distribution of the specific duration of Hungarian voiceless stops (white) and voiced ones (dark) in V**C**V combinations in continuous speech

Table 8. The specific duration values in ms for consonants in V**C**V positions in continuous speech

C	(b) [b]	(p) [p]	(d) [d]	(t) [t]	(g) [g]	(k) [k]	(G)[—]	(T) [X]	(m)[m]	(n) [n]	(N)[ᒍ]
Mean	65	77	70	76	62	74	68	76	67	48	66
Min.	55	61	53	61	47	59	53	47	51	36	45
Max.	78	94	80	96	78	92	87	88	82	64	88

(j) [j]	(h) [h]	(v) [v]	(f) [f]	(z) [z]	(s) [s]	(c) [ts]	(Z) [3]	(S) [Σ]	(C)[tΣ]	(l) [l]	(r) [r]
59	62	61	85	68	82	92	67	83	98	52	37
36	42	36	69	57	62	77	46	76	77	37	18
102	82	76	96	76	103	106	82	100	112	68	46

Comparing the results of Table 8 with the results of Kassai (1979) and Olaszy (1985) the order of the mean values of consonants coincides. Kassai gave the length order as: liquids< nasals< voiced stops< voiced spirants< voiceless stops< voiceless fricatives< voiceless affricates. If we follow this order, the data from Olaszy (1985) are: 45, 67, 69, 65, 117, 120, 125 ms, and the present data are: 44, 61, 66, 65, 76, 79, 95 ms. The difference between the data from 1985 and now can be explained with the material of the experiment. Olaszy (1985) measured the data mainly in two-syllable words, the present data were defined for continuous speech.

The average duration for all consonants in V**C**V position ranges from 37 ms to 98 ms.

Consonants in V**C**C and C**C**V combinations

Consonant clusters were examined only in V**C**C and in C**C**V combinations where the duration of the C in the middle position was defined. The results contain 2x4761 specific duration values for the 23 consonants for both types of combinations. The matrix for the sound (b) in V**C**C combination is shown in Table 9. Table 10 shows the specific durations of (b) in C**C**V combinations.

Table 9. The specific durations for (b) in V**C**C combinations in ms., in continuous speech

b	b	p	d	t	g	k	G	T	m	n	N	j	h	v	f	z	S	c	Z	S	C	l	r
A	87	79	71	78	78	76	80	69	59	71	89	69	79	68	69	68	76	89	76	79	73	67	79
a	88	80	72	80	79	77	81	70	60	72	90	70	80	69	70	70	77	90	77	80	74	68	80
o	93	85	77	85	84	82	86	75	61	77	95	75	85	68	71	75	82	95	82	85	79	73	85
u	96	88	80	88	87	85	89	78	62	80	98	78	88	68	70	78	85	98	85	88	82	76	88
U	87	79	71	79	78	76	80	69	59	71	89	69	79	69	69	69	76	89	76	79	73	67	79
i	93	85	77	85	84	82	86	75	60	77	95	75	85	65	70	75	82	95	82	85	79	73	85
E	86	78	70	78	77	75	79	68	58	70	88	68	78	68	68	68	75	88	75	78	72	66	78
O	86	78	70	78	77	75	79	68	58	71	88	68	78	68	68	68	75	88	75	78	72	66	78

e	95	87	79	86	86	84	88	77	60	79	97	77	87	66	72	76	84	97	84	87	81	75	87

Comparing the data with the durations of (b) in C**V**C combinations (Table 7) the conclusion is that the duration of (b) is longer in V**C**C and C**C**V combinations than in V**C**V position. The effect of articulation can be seen for example in the **(m)** column in Table 9., where the duration of (b) is shorter than in other columns. The same is the case in the **(m)** row of Table 10. This shorter duration of (b) in the (b)(m) and (m)(b) combinations may be explained by the fact that (b) loses its burst in this V**C**C combination because of the same bilabial articulation. In the mentioned C**C**V combination the voiced stop portion of (b) is shorter because of the shared articulation point. Similar but not so strong reduction can be seen in the columns of **(b)(v)** and **(b)(f)** in Table 9.

In general there is no significant difference between the durations of consonants in V**C**C and in C**C**V combinations.

