UvA-DARE (Digital Academic Repository)

Modelling of phone duration (using the TIMID database) and its potential benefit for ASR

Pols, L.C.W.; Wang, X.; ten Bosch, L.F.M.

DOI

[10.1016/0167-6393\(96\)00033-7](https://doi.org/10.1016/0167-6393(96)00033-7)

Publication date

1996

Published in

Speech Communication

[Link to publication](#)

Citation for published version (APA):

Pols, L. C. W., Wang, X., & ten Bosch, L. F. M. (1996). Modelling of phone duration (using the TIMID database) and its potential benefit for ASR. *Speech Communication*, 19(2), 161-176. [https://doi.org/10.1016/0167-6393\(96\)00033-7](https://doi.org/10.1016/0167-6393(96)00033-7)

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

ELSEVIER

Speech Communication 19 (1996) 161–176

SPEECH
COMMUNICATION

Modelling of phone duration (using the TIMIT database) and its potential benefit for ASR ¹

Louis C.W. Pols ^{a,*}, Xue Wang ^a, Louis F.M. ten Bosch ^b

^a *Institute of Phonetic Sciences / IFOTT, University of Amsterdam, Amsterdam, The Netherlands*

^b *Lernout and Hauspie Speech Products N.V., Brussels, Belgium*

Received 23 January 1996; revised 21 May 1996

Abstract

As indicated by Boulard et al. (1996), the best and simplest solution so far in standard ASR technology to implement durational knowledge, seems to consist of imposing a (trained) minimum segment duration, simply by duplicating or adding states that cannot be skipped. We want to argue that recognition performance can be further improved by incorporating “specific knowledge” (such as duration and pitch) into the recognizer. This can be achieved by optimising the probabilistic acoustic and language models, and probably also by a postprocessing step that is fully based on this specific knowledge. The widely available, hand-segmented, TIMIT database was used by us to extract duration regularities, that persist despite the great speaker variability. Two main approaches were used. In the first approach, duration distributions are considered for single phones, as well as for various broader classes, such as those specified by long or short vowels, word stress, syllable position within the word and within an utterance, post-vocalic consonants, and utterance speaking rate. The other approach is to use a hierarchically structured analysis of variance to study the numerical contributions of 11 different factors to the variation in duration. Several systematic effects have been found, but several other effects appeared to be obscured by the inherent variability in this speech material. Whether this specific use of knowledge about duration in a post-processor will actually improve recognition performance still has to be shown. However, in line with the prophetic message of Boulard et al.’s paper, we here consider the improvement of performance as of secondary importance.

Résumé

Comme mentionné dans l’article récent de Boulard et al. (1996) publié dans ce journal, la solution, à ce jour la meilleure et la plus simple, pour introduire des connaissances sur la durée dans les systèmes de reconnaissance de parole standard serait d’imposer une durée minimale (apprise) par segment, simplement en dupliquant ou en ajoutant des états qui ne peuvent pas être omis. Dans notre article, nous défendons le point de vue que les performances de reconnaissance peuvent être encore améliorées en incorporant des connaissances “spécifiques” (comme la durée et le pitch) au sein des systèmes. Ceci peut être obtenu en optimisant les modèles acoustiques et de langage, et probablement aussi par un post-traitement entièrement basé sur cette connaissance spécifique. Nous avons utilisé la base de données TIMIT, largement répandue et

* Corresponding author.

¹ This paper should, at least partly, be considered as a reaction to some specific remarks in the stimulating paper of Boulard et al. (1996).

segmentée manuellement, pour extraire les régularités de durée qui persistent, malgré la grande variabilité inter- et intra-locuteur. Deux approches ont été principalement mises en oeuvre. Dans la première approche, on considère les distributions de durée pour les phones individuels, ainsi que pour des classes plus larges, comme celles spécifiées par les voyelles longues ou courtes, l'accent lexical, la position de la syllable au sein du mot et du syntagme, les consonnes post-vocaliques et la vitesse d'articulation. L'autre approche utilise une analyse hiérarchique de la variance pour étudier la contribution numérique de 11 facteurs différents aux variations de durée. Plusieurs effets systématiques ont été identifiés. D'autres effets semblent être obscurcis par la variabilité inhérente au matériau de parole utilisé. Quant à savoir si l'exploitation de cette connaissance sur la durée dans un étage de post-traitement va réellement améliorer les performances de reconnaissance, ceci reste à montrer. Cependant, conformément à l'esprit du message prophétique de l'article de Boulard et al, nous considérons ici l'amélioration des performances comme un enjeu d'importance pour l'instant secondaire.

