

A MODEL OF PHONEME DURATIONS BASED ON THE ANALYSIS OF A READ DUTCH TEXT

B. Van Coile⁰

ABSTRACT

This paper describes the vowel part of a durational model for Dutch. A Dutch text, with a total duration of more than 8 minutes, was analysed. Previously observed durational phenomena were confirmed: short/long vowels, word-final lengthening, prepausal lengthening, the influence of prominence, ... Most of these trends were also quantified.

INTRODUCTION

In order to enhance the naturalness of synthetic speech, we need a durational model that attributes a well estimated duration to each phoneme. A lot of studies on durational phenomena use corpora of words (real words or nonsense words). In some of these corpora the words are spoken in isolation, in others they are pronounced in a carrier phrase (among others ref 1,2,3). In order to develop a durational model for continuous speech, some researchers have analysed the phone durations measured in a read text (among others ref 4,5,6). This paper describes such an approach for Dutch.

MATERIAL

A Dutch text with a duration of 8'30'' was read by a female, native speaker. The text was tape-recorded, digitised and stored on computer for further processing. The segmentation of the corpus was performed manually through visual inspection of computer-graphics. Each graphic showed a speech waveform together with classic parameters such as energy contours, the zero crossing rate, the spectral centre of gravity, ... If needed, a digital playback system was used in order to listen simultaneously to the signals. A broad phonetic transcription of the text was made.

Two different stress marks were used: one for lexical stress and a second for the sentence accents. Each word received (at least) one stress mark. Sentence accents were determined independently by the speaker and the author. While listening to the tape-recorded text, they each marked those words that sounded dominant. The agreement between their results was remarkable. In case of disagreement, the lexical option was chosen. Subsequently, the constituent structure of each sentence was determined together with part-of-speech information. The 'Uit den Boogaart' corpus (ref 7) was used in order to determine each word's frequency of occurrence in Dutch.

⁰ Laboratory of Electronics and Metrology, State University of Ghent,
St Pietersnieuwstraat 41, 9000 Ghent, Belgium

METHOD

A stepwise factor selection method was used to create several durational models with increasing complexity and prediction accuracy. During this process, the theory of least squares was used to determine the optimal parameters of the models.

AN ANALYSIS OF VOWEL DURATIONS

Figure 1 shows the durational histogram of all vowels pooled. The variance of the distribution is used as a reference value for the evaluation of the durational models. The simplest model we can propose, attributes the same value of 69 ms to each vowel.

(1) Intrinsic vowel duration

Each phoneme has its own intrinsic duration. Figure 2 shows the mean values (and the standard deviations) for the duration of the different Dutch vowels. All occurrences of the same vowel were pooled. A durational model that assigns each vowel its own intrinsic duration (i.e. the mean value from the figure) explains 56 % of the original variance. It is important to note the clear durational distinction between two sets of vowels: the long vowels [e, ø, o, a, ei, ay, au] and the short vowels [i, y, u, I, A, ɔ, E, ʌ]. The schwa is clearly the shortest vowel. The simple durational model (no 1) that assigns the long vowels, the short vowels and the schwa a different duration accounts for 54 % of the total variance.