Table 10. The specific durations for (b) in C**C**V combinations in ms., in continuous speech

b	**A**	**a**	**o**	**u**	**U**	**i**	**E**	**O**	**e**
b	80	82	82	75	86	88	85	76	80
p	76	77	77	71	81	83	81	72	76
d	55	57	57	50	60	63	60	51	55
t	52	54	54	47	57	60	57	48	52
g	73	75	75	68	79	81	79	70	73
k	52	54	54	47	57	59	57	48	52
G	72	74	74	67	78	80	77	68	72
T	83	85	85	78	89	91	88	79	83
m	42	44	44	37	47	49	47	38	42
n	62	63	63	57	67	69	67	58	62
N	60	62	62	55	66	68	65	57	60
j	49	51	51	44	55	57	54	45	49
h	77	79	79	72	82	85	82	73	77
v	72	74	74	67	77	80	77	68	72
f	66	67	67	61	71	73	71	62	66
z	71	73	73	66	77	79	76	68	71
s	61	63	63	56	66	69	66	57	61
c	73	75	75	68	79	81	78	70	73
Z	75	77	77	70	80	83	80	71	75
S	76	77	77	71	81	83	81	72	76
C	62	64	64	58	68	70	68	59	62
l	60	62	62	55	65	68	65	56	60
r	87	89	89	82	93	95	92	83	87

Comparisions with natural speech

As it was seen the results of the introduced inverse measurement, gave relevant duration data. The defined specific duration values are characteristic for Hungarian continuous speech. Using these data the basic, segmental level duration of every sound in an utterance can be given. On figure 4 the specific duration data and the measured ones of the beginning part of the sample sentence (1) *A tervezett tárgyalás után…*

[□] [t] [E] [r] [v] [E] [z] [E] [t:] [t] [a:] [r] [—] [□] [l] [a:] [Σ][u] [t] [a:] [n] are shown.

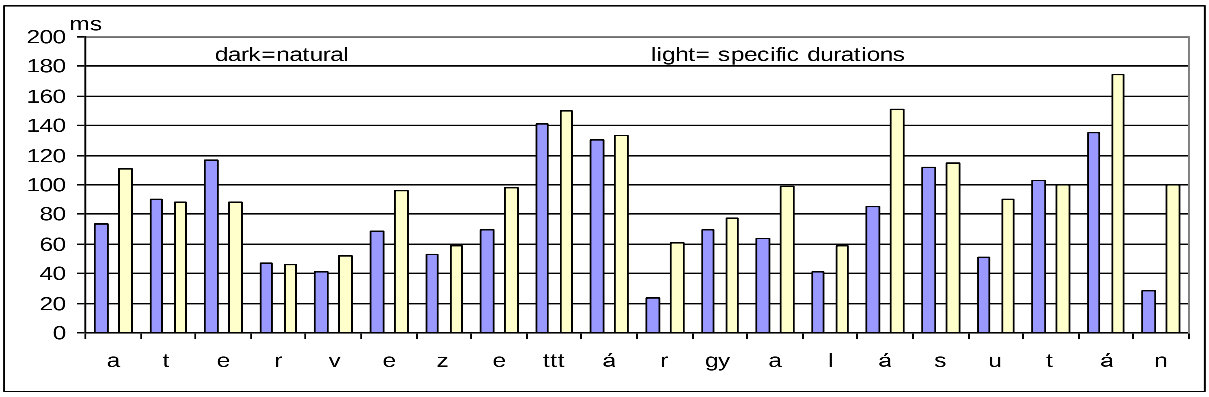

Figure 4.

The difference of specific and natural durations in the first part of the sample sentence

The main tendency of the two representations is similar, the most differences are in vowels. These differences will be eliminated by the suprasegmental level rules (2nd and 3rd step of the model).

