A duration model for phonetic units in isolated Dutch words

H. Strik & E. Konst

Introduction

The work described in this paper was done as part of the ESPRIT project POLYGLOT. The aim of the project is to develop a multi-lingual Speech-to-Text and Text-to-Speech system. Part of the work comprises the adaptation of a large vocabulary isolated word speech recognition (IWSR) system, originally developed for Italian (Billi et al., 1989), to a number of other European languages, including Dutch.

The IWSR system runs on a MS-DOS PC that uses one or two special purpose plug-in boards. The speech is picked up by a table-mounted microphone and A/D-converted with a sampling frequency of 16 kHz and a 12 bit resolution. For each 10 ms frame 20 LPC Cepstrum coefficients are calculated, which are used to calculate the acoustic distance to a set of stored prototypes. A prototype represents a phonetic unit, and because these phonetic units are not always phonemes we will use the terms prototype and phonetic unit throughout this article. The resulting lattice of prototype labels is then submitted to a dynamic programming procedure that outputs the most likely string of prototype labels. To calculate the optimal path in the lattice, the dynamic programming algorithm uses the acoustic distances (which are stored in the lattice) and statistics on prototype frequency, prototype-pair frequency, and the duration of prototypes. The prototype string is then used for fast lexical access, to retrieve the most likely word candidates. In a later stage of the recognition process, called Fine Phonetic Analysis (FPA), a top-down algorithm is used to find the best candidate. For FPA the statistics on the duration of prototypes is also required (Billi et al., 1989).

The statistics on prototype frequency and prototype-pair frequency can be derived from large text corpora with tools that were developed in a previous ESPRIT project, entitled "Linguistic Analysis of European Languages" (Boves, 1989). At the start of the POLYGLOT project the statistics on the duration of prototypes in isolated Dutch words were not available. In this article a description is given of the material and the procedure that was used to derive a duration model for phonetic units in isolated Dutch words.

1 Material and segmentation

The speech material consists of 500 different, meaningful Dutch words. The database includes the 200 most frequent words of Dutch (obtained from the CELEX lexical database), plus 300 words that represent the phonemic distribution of Dutch. These 500 words were divided in 10 subsets of 50 words. All the words were pronounced clearly in a normal speech rate by one male untrained speaker. After each word there was a pause of at least 1 second, and after each subset there was a pause of several minutes. The speech signals were A/D-converted with a 16 kHz sampling rate.

The 500 words were segmented manually by looking at the oscillogram and the spectrogram of the signal, and by listening to the signal. The segmentation was exhaustive, no part of the utterances was left unsegmented. For labelling, a computer phonetic alphabet (CPA) was used, comprising 77 different symbols. The CPA symbols are built up of one or two ASCII characters. Only 42 of the 77 symbols were used for the duration model, for reasons that are beyond the scope of this article (see e.g. Drexler et al., 1991). These 42 symbols are listed in Table II, together with a Dutch example word for each symbol. The 35 symbols that were also used during segmentation, but are not incorporated in the duration model are listed in Table I.

Table I. The 35 symbols that were not used for the duration model.

- /-/ silence, usually at the beginning and end of words
- /b0,d0,g0,G0,l0,v0,w0,z0/ voiceless part of a (voiced) consonant
- /hv/ voiced part of the (voiceless) consonant /h/
- /ã/ nasalized vowel /a/
- /Z/ voiced palato-alveolar fricative, as in the word 'jacquet'
- /?/ glottal stop
- /@#,a#,b#,d#,e#,f#,i#,j#,l#,m#,n#,N#,o#,q#,r#,s#,u#,w#,y#,X#,z#/ the final parts of some words in which there is still a small signal, but it is not characteristic for the preceding phoneme anymore
- /#/ unknown symbol, for parts of the words that could not be classified by any of the 76 other symbols

In Table II the number of occurrences are given for each of the 42 symbols that were used for the duration model. In all, 3276 symbols were used for deriving the duration model (see row 1 of Table II). Furthermore, there were 1000 silences at the word boundaries and 409 other occurrences of the 35 symbols listed above, which makes a grand total of 4685 symbols in this database.