1. Introduction

Many phoneticians feel challenged by the surprisingly high performance of several HMM- and/or ANN-based continuous speech recognition systems. In a probabilistic approach, the possible invariance and the apparent variability problem is mainly handled by systematically analysing large data sets, in which all possible sources of variation are properly represented, and by subsequently considering the involved probability distributions. However, for instance, for well articulated stressed vowels a more peripheral spectral quality, a more extended formant transition, a longer duration, and a higher energy are evident compared to those characteristics for similar but unstressed vowels. Such effects are not just accidental but are rather consistent and predictable. This "specific knowledge" could be added to the statistically defined "general knowledge" (provided by the structure of the conventional monophone-based HMM recognisers), thus improving the overall system performance. This is of course not at all a new idea, it is just more easily said than done.

Two main levels may be considered at which such specific knowledge can be integrated into statistically-based ASR-systems:

(1) At one level one makes intelligent choices for such aspects as the pre-processor (e.g. critical band type filtering, or RASTA or PLP), the parametric representation (e.g. cepstral coefficients plus overall energy, but also delta and delta-delta components), the descriptive units (e.g. monophones or triphones), and the language model used (e.g. bigram). Boulard et al. (1996) put much emphasis on this level.

(2) At the other level one may attempt to model certain sources of variation. Actually the above-men-

tioned triphone model also belongs here, since it is one, perhaps not very intelligent but still very successful way to cope with coarticulation. Gender-specific phone models are another example, just as double representations for each monophone or triphone model, depending on whether the central vowel is stressed or unstressed. Certain forms of tied mixtures also belong to this level, albeit in an opposite fashion, since here *lack of variation* is used to create broader categories. Rate-specific models could also belong here. At best the aforementioned approaches manage to extract from external sources specific information, such as gender (male versus female), word stress (unstressed or stressed, or perhaps even primary and secondary stress), or speaking rate (slow, medium, fast) and then selectively choose the appropriate reference models. It would be more challenging to be able to make use of continua rather than two or three discrete classes.

Duration modelling could in principle also contribute to better recognition performance. As Boulard et al. (1996) already indicate, so far the best and simplest solution seems to consist of imposing a (trained) minimum segment duration, simply by duplicating or adding states that cannot be skipped in the dynamic programming. Sometimes explicit state duration pdf, or so-called hidden semi-Markov models (HSMM) are used for duration modelling (Levinson, 1986; Hochberg and Silverman, 1993). What we have in mind, though, implies an additional post-processing step in which, starting from the word graph or the *N*-best hypotheses, a rescoring is done based on various durational characteristics, such as long versus short vowel, presence or absence of lexical stress, being in utterance-final or non-final position, and certain more subtle aspects, such as a

longer duration for a vowel preceding a voiced rather than an unvoiced plosive.

In order to be able to evaluate whether the implementation of such durational knowledge sources is useful (at least in terms of improving recognition performance, but, as far as phoneticians are concerned, also in terms of a better insight in the features controlling duration), one first of all has to retrieve duration characteristics from a database, next one has to implement the found regularities in terms of a post processing, after which one can test whether recognition performance has improved. Once this is shown to be successful for duration modelling, one could try to extend this approach to other areas, such as pitch modelling, as well.

In the present paper we will mainly present data concerning the first step (collect and study data on durational characteristics). In line with the adagium of the Boulard et al. (1996) paper, the improvement of the recognition performance is now of secondary importance. Our present focus is the study of incorporation of knowledge about duration into HMM-based recognizers.

In this study, more extensively described in a doctoral thesis by the second author, we have made use of the HMM Toolkit (HTK) (Young, 1992; Woodland and Young, 1993). The chosen features are the 12 mel-based cepstral coefficients plus utterance-normalised energy (13 features themselves, delta, and delta-delta, 39 features altogether). 50 monophone, multi-state linear models were used with an 8 component Gaussian mixture per state and a diagonal covariance matrix. No triphone models were used. The frame rate is 8 ms. As databases both Dutch material from one speaker as well as American-English material from many different speakers (the TIMIT database) were used. Since the TIMIT-database is extensively studied (e.g. Byrd, 1992; Keating et al., 1994; Gauvain et al., 1994; Lamel and Gauvain, 1993a,b; Young and Woodland, 1993; Lee and Hon, 1989), we will limit ourselves in the present paper to TIMIT.