Suprasegmental level duration modification rules

The second phase of the model contains 2 levels, e.g. word and phrase level modifications of the specific durations. The main goal here is to determine where and to what extent should we lengthen or shorten the specific duration of the given sound. The modification in the model is performed by using multiplying factors ranging: for shortening between 0.5 and 0.95, and for lengthening from 1.1 to 2. A certain factor is determined for each sound of the utterance and applied on the specific durations of the sounds. Comparing the specific durations and the natural ones in the sample sentences and taking into consideration of earlier results it was assumed that the further modification level is defined by the word. It was found that the length of the word and the inherent sound types and sound distribution influences the sound durations. The highest modification (3rd part of the model) concerns the effects of phrase structure.

Word level duration modifications

In this level the differences between natural and specific durations have been studied by making such duration pictures as was shown on Figure 4. It was found that in most cases the duration of vowels must be shortened, in less cases lengthened. 44 test sentences have been selected from the basic text material and these sentences were used for perceptual evaluation. The measurement set up for the test was basically the same as was shown on Figure 1. The only difference was that the test sentences were played with falling intonation (but without accent). Test persons have to compare the same sentence with two duration structures. The first was produced with specific durations, the second with modified one (using M2 factors and adjusting them to change the duration towards the natural values). The final M1 factors have been determined from the results of these listening tests. The perceptual test showed that word level modifications are more important than those of on sentence level. After word level modifications the duration structure of the utterance reaches in most cases the stage of 90% of the final, desired one. Another conclusion was that accents do not influence the duration map of the word, 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). Similar results are reported by Fónagy (1958) and Kovács (2002). Strong accents (e.g. focus) may be exceptions.

It was found that two features define the duration modification on word level. The sound map of the word and the length of the word. The sound map of the word shows the types of vowels, the consonant clusters, the place of sounds inside the word. All together twenty-five basic rules have been defined for the modification of short vowels. Examples are shown for the first short vowel of the word in Table 10. The data of this table show two things, i.e. the modifications are mostly shortenings, and the modification factors are vowel dependent. Separate rules (altogether 48) define the modification factors for long vowels. An example rule set is shown in Table 11 for the sound [a:]. Here separate rules define the modification as a function of the number of syllables in the word. The values of the modification factors express that the [a:] is shortened in the function of the number of syllables of the word consequently.

Table 10. Modifying multiplication factors for short vowels in the first syllable of the word longer than two syllables

Vowels Sequence	(i) [i]	(u) [u]	(U) [y]	(o) [o]	(a) [□]	(e) [E]	(O) [O]
# C V C1	1	0.8	0.8	0.8	0.8	0.8	0.8
# C V C1 C	1	1	1	1	1	0.8	1
# C V C2	1	0.8	0.8	0.9	1	0.8	1
#□ C V C1	0.8	0.8	0.8	1	0.9	0.9	0.8
#□ C V C1 C	1	1	0.8	1	1	0.9	1
#□ C V C2	0.8	0.8	0.8	1	1	0.9	1
# V C	0.8	0.8	0.8	1	0.9	1	1
# V C1 C	0.8	1	1	1	1	1	1
#□ V C	1	1	0.8	1	0.9	1	1
#□ V C1 C	1	1		1	1	1	1.1
# V C2 C	1	1	0.8	1	1	1.3	1

V = the short vowel in question, C = any consonant, C1 = any consonant but not [r, l], C2 = [r, l],
□ = article, # = absolute beginning position
multiplication factor = for example 0.8

Table 11. Modifying multiplication factors for [a:] if it is the only long vowel in the word (for 1,2,3,4,5 and 6 syllable words)

[a:] in the	Syll. Sequ.	1	2	3	4	5	6
1st syll.	VC1	_	1	0.9	0.8	0.8	0.75
	VC2	_	1.3	1.2	1.2	1.1	1
2nd syll.	VC1	_	_	0.9	0.85	0.85	0.8
	VC2	_	_	1	1	1	1
3rd syll.	VC1	_	_	_	0.9	0.8	0.8
	VC2	_	_	_	1	1	1
4th and more syll.	VC1	_	_	_	_	0.8	0.8
	VC2	_	_	_	_	1	1
Last syll.	VC1	1.2	0.9	0.85	0.8	0.8	0.8
	VC2	1.3	1.3	1.3	1.2	1.1	1.1

V = sound [a:] ,C1 = any consonant but not [r, l], C2 = [r, l]

multiplication factor = for example (1.3)

The specific duration of consonants is modified by 8 rules like: shorten the specific duration of CC and CCC clusters if they are not in the last word of the sentence ([ng] and [nk] combinations exceptions); shorten the specific duration of long stop consonants being at the end of the word (in sentence internal position).