2 Method

The goal of our procedure was to group the prototypes in classes, and to obtain, for each class, the conditional probability of a certain duration given the class: P(duration|class).

The first step in achieving this goal was to write a program that could read the label files, and calculate duration distributions for classes of phonetic units. A program, called PHONSTAT (PHONetic unit STATistics), was written. The input of PHONSTAT is a list with the names of the label files, a list with label symbols that defines the classes, and the bin width in milli-seconds (the width of a bar in the duration histogram). A bin width of 10 ms is required by the POLYGLOT IWSR system. The output of the program PHONSTAT is a histogram for each class defined. By way of example, the histogram for one class of phonetic units, viz. the schwa vowels, is given in Fig. 1a (the definition of the classes is given in section 4). The horizontal axis consists of 40 bins of 10 ms each, so statistics are presented for durations from 10 ms up to 400 ms. To obtain the conditional probability of a certain duration given the class, the number of occurrences in each bin is divided by the total number of occurrences for the given class.

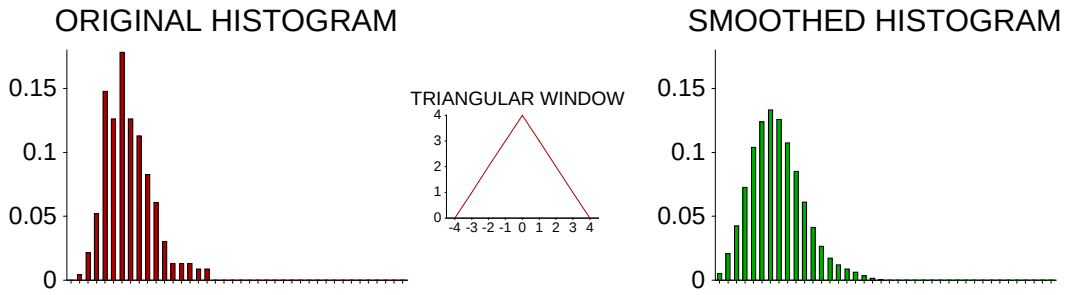

Figure. 1. Histograms and smoothing window.
(a) Original histogram for schwa vowels; (b) Triangular window;
(c) Smoothed histogram for schwa vowels.

The classes were not defined a priori. The main criterion that was used in grouping the prototypes into classes was that the duration distributions of the prototypes in 1 class should be comparable. Another criterion was that classes should be created which were phonetically meaningful. The procedure that was used can be divided in 4 stages:

1. calculate a histogram for each of the 42 prototypes
2. compare histograms of individual prototypes, and group them into classes
3. calculate a histogram for each of the classes defined;
   if the result is not satisfactory go back to stage 2.
4. smooth the histograms

The duration distribution of a class of prototypes (viz. for the schwa vowels) is given in Fig. 1a. If the number of occurrences of the prototypes in this class had been larger, the histogram would probably have been smoother. To approximate the probable distribution (i.e. the distribution if the number of occurrences had been much larger), the

histograms are smoothed (low-pass filtered) in stage 4. Smoothing is done by calculating the convolution of the histogram with a triangular window (see Fig. 1b). The resulting, smoothed histogram is shown in Fig. 1c.

3 Results and discussion

In Table II statistics are presented for three kinds of classes. First of all, for classes that consist of only one phonetic unit. These classes will be called simple classes, and their names are the corresponding CPA symbols. Secondly, for classes that result from joining several simple classes. These classes will be called compound classes. They are given names that are written in capitals (like 'BURSTS', 'VOICE BARS', etc.). And thirdly, for the class that results from joining all 42 simple classes, which is called 'ALL 42 SYMBOLS'.