2. Phone duration distributions

The TIMIT database, which is available on CD-ROM (Garofolo et al., 1993; Lamel et al., 1986;

Pallett, 1990; Zue et al., 1990), is fully (hand) segmented and labelled. The *training set* consists of 3,696 utterances (4,891 different words), 8 sentence utterances each (the so-called *si* and *sx* sentences) of 462 American-English speakers (326 male/136 female, from 8 major dialect regions of the U.S.). The *sx* sentences were read from a list of 450 phonetically balanced sentences selected by MIT, whereas *si* sentences were randomly selected by TI. The two other so-called *sa* sentences, which are the same for all speakers, were not used by us because they might introduce certain specific duration characteristics based on over-representation. The *test set* consists of 1,344 utterances (2,373 different words) from 168 speakers (112 male/56 female). A total of 61 different phone symbols is used in TIMIT, of which some are very rare. It is common practice to reduce this set to 50 (17 vowels and diphthongs, 5 liquids and glides, 4 nasals, 11 fricatives and affricates, 8 plosives and a flap, 3 closures, and 2 pauses including non-speech). For practical reasons, the so-called TIMITBET phonetic notation will be used in this paper. Each of these 50 phones appears at least a few hundred times. Since certain phones are very similar, the phone inventory has been further reduced to only 39 phone categories.

Fig. 1 gives the duration distribution of all 134,627 non-silent phone segments (with the 3 closure types included) in the 3,696 training utterances. For a few selected phones their individual distribution is given as well.

As observed earlier, monophone models (rather than bi- or triphone models) are used. The *total* number of states per model is determined by duration constraints and varies in our approach between 3 and 10. Some of the states are tied to one observation pdf (for detailed ways of tying, see Wang, 1996), so that each model has only three *different* pdf's. The central pdf models the steady state part of the phone and the other two model the beginning and end transitions, respectively. The smooth curves in Fig. 1 are duration pdf's calculated with the HMM transition probabilities. It can be seen that, even with well chosen numbers of states, the HMMs do not always model the duration histograms well. Results on optimisations at this level can be found in (Wang, 1994, 1995).

It is important to keep in mind that there are

Fig. 1. The bars in each panel give the normalised duration histogram (in bins of 8-ms) of hand-labelled phones in the 3,696 training utterances of the TIMIT database. The first panel shows the histogram of all the non-silence phones, whereas the other 3 panels show the histograms of 3 example phones /aa/, /s/ and /t/, together with their HMM modelled durational pdf (dark curves). The counts of the phone instances, and the number of states (n) in the linear HMM for the 3 example phones, are also given.

actually at least three types of phone sequences per utterance (and thus also for the whole database):

(1) The hand-labelled sequence, this contains 142,910 segments altogether, of which 8,283 are long pauses, leaving 134,627 real phone (and closure) segments.

(2) The lexical form, representing the ideal or norm pronunciation of each word in isolation. Concatenation of these word transcriptions to form the utterance transcription, produces a larger number of segments (154,133), of which 7,392 are pauses (one following and one preceding each of the 3,696 utterances), leaving 146,741 phone (and closure) segments. Of the total of 12,114 deleted phones, 7,272 are non-released stop bursts, and 3,237 are the closures. This is a substantial part of all deletions (86.8%).

(3) The one resulting from the actual automatic word or phone recognition.

In order to indicate how different these three representations can be, we only have to mention the

present level of phone recognition performance (Lee and Hon, 1989; Robinson, 1991; Lamel and Gauvain, 1993b; Ljolje, 1994; Young and Woodland, 1994), which is about 75% correct. Another indication of the difference between representations (1) and (2) is represented in Table 1. This table presents (a part of) a confusion matrix between (2) the phone sequence taken from the lexicon, and (1) the manual labelling. For the TIMIT training set, only 78.2% phone instances in the lexicon (120,599 out of 154,133) are completely correctly matched (c) with the hand-labelled realisations, after applying appropriate dynamic programming (DP) at symbolic level! Part of the mismatches are of course related to insertions (i , altogether 4,322) and deletions (d , altogether 15,545), the rest are substitutions (s , altogether 17,989), partly related to alternative pronunciations and word boundary effects, for an example see Fig. 2. Partly because of the greater number of similar vowel classes, the vowel matches are worse than for the other phone categories.

Table 1

Part of a matrix indicating the confusion (absolute numbers of occurrence), after DP match, between on the one hand the phone sequence taken from the lexicon (left column entry), and on the other hand the manual labelling (top row entry). The total numbers of deletions and insertions are also indicated, as well as the percentage covered per phone

	iy	ih	ix	uw	aa	ay	ow	er	y	r	l	n	s	dh	p	b	vcl	...	Del	%c
iy	3920	129	257	6				1	15										63	90.1
ih	173	3068	1833	2	1	2	1	43		1	2	2	4				2		513	54.5
ix	148	284	1686	1				13				22	1				2		345	60.8
uw	1	7	202	1907				9			1								170	81.5
aa		1	4		1860	9	1	94											17	86.8
ay	1	16	9		72	1913				1									11	93.3
ow			5	6	1		1607	2											8	93.0
er		8	51	3	1			2912		86									27	92.4
y	5	10	23	3				30	924										127	91.8
r	6	8	44		4			465	1	4504							1		735	87.6
l			11		3		2	1			4274								282	98.7
n	1	3	10		1			5			2	6652	3				2		541	96.9
s			3					3						5861					204	98.1
dh	27	6	53									12	1	2363					99	89.5
p			3					3							25863				345	98.5
b	1	5	14		1			7		2			2			2176			268	98.0
vcl	6	12	49		5			4	2	2	1	7	3	3			7079		1567	87.7
...																				
Ins	55	41	170	6	16	1	3	14	46	73	23	16	16	2	57	1	43			