The difference in number of both rule groups shows that in continuous speech the duration of vowels varies more dynamically than that of the consonants. The result of word level modification is expressed in the model as follows: every sound of the word gets a multiplication factor. For example the duration map (the series of M1 factors) of the word *láthatatlan* [l][a:][t][h][□][t][□][t][l][□][n] ‘invisible’ will show the following picture:

l(1) a:(0.8) t(0.9) h(0.9) □(0.9) t(1) □(1) t(0.9) (l)0.9 □(1) n(1)

Comparative measurements have been performed between natural and synthesised durations at this level. It was found that 90% of the modelled durations was very close to the natural one. Figure 5 shows again the duration map of the first part of the sample sentence (1) after performing the duration modification on word level (according to step 2). It can be seen, that the duration of vowels have been corrected towards the values of the natural sample. 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).

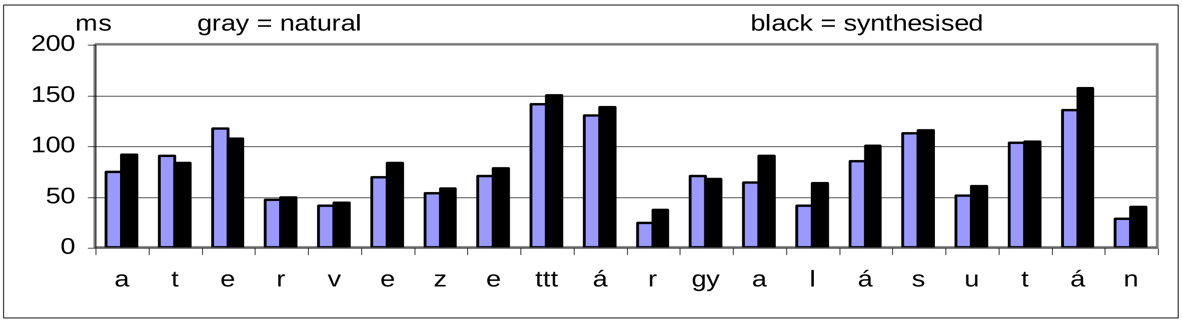

Figure 5. The corrected sound duration values of the sentence part of Figure 4. The horizontal axis contains the sounds with Hungarian letters. The articulation rate was 14 sounds/s in natural speech and 13 sounds/s in the synthesised one. Therefore most of modelled durations are slightly longer than the natural ones

Sentence level duration modification rules.

In the third step of the model only slight modifications are performed, mainly concerned lengthening: in the last word of the sentence and in the last syllable of the word in phrase boundaries and also in the first syllable of certain questions.

Conclusions

The proposed 3 level model gives very similar duration data than natural pronunciation. The most important part of the model is the module of first level, where the specific (segmental level) durations are determined. The presented indirect method for the definition of specific durations gives relevant data for the basic duration structure of the given language. Furthermore, this method gives us the possibility to define the only theoretically existing, segmental level specific sound durations in the form of exact data for every sound in every sound combination for the given language. The results for Hungarian showed that sound duration values defined with this inverse method correlate with the results of earlier investigations not only for Hungarian but also for English and French. This means that the inverse method presented can be successfully used for the definition of sound durations for continuous speech.

For Hungarian close to 20,000 individual specific sound durations (in triphone sound sequences for the middle sound) have been determined.

Specific duration values can represent a good basis for further (suprasegmental level) duration modifications. The second step of the model represents a semi-suprasegmental level in which fine modifications of specific durations in the word are summerised. Rules can be determined at this level to characterise the value of shortening or lengthening of the sounds. Measurement results show, that the sound durations in Hungarian are formed mostly on segmental level, and on word level. Phrase and sentence level modifications take less important role in forming final durations.

The advantage of this model can be summerised in five points:

(1) The sound durations can be determined in the function of adjacent sounds for continuous speech (independently of the speaker or the type of text in question). The specific durations give a good basis to make further studies about the organisation of the time structure of speech.