Table II. Statistical data for the class 'ALL 42 SYMBOLS' and for 10 compound and 42 simple classes (for an explanation, see text).

class	example	nr.	med	mean±sd	KvB
ALL 42 SYMBOLS		3276	82	92±61	
BURSTS		455	20	23±16	
bh	see b	15	12	12± 6	
dh	see d	58	12	13± 6	
ph	see p	61	17	21±14	
th	see t	227	20	22±15	
kh	see k	94	33	34±18	
VOICE BARS		151	52	61±39	
G	gal	11	27	34±18	
g	goal	6	61	50±20	
d	dak	78	45	57±39	
b	bad	56	71	72±39	
LIQUIDS & GLIDES		517	62	69±31	
h	had	41	58	59±21	
r	rat	224	58	63±27	
l	lat	161	66	75±34	
j	jat	29	72	77±37	
w	wat	62	75	78±30	

Table II. (continued)

class	example	nr.	med	mean±sd	KvB
SCHWA VOWELS		230	67	72±28	
@	all<u>e</u>	230	67	72±28	
OCCLUSIONS		415	66	78±40	
t	<u>t</u>est	244	65	75±39	
k	<u>k</u>ak	100	64	77±39	
p	<u>p</u>ut	71	83	90±40	
VOICED FRICATIVES		79	85	92±39	
v	<u>v</u>at	39	82	91±39	
z	<u>z</u>at	40	95	93±38	
SHORT VOWELS		454	99	102±28	
I	b<u>i</u>d	76	89	91±24	124±24
U	p<u>u</u>t	25	95	98±29	108±22
O	b<u>o</u>d	56	97	99±25	108±23
A	b<u>a</u>d	128	100	103±23	120±19
i	b<u>ie</u>t	53	95	105±37	140±12
E	b<u>e</u>d	79	105	107±25	124±15
u	b<u>oe</u>k	30	114	111±25	150±17
y	b<u>uu</u>rt	7	154	140±55	136±23
NASALS		288	110	110±41	
m	<u>m</u>at	81	106	108±38	
n	<u>n</u>at	164	110	110±43	
N	la<u>ng</u>	43	117	116±36	
UNVOICED FRICATIVES		402	130	147±67	
f	<u>f</u>iets	65	114	126±46	
X	la<u>ch</u>	134	125	132±56	
s	<u>s</u>ap	193	139	162±76	
S	<u>sj</u>aal	10	158	169±52	
LONG VOWELS		285	185	182±66	
o	b<u>oo</u>t	55	163	162±61	178±15
e	b<u>ee</u>t	78	180	176±55	180±25
AU	b<u>ou</u>t	26	171	178±93	
a	b<u>aa</u>t	68	193	186±61	204±11
EU	b<u>eu</u>k	7	195	193±30	180±21
EI	b<u>ij</u>t	41	205	204±76	
UI	b<u>ui</u>t	10	226	205±46	

In the first row the statistics are given for all 3276 occurrences of the 42 symbols. The rows with the statistics for the compound classes are ordered according to their average value (in ascending order). Directly following a row with statistics for a compound class, are rows with statistics of the simple classes that make up the compound class. Within a compound class, the rows for the simple classes are also ordered according to their average value (again, in ascending order).

The name of the class is given in the first column, as explained above. For simple classes a typical Dutch example word is given in the second column. The third column states the magnitude of a class; in the fourth column the median value of the durations; in the fifth column the average duration with the standard deviation; and finally the sixth column are lists the average and standard deviation of the duration of vowels, as measured by Koopmans van Beinum (1980). All duration values are given in milliseconds.

The purpose of the current study is the development of a duration model for phonetic units that can be used in an isolated word speech recognition system. For that purpose the best division is the one given in Table II. The smoothed histograms for the 10 compound classes are given in Figure 2. From Table II and Figure 2 we can infer that the duration distributions of the phonetic units in the different compound classes are coherent, except for the compound class 'UNVOICED FRICATIVES'. The smoothed histogram for the 'UNVOICED FRICATIVES' is bimodal (see Fig. 2). At first sight, the values in Table II suggest that it would have seemed better to divide this class into two classes: one with /f/ and /X/, and one with /s/ and /S/. We tried this, but the resulting histograms of both classes were still bimodal. Therefore, we decided to make one compound class for the unvoiced fricatives. The reason of the bimodality probably is that unvoiced fricatives in word final position are often significantly longer than in other positions.