This substantial discrepancy at transcription level is a complicating factor in the collection of duration statistics from the TIMIT database. For such duration statistics, not only the intended phone and its surrounding phones, but also its position in the word and in the sentence, as well as its word stress and sentence accent, are relevant. However, the TIMIT manual labelling unfortunately does not offer the actually realised sentence accent or word stress. So, the best one can do is to rely on the lexical stress given for the norm pronunciation of each word specified in the lexicon. In fact, this is the only sensible information that is available for improving performance in actual word recognition. As is found in the example sentence of Fig. 2 and in most other situations, vowels are mostly matched with (other) vowels, unless they are themselves deleted. When such a

vowel match is found, the realisation of the *transcribed* vowel is used to add to the duration statistic of that vowel, whereas the stress and location information is inherited from the norm form in the lexicon. So, in the example sentence above, the stress mark “1” (primary stress) of /ae/ is “copied” to /ix/. If the DP does not find a match between the transcribed vowel and its lexical form, that vowel instance is used with “unknown” stress and location information. Of the 45,572 vowel instances, DP finds a match for 42,912 (94.2%) of them.

Crystal and House (1982, 1988a,b, 1990) have collected somewhat similar duration statistics on two read scripts containing 592 words altogether (about 1,740 phones). These scripts were read 14 times by 32 talkers. Two sets of 7 talkers (12 male, 2 female) were selected to form a slow and a fast group (based

```

si1865 by mptf0: now      there's      nothing      left      of      me
norm:             ns n awl dh ae1 r z      n ahl th ix n l eh1 f cl t ax v m iyl ns
actual:           ns n awl dh ix1 z epi n ahl th ix n l eh1 f cl t ax v m iyl ns
(mis)match:       c c c c s d c i c c c c c c c c c c c c c c c

```

Fig. 2. Example of a DP match between the lexical form (norm) of the words in this short sentence and the actual labelling. The labels *c*, *s*, *d* and *i* indicate a correct match, a substitution, a deletion and an insertion, respectively. The number 1 connected to some of the vowels, indicates primary stress. An epenthetic closure is indicated with *epi*.

on total reading time). First the first four sentences by all 14 talkers, and later the whole script as read by 6 talkers, was hand-segmented into phones (5,636 versus 10,303). Word stress (primary and secondary stress, and unstressed) and syllables were later marked as well. The number of words, speakers and phones are smaller than in TIMIT, whereas also the same script has been read repeatedly, but all data analyses have been done with much care and in great detail, including various modelling attempts. Below, some of our results will be compared with those of Crystal and House.

In collecting duration statistics from the TIMIT database, we are going to distinguish, next to individual phones, the following categories as well as various combinations:

- short versus long vowels, including diphthongs,
- stressed (primary and secondary) versus unstressed,

- position in the word (word-final, non word-final, monosyllabic word),
- position in the sentence (utterance final versus non-final),
- vowels preceding voiced or unvoiced plosives,
- subcategories of speakers (gender; dialect),
- utterance-specific speaking rate (slow, average, fast).

The above categories are partly based on phoneticians' interest (such as the effect of *post-vocalic consonants*) but mainly on potential usefulness for improving recognition performance. This implies that only those subdivisions make sense that can be derived from the speech data of the unknown sentences after the usual recognition but before post-processing (such as *speaking rate* (Jones and Woodland, 1993) and phone positions in *word* and *sentence* (Wang, 1996)). In that respect word *stress* is already a difficult parameter, since it is not at all

Fig. 3. Effect of pre-pausal lengthening (utterance-final versus non-final, dark line versus thin line in each panel) and word stress (unstressed versus stressed, upper panels versus lower panels) on the duration (in ms) of short and long vowels (left versus right panels). The number gives the actual number of occurrences in the training set. The mean durations for the two distributions in each panel are indicated by open or filled bars, respectively.

clear that realised word stress can be consistently derived from the speech signal. That is why we have presently linked it to the stress mark in the norm pronunciation as specified in the lexicon. Let us first look at some vowel statistics (Umeda, 1975).