(2) The influence of articulation on duration can be separated from other possible factors;

(3) The results are recontrollable at any time

(4) The results show that sound durations are determined basically by the articulation and by the sound map of words in Hungarian. Phrase and sentence level influence on sound durations is small.

(5) The intrinsically existing specific durations are firstly expressed by actual numerical values. It could not be derived till now on the basis of direct measurements.

Acknowledgements

This research was partly supported by the Hungarian Research Found (OTKA T 037288). The algorithms for the synthesis supported perceptual tests and for the distribution measurements were programmed at the Budapest University of Technology and Economics, Depatrment of Telecommunications and Telematics by Péter Olaszi and Géza Kiss.

References

Allen, J. Hunnicut, S., Klatt, D. H. (1987) *From Text to Speech: the MITalk System.* Cambridge, U.K.: Cambridge Univ. Press,

Campbell, N. (1992). Syllable based segmental duration. In: *Talking Machines:Theories, Models, and Designs* (G. Bailly, C.Benoit and T.R. Sawallis, Editors.), pp. 211-224.Elsevier Science Publishers.

Fónagy I.: A hangsúlyról. *(About the accent)* NytudÉrt 18. 1958.

Kassai I.:(1979) *Időtartam és kvantitás a magyar nyelvben (Duration and quantity in Hungarian)*, Nyelvtudományi Értekezések 102. Budapest.

Huggins, A. W. F. (1972). Just noticable differences for segment duration in natural speech. *Journal of the Acoustic Society of America,* 51, 1270-1278.

Kovács Magdolna (2002): Tendenciák és szabályszerűségek a magánhangzó-időtartamok produkciójában és percepciójában. Debreceni Egyetem, Hungary pp. 118

Magdics, K. (1966). A magyar beszédhangok időtartama (Duration of Hungarian speech sounds), *Nyelvtudományi Közlemények 68*. pp.125-139. Budapest

O'Shaughnessy D. (1981). A study of French vowel and consonant durations. *Journal of Phonetics* 9, 385-406.

Olaszy G.(1985). *A magyar beszéd leggyakoribb hangsorépítõ elemeinek szerkezete és szintézise (The structure and synthesis of the most important building elements of Hungarian speech)*. Nyelvtudományi Értekezések 121. Budapest 1985.

Olaszy G.:(1991) The inherent time structure of Hungarian speech sounds. In.: *Temporal factors in speech,* (M. Gósy,.Editor).pp.107-138. MTA Nyelvtudományi Intézet, Budapest,

Olaszy, G., Olaszi, P. (1998). Hangidőtartamok mesterséges változtatása periódusok kivágásával, megismétlésével. (Changing the sound duration by inserting and deleting pitch periods) In: *Beszédkutatás'98* (M. Gósy, Editor) pp.151-162.MTA Nyelvtudományi Intézet, Budapest,

Olaszy G. – Németh, G. – Kiss G. (2002): Hungarian audiovisual prosody composer and TTS development tool. In: Prosody 2000. Editors: Puppel Stanislaw, Grazina Demenko. Poznan, 2001. 167-178.

H. Kato, M. Tsuzaki, and Y. Sagosaka. (1998). Effects of phonetic quality and duration on perceptual acceptibility of Temporal changes in speech. *Proc. of the 5th International Conference on Spoken Language Processing*, pp. 892–895. Sydney,

Santen van, J. P. H. (1992). Contextual effects on vowel duration. *Speech Communication* 11. pp. 513-546.

Santen van, J. P. H. (1998). Timing. In: *Multilingual text-to-speech synthesis: The Bell Labs Approach*. (R. Sproat Editor). pp. 115-139. Kluwer Academic Publishers

Shiga Y., Matsuura H., and Nitta T. (1998).Segmental duration control based on an articulatory model *Proc. of the 5th International Conference on Spoken Language Processing*, pp. 1244–1247. Sydney,

Zellner, B. (1994). Pauses and the temporal structure of speech. In *Fundamentals of Speech Synthesis and Speech Recognition,* (E. Keller Editor). pp. 42–62. John Wiley and Sons, New York,