The main criterion that was used in defining the classes was that phonetic units with similar duration distributions were grouped together. The result was that phonetically similar events were automatically grouped together. However, there were two exceptions. First, the unvoiced fricative /h/ is generally much shorter than the other unvoiced fricatives (see Table II). Because its duration distribution most closely resembles that of the liquids and glides, the /h/ was put into the class 'LIQUIDS & GLIDES'. And secondly, the average duration of the occurrences of the /G/ in our database was much shorter than that of the other voice bars. This was probably caused by the fact that 7 ou of 11 occurrences of the /G/ were partly devoiced, and therefore labelled as /G0/. Nevertheless, on phonetic grounds, we decided to put the /G/ in the class 'VOICE BARS'.

In phonology, the Dutch vowels are usually divided into two groups according to their length: the short vowels /I,U,O,A,E/, and the long vowels /i,u,y,o,e,a,AU,EI,UI,EU/ (see e.g. Booij, 1981). However, the phonetic realization of the vowels /i,u,y/ is generally short, except before /r/ (Nooteboom, 1972). In our database 4 out of 7 occurrences of an /y/ are followed by /r/. Therefore, the average duration of the /y/ is probably relatively

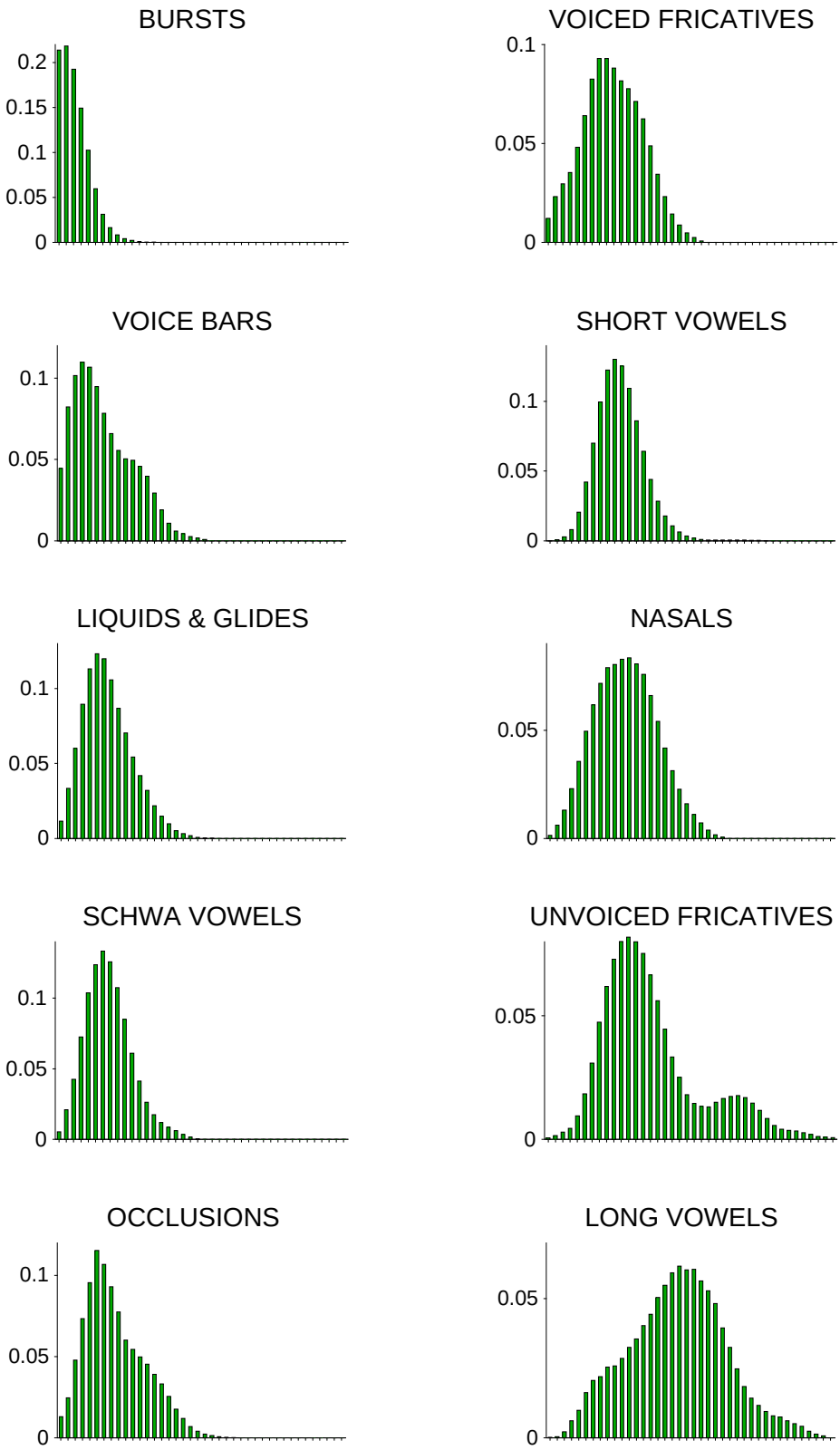

Figure 2. Smoothed histograms for the 10 compound classes.

high. The division of Dutch vowels into short and long vowels as given in Table II, is in accordance with other measurements of the duration of Dutch vowels that have been reported in the literature (Van Coile, 1990; Koopmans van Beinum, 1980; Nooteboom & Slis, 1972).

Koopmans van Beinum (1980) measured durations of vowels in different speech conditions. Of these conditions, the monosyllabic word condition most closely matches the speech condition used in the current experiment. The numbers in the sixth column of Table II are the average and the standard deviation of 5 measurements of the duration of vowels in isolated monosyllabic words spoken by a male untrained speaker (Koopmans van Beinum, 1980: 