2.1. Vowel duration distribution affected by stressing and location

In order to get more consistent representations, we will first of all distinguish long versus short vowels. This grouping is based on the actual mean duration for each vowel under all conditions. This generally coincides with the common definition of short and long vowels (with example words):

short:	iy	(beat)	long:	ae	(bat)
	ih	(bit)		aa	(cot)
	eh	(bet)		ao	(about)
	ix	(roses)		ey	(bait)
	ax	(the)		ay	(bite)
	ah	(butt)		oy	(boy)
	uw	(boot)		aw	(bough)
	uh	(book)		ow	(boat)
				er	(bird)

Fig. 3 gives a first indication of the duration distributions for long and short vowels with word stress and position in the utterance as additional factors (Oller, 1973). Not surprisingly there is a substantial difference between *short* and *long* vowels (left-hand versus right-hand side of Fig. 3). Also the *stressed* vowels in general are substantially longer than the *unstressed* ones (lower versus upper

Table 3

Mean vowel duration (in *ms*) of TIMIT training set, for the short and long vowels separately, in stressed and unstressed syllables, in word-final, non-word-final positions and in monosyllabic words. *n* shows number of tokens

	Short vowel				Long vowel			
	unstressed		stressed		unstressed		stressed	
	<i>ms</i>	<i>n</i>	<i>ms</i>	<i>n</i>	<i>ms</i>	<i>n</i>	<i>ms</i>	<i>n</i>
word-final	71	5022	113	1268	97	1583	148	1753
non-word-final	55	5858	82	4728	89	1679	123	5402
monosyllabic word	52	3085	86	8170	86	288	138	6736

part of Fig. 3), whereas also *utterance-final lengthening* is a consistent effect, at least for the stressed vowels (dark line versus thin line in lower panels of Fig. 3). All three effects are potentially useful to improve automatic speech recognition performance. These main effects are summarised in Table 2 together with data from (Crystal and House, 1988b) in terms of mean duration and number of tokens involved. Crystal and House distinguish 15 monophthong vowels (and 3 diphthongs), of which 7 are long and 7 are short, however this last category then also contains reduced vowels that can only occur in unstressed syllables, so they prefer a short vowel category of only 4 types /IeAu/. Although neither the vowel sets nor the number of tokens are really comparable, it is clear that both data sets exemplify the lengthening effect of stress and utterance-final (or pre-pausal) position.

Table 3 presents similar subdivisions of the vowel data, but this time separated out in terms of the

Table 2

Mean vowel duration in *ms*, for the short and long vowels separately, in stressed and unstressed position, and in utterance-final (or pre-pausal) and non-final positions. Both the TIMIT data as well as data extracted from (Crystal and House, 1988b) are presented. Each time the number of tokens involved (*n*) is indicated as well

	TIMIT short V (8)		Crystal and House short V (4)		TIMIT long V (9)		Crystal and House long V (7)	
	<i>ms</i>	<i>n</i>	<i>ms</i>	<i>n</i>	<i>ms</i>	<i>n</i>	<i>ms</i>	<i>n</i>
unstressed	60	13,965	56	842	93	3,550	84	286
stressed	87	14,166	93	601	133	13,891	151	1,411
utt. final unstressed	78	1,199	81	39	109	408	110	7
utt. final stressed	142	954	147	78	177	1,135	202	125
utt. non-final unstressed	59	12,766	56	727	90	3,142	77	224
utt. non-final stressed	83	13,212	85	253	129	12,756	134	628

position of the vowel within a word (word-final versus non-final), whereas the duration mean for vowels in mono-syllabic words is given separately. The distinction between stressed and unstressed is still very apparent, however, the effect of the position within the word is not very consistent at all, perhaps apart from the somewhat longer duration for short vowels in word-final positions. A partial explanation for these marginal effects is of course that word boundaries frequently are not actually realised in read speech.

2.2. Effect of post-vocalic plosives on vowel duration

Most phonetic handbooks tell us that in languages like English (e.g. Peterson and Lehiste, 1960; Chen, 1970) and Dutch (Nooteboom, 1970) a vowel preceding a voiced plosive is generally longer than the same vowel preceding an unvoiced plosive, and actually the reverse seems to be true for the closure time. Since plosive identification is an inherently difficult task, and since the (lack of) voiced/unvoiced distinction adds to that difficulty, such additional duration features could be very helpful to improve plosive identification. Also in rule-based synthesis these vowel- and closure-duration features are successfully applied to improve quality and intelligibility (Klatt, 1987; Van Heuven and Pols, 1993). Van Santen (1992) presents very nice data for word-penultimate stressed vowel duration as a function of the post-vocalic consonant for a single male speaker in a 71-minutes database of read sentences (containing 2,162 sentences, 13,048 words and 18,046 vowel segments). Differences in vowel duration of about 100 ms were found for voiced versus voiceless plosives.

We wondered whether similarly consistent effects could be found for a much less homogeneous database of many different speakers such as TIMIT.

Also given Van Santen's experience, we started to limit ourselves to stressed vowels only. It can nevertheless be seen in Fig. 4 that the distributions of vowels followed by voiced and unvoiced plosives are very similar indeed! Only the tails of the distributions give some indication of a lengthening effect for voiced plosives.

Crystal and House (1988b) also study the *lengthening-before-voicing effect* for word-final obstruents, and conclude that the data are rather sparse and most of the time non-consistent. Also in (Crystal and House, 1988a) this phenomenon is analysed and there it is shown that the lengthening effect is largest for prepausal word-final consonants, especially plosives.

2.3. Effect of speaking rate on vowel duration

One does not even need a phonetics handbook to realise that, on average, fast-rate speech will have shorter phone durations than slow- or normal-rate speech (e.g. Van Son, 1993; Eefting, 1991). So far hardly any ASR system takes that systematic effect into account. In the 1994 Benchmark test for the ARPA Spoken Language Program clear evidence was presented (Pallett et al., 1995) that speakers with a high speaking rate (in terms of number of words per minute) almost unanimously showed a higher word error rate for all 20 systems that participated in the so-called baseline Hub1 C1 test. This certainly makes it interesting to see whether rate-specific duration information could help in improving recognition

Fig. 4. Effect of voicing of the post-vocalic plosives on vowel duration, for short and long vowels separately.

Fig. 5. Histogram of the relative speaking rate (for a definition, see text) for all 3,696 utterances of the TIMIT training set. The dark curve represents the utterance-averaged phone duration in ms (against right axis) in each histogram bin.

performance. Jones and Woodland (1993) already showed that speaking rate can be measured at recognition time and that it can be used to improve the speech recognition behaviour through e.g. post processing.

Although speaking rate may indeed vary within an utterance, we here opt for the average rate measured at the utterance level. The *normalised phone duration* (Jones and Woodland, 1993) is chosen as the basic unit, this way we correct for the intrinsic

cally long or short duration of specific phones. The *relative speaking rate* of an utterance is indicated by the average normalised phone duration in that utterance. (Actually this is similar to the *reciprocal* of speaking rate.) Fig. 5 gives a histogram distribution of the relative speaking rate for all 3,696 utterances. It is similar to the usual duration pdf of most phones, and has a binomial-like distribution. For comparison, the utterance-averaged absolute phone durations in the corresponding histogram bins are also shown. It

Fig. 6. Histogram of the actual duration of two example phones /f/ and /iy/, separate for the three rate groups: fast (thin line), average (thick line) and slow (broken line), respectively. The number of tokens in each group is given, whereas the mean duration in each group is indicated by a bar.

can be seen that the averaged absolute phone duration has a near-linear relation with the relative utterance speaking rate, this is particularly true in the middle region, where counts are large. The irregularities in the periphery are basically due to the fact that these peripheral data represent relatively few utterances for which the intrinsic phone duration may vary a lot.

Based on this rate measure per utterance, the whole training set is divided in three groups (Jones and Woodland, 1993) of 1,232 utterances each, indicated as fast, average and slow rate. It has to be kept in mind that each speaker may have utterances belonging to different rate groups. After such grouping based on *utterance* rate, we can have a look at the *phone* duration histograms per group, this is exemplified in Fig. 6 for two phones. As these two

examples already show, there is a general tendency for most phones in an utterance to be short if the utterance rate is fast, and long if the rate is slow. These promising results seem to be potentially useful to improve recognition performance. Crystal and House (1982) also come to the conclusion that average phone duration varies with the speaking rate, although there seems to be a tendency for the most prolongable sounds to be lengthened most, with the pause times at the top.

3. Hand-labelling versus automatic segmentation

The duration statistics derived in the preceding section, were all based on the hand-labelled TIMIT material. This could be the basis for an initialisation

Fig. 7. For one-tenth of the segments in the TIMIT database, the duration in ms of the manually-labelled segments (horizontal axis) is presented against the automatically derived segment durations in ms (vertical axis) via Viterbi search in the segmentation mode.

process to add duration-specific knowledge to a recognition system. However, the HMM-based recogniser will never be able to recognise and reproduce all these phone segments with their correct duration. So, our next step is to see how close the phone durations after Viterbi search match the hand-labelled segments. In Fig. 7, for one-tenth of the total data set of 134,627 points (this limitation is for graphical and memory reasons only) the duration of the manually labelled segments is plotted against that of the segments after Viterbi search. The ‘‘Viterbi’’-duration can of course only be expressed with a resolution of 8 ms (frame duration), which explains the discrete character along that dimension. A first deviation between both durations is the empty start region (bottom) along the Viterbi duration, which is caused by a minimum duration because of the actual number of non-skippable states per phone model. It can be seen in Fig. 7 that the number of outliers can be quite serious. If a 20-ms deviation around the diagonal is used, then some 85% of all data points fall in that region. The TIMIT database has been used repeatedly to test automatic segmentation procedures (Brugnara et al., 1993; Ljolje and Riley, 1991; Vorstermans et al., 1995) with comparable results. This still leaves 15% for which the phone duration, with full knowledge of the phone sequence, deviates more than 20 ms from the actual duration. For all 50 phones together the figure shows a rather even distribution around the diagonal, however for individual phones the deviation can be much more irregular. In any initial training based on the hand-labelled data, such deviations cannot be taken into account and perhaps should be taken care of in some later recognition step.

4. Analysis of variance

By studying each effect individually, as done above, a much better view on the significance of various parameters is obtained. However, a quantitative comparison is not very well possible along those lines. Inspired by Sun and Deng (1995) we have therefore developed a hierarchically structured Analysis of Variance (ANOVA), that in principle should allow the analysis of the contribution of various identifiable factors to the overall durational variability

in the TIMIT database. Sun and Deng performed such an analysis for the spectral variability in the form of Mel-Frequency Cepstral Coefficients (MFCC).

There is no straightforward ANOVA that can solve this problem. The complications lie at various levels:

- the inability to model this complex factorial design in a fully satisfactory way,
- the sheer size of the data, which leads to memory problems,
- the problem of nested factors,
- empty cells and singletons,
- the ordering problem.

We intend to analyse the variation in phone duration as explained by the following 11 factors:

<i>R</i>	speaking rate
<i>Cl</i>	broad phonetic class
<i>Ph</i>	phone
<i>Pt</i>	phone in context (only vowel)
<i>S</i>	stress
<i>Lw</i>	location of syllable in word
<i>Lu</i>	location of syllable in utterance
<i>G</i>	gender of speaker
<i>Dr</i>	dialect region of speaker
<i>Sp</i>	speaker
<i>Sg</i>	phone segment

The speaking rate *R* has 3 levels (slow, average, fast) and is measured per utterance as specified in Section 2.3. The broad phonetic class *Cl* has 8 levels (plosives, affricates, fricatives, nasals, semi-vowels and glides, long vowels, short vowels, and pauses), *Ph* are the 50 reduced TIMIT phones. *Pt* only considers the right-hand context of vowels as being either /p,t,k/, /b,d,g/, or the rest. For all other phones *Pt* has only one level. Such a restriction for factors also applies to *S*, *Lw* and *Lu* because stress and location of syllables in words and in utterances, only applies to vowels. Stress *S* has 3 levels (primary and secondary stress, and unstressed). *Lu* has also three levels (last, penultimate, and rest position within the utterance), whereas *Lw* has one more level than *Lu*, namely that of vowels in monosyllabic words. *G* (male, female), *Dr* (8 dialect regions) and *Sp* (462

speakers) are obvious and are all three specified in the TIMIT database. *Sg*, finally, represents the actual instances of the phone segment durations of a given phone. The number of observations at that level varies, but is on average less than 1.

The relations between all the levels in all the $K = 11$ factors can be shown in a tree, part of which is represented in Fig. 8.

In order to calculate the variation in each factor in terms of sum of squares (*SS*), a fully nested model for any two adjacent factors is used. The total *SS* can be decomposed into a sum of *SS*s of each factor. At each factor the variation is decomposed into *SS*_{between} and *SS*_{within}, and the latter is further decomposed into *SS*_{between} and *SS*_{within} at a lower factor. For more details see (Wang, 1996).

A special problem encountered in calculating the *SS* terms is the computer memory problem caused by the large number of nodes, each requiring an accumulator and a counter to be allocated. Down to the factor *Dr*, the total number of nodes is already 67,680, given the number of levels in all the factors above. Actually this number 67,680 is achieved in the following way:

$$\begin{aligned}
 &3 \times [33 + 17 \times 3 \times 3 \times \\
 &R \quad \text{non-V} \quad V \quad Pt \quad S \\
 &(1 + 3 + 2 + 3)] \times 2 \times 8 = 67,680 \\
 &Lu \times Lw (\max 4 \times 3) \quad G \quad Dr
 \end{aligned}$$

Plain implementation down to the next factor *Sp* would require a further multiplication with about 29, being the average number of speakers at this level. In order to avoid this, a special algorithm is designed which allocates some temporal accumulators and counters for a single speaker, after which the result is put into other global accumulators and counters once all the data from one speaker are used up. Then the temporal accumulators and counters are reset for use by the data of the next speaker.

Table 4

Percentages of variations in terms of *SS* in the 11 factors calculated from the TIMIT training set. The upper section shows the results of a factor ordering with the "speaker factors" *G*, *Dr* and *Sp* after the "phonetic factors" from *Cl* down to *Lu*. The lower section shows the results of swapping these two groups of factors. The last column shows the loss of calculated *SS* due to singleton cells

<i>R</i>	<i>Cl</i>	<i>Ph</i>	<i>Pt</i>	<i>S</i>	<i>Lw</i>	<i>Lu</i>	<i>G</i>	<i>Dr</i>	<i>Sp</i>	<i>Sg</i>	loss
2.31	15.18	26.06	0.21	0.39	0.77	0.91	0.30	2.16	16.02	34.93	0.75
<i>R</i>	<i>G</i>	<i>Dr</i>	<i>Sp</i>	<i>Cl</i>	<i>Ph</i>	<i>Pt</i>	<i>S</i>	<i>Lw</i>	<i>Lu</i>	<i>Sg</i>	loss
2.33	0.01	0.03	0.59	19.57	37.46	1.05	1.21	1.52	0.57	34.93	0.74

Fig. 8. Part of the tree for ANOVA for the 11 indicated factors. Also the number of possible splits of levels per factor, as well as the cumulative numbers of levels, are indicated.

The result of calculating the *SS* terms in percentages in each of the subsequent 11 factors is shown in Table 4 for two different orderings of the factors. The most tangent phenomenon seen from these percentages is, that when *Sp* (and *G* and *Dr*) are put before *Cl* and *Ph* (see lower section of Table 4), the

Table 5

From top to bottom the percentages of the variations per factor are presented, while omitting more and more factors from the analysis. The upper row contains all the factors, whereas the lowest row only contains two factors. The order of the factors for all analyses are the same (from left to right). Losses due to singleton cells are shown in the last column. For clarity, entries with the same value (within two decimals precision) as the entry below are shown with a downward arrow

<i>R</i>	<i>Cl</i>	<i>Ph</i>	<i>Pt</i>	<i>S</i>	<i>Lw</i>	<i>Lu</i>	<i>G</i>	<i>Dr</i>	<i>Sp</i>	<i>Sg</i>	loss
2.31	15.18	26.06	0.21	0.39	0.77	0.91	0.30	2.16	16.02	34.93	0.75
↓	↓	↓	↓	↓	↓	↓	0.30		18.19	↓	↓
2.31	15.18	26.06	↓	0.39	↓	0.91			18.48	34.93	0.75
2.33	15.52	26.20	↓	0.41	0.77				18.98	35.57	0.00
↓	↓	↓		0.41					18.11	37.21	↓
↓	↓	↓	0.21						17.28	38.54	↓
↓	↓	26.20							16.41	39.54	↓
↓	15.52								5.00	77.15	↓
2.33									0.62	97.05	↓
									1.64	98.36	0.00

variation in *Sp* is rather small, while when *Sp* (and *G* and *Dr*) are put after the splitting of the data by *Cl* and *Ph*, *Sp* explains a much larger percentage of variation. This is further confirmed by the data in Table 5, in which the initial order of the factors is kept while subsequently omitting factors. When *Cl* and *Ph* are omitted, *Sp* then only explains a very small percentage any more, whereas *Sg* takes the major share.

In principle, when a factor is omitted from the analysis, other factors should account for more variation than otherwise. However, the above analysis shows that this is not always true for the factor *Sp*. Going from bottom up in Table 5, the *SS* of *Sp* starts to increase (from 0.62% to 5.00% and then to 16.41%) from the third row on when the “phonetic factors” *Cl*, *Ph* and *Pt* become included. Such a phenomenon, that might be typical for a database like TIMIT with so many different speakers, can better be understood by looking at some example histograms in Fig. 9.

In the right-most column it can be seen that, when the whole speech data is not split into phones, the durational histogram of one speaker tends to be the same as that of any other speaker, and consequently the variation across speakers is rather small. This is also clear from the very narrow distribution per speaker of the mean duration, as given in the lower right-most histogram in Fig. 9. However, the strongly varying distributions for each individual phone over all the speakers, shown at the bottom row, indicates

that, if the data is split by phones, there is a large durational variation over speakers. So, averaging over phones, strongly reduces the speaker variability, whereas keeping the phone variability in the data, also substantially increases the speaker variability (see histograms in the bottom row of Fig. 9, where the distributions are relatively wider for the individual phones than for “All” phones). If one looks again at the numbers under *Sp* in the upper and lower section of Table 4, one will also see this process illustrated.

The above discussion indicates a general point to be noticed in such an analysis of variation: No factor should be omitted (such as *Ph*) that actually plays an essential role in the data. This also means that the order of factors to be analysed as given in the lower part of