figure 1: the durational histogram of all vowels pooled

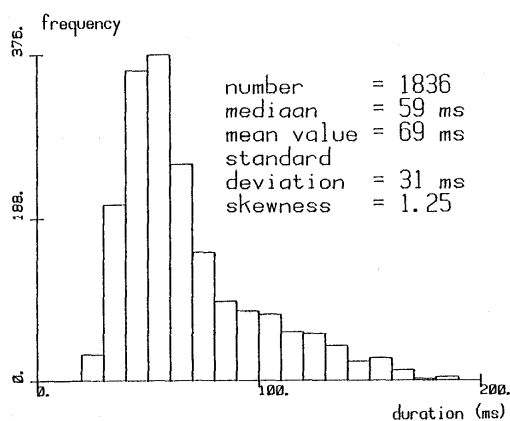
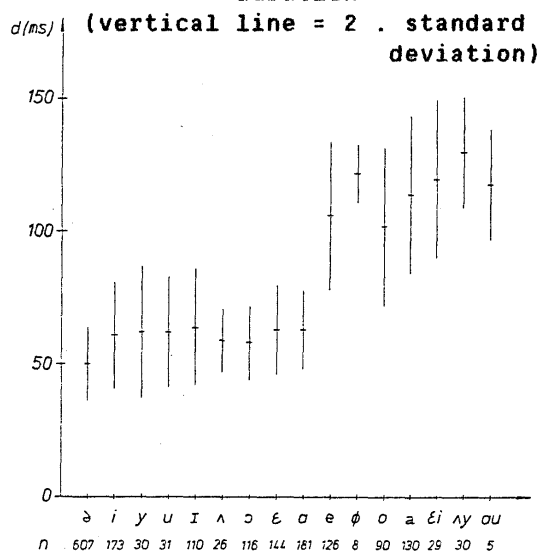


figure 2: the intrinsic vowel duration



(2) Stress, function word versus content word

An important factor which affects the duration of vowels and which is related to the aspect of stress, is the distinction between function words and content words. Vowels in function words are shorter than those in content words. A possible classification into both word classes is based on part-of-speech information: closed-class words

(prepositions, auxiliary verbs, pronouns, articles and conjunctions) are considered to be function words. The word's unconditional frequency of occurrence can also be used as a classification criterium. In order to determine the optimal classification method, we used the following procedure: all non-prepausal, lexically stressed vowels were selected. The variance of this group of vowels was taken as a reference during the experiment. Word frequencies were taken from 'Uit den Boogaart' (ref 7). According to a word frequency threshold, the vowels were subdivided into two classes. Figure 3 shows the remaining portion of the variance for different word frequency thresholds. A frequency threshold of 2000, explains 29 % of the variance. A slightly better result (31.5%) is obtained if all personal pronouns are considered as function words too. A classification based on part-of-speech information, explains only 22 % of the total variance. In what follows, the optimal distinction between the two word classes was used. Figure 4 clearly shows the influence of prominence and stress on vowel duration. The corresponding durational model (no 2) which explains 64 % of the total variance is shown in table 1, rules 1 to 12.

figure 3: the remaining portion of the variance for different frequency thresholds

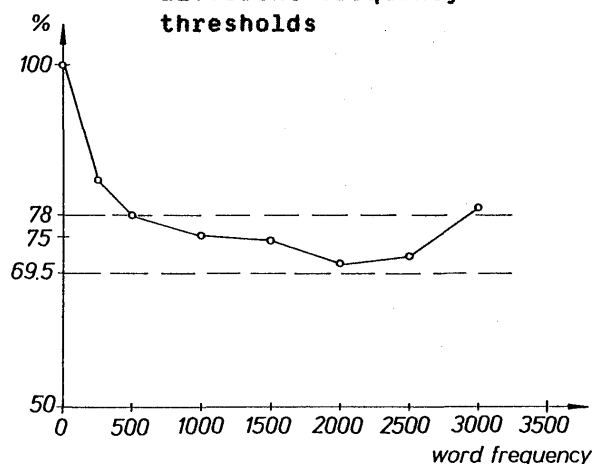
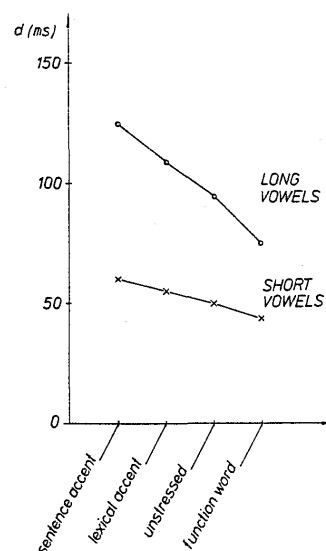


figure 4: durational mean values



(3) Positional conditions

The third model which explains 73 % of the total variance takes into account the position of the vowel in the word and its position in the phrase. Rules 1 to 21 (see table 1) constitute this model (no 3). Several modification rules are introduced: they modify a previously assigned duration with a certain factor. For the calculation of the optimal parameters shown in table 1 we used the term phrase-final as a synonym of prepausal. The following influences are clearly seen:

- Vowels tend to be larger in phrase final positions than in others.
- Vowels with stress become shorter when the number of syllables which remain to be spoken increases. This effect is especially clear in phrase final positions.
- Unstressed vowels are longer in word final positions than in non word final positions.

These observations are in agreement with the results described in literature (among others: ref 2,3).

(4) Consonantal influence

The influence of the following consonant on vowel durations is clearly seen in phrase final situations. In other circumstances the effect is very small. Model 4 uses the same rules as the previous model, supplemented by modification rules for different consonantal conditions. This model explains 76 % of the total variance.

TABLE 1: durational rules with optimal values for model 4.
Syl=X means that X syllables remain to be spoken

(1)	{i,y,u,I,ʌ,ɔ,ɛ,ɑ}	-->	[+ short]	
(2)	{e,ɸ,o,a,ɛi,ʌy,au}	-->	[- short]	
(3)	{i,y,u}	-->	[- short]	/_r
(4)	[+ short, sentence accent]	-->		69 ms
(5)	[+ short, lexical accent]	-->		64 ms
(6)	[+ short, - accent]	-->		59 ms
(7)	[+ short, function word]	-->		52 ms
(8)	[- short, sentence accent]	-->		120 ms
(9)	[- short, lexical accent]	-->		104 ms
(10)	[- short, - accent]	-->		92 ms
(11)	[- short, function word]	-->		70 ms
(12)	[schwa]	-->		47 ms
(13)	[+ accent, + phrase final, Syl=0]	-->	*	1.40
(14)	[+ accent, + phrase final, Syl=1]	-->	*	1.25
(15)	[+ accent, + phrase final, Syl>1]	-->	*	1.02
(16)	[+ accent, - phrase final, Syl=1]	-->	*	0.96
(17)	[+ accent, - phrase final, Syl>1]	-->	*	0.94
(18)	[- accent, + phrase final, + word final]	-->	*	1.56
(19)	[- accent, + phrase final, - word final]	-->	*	1.04
(20)	[- accent, - phrase final, - word final]	-->	*	0.93
(21)	[schwa, phrase final, word final]	-->	*	1.20
	{[accent, phrase final, Syl<2],			
	[- accent, phrase final, word final]}			
(22)		/_[unvoiced plosive]	-->	* 0.85
(23)		/_[{nasal},{liquid}]	-->	* 0.95
(24)		/_[voiced plosive]	-->	* 1.04
(25)		/_[voiced fricative]	-->	* 1.14

CONCLUSION

Several durational models with increasing complexity and accuracy were proposed. They are based on the analysis of production data measured in a read Dutch text. No attempt is made to justify the models on perceptual grounds. The better performing models (no 3,4) account for the bulk of the durational variations seen in the text. As a consequence they are useful in text-to-speech synthesis.

REFERENCES

1. A.S. House, JASA 33, 1174 (1961)
2. S.G. Nootboom, Philips Research Reports, Suppl. 5 (1972)
3. D. Klatt, J. Speech Hearing Res. 17, 51 (1974)
4. N. Umeda, JASA 58, 434 (1975)
5. D. Klatt, J. Phonetics 3, 129 (1975)
6. D. O'Shaughnessy, JASA 76, 1664 (1984)
7. P.C. Uit den Boogaart, Woordfrequenties (Oosthoek, Scheldema & Holkema, Utrecht, 1975) p 471.