137). Except for /y/ and /EU/, the average values found by Koopmans van Beinum were larger than the average values obtained for the current database. This is not surprising, as the current database contains both mono- and polysyllabic words. It was explained above why the average value of /y/ is probably relatively high. Koopmans van Beinum also mentioned that /y/ and especially /EU/ turned out to occur much less frequently than the other vowels. This is reflected in the number of occurrences of these vowels given in Table II. Furthermore, she found that the reduction in the duration of the vowel /EU/ in unstressed positions was much less than the reduction of other vowels in unstressed positions (compared to stressed positions). This could explain why the average duration of the 7 occurrences of the vowel /EU/ in our database is not smaller than the average duration found by Koopmans van Beinum.

4 References

Billi, R., Arman, G., Cericola D., Massia, G., Mollo, M., Tafini, F., Varese, G. & V. Vittorelli (1989) A PC-based large vocabulary isolated word speech recognition system. *Proceedings of EUROSPEECH-89*, 2, 157-160.

Booij, G.E. (1981) *Generatieve fonologie van het Nederlands.* Utrecht/Antwerpen: Het Spectrum.

Boves, L. (1989) A multi-lingual language model for large vocabulary speech recognition. *Proceedings of EUROSPEECH-89*, 2, 168-171.

Coile, B. van (1990) *Tekst-naar-spraak omzetting: een taalkundig, fonetisch en akoestisch probleem.* Unpublished Ph.D. thesis.

Drexler, H., Roddeman, R., Boves, L., & H. Strik (1991) Optimizing lexical fast search in a large vocabulary isolated word speech recognition system. *Proceedings of EUROSPEECH-91*, 3, 1401-1404.

Koopmans van Beinum, F.J. (1980) *Vowel contrast reduction: an acoustic and perceptual study of Dutch vowels in various speech conditions.* Unpublished Ph.D. thesis.

Nooteboom, S.G. (1972) *Production and perception of vowel duration: a study of durational properties of vowels in Dutch.* Unpublished Ph.D. thesis.

Nooteboom, S.G. & I.H. Slis, (1972) The phonetic feature of vowel length in Dutch. *Language & Speech*, 15, 301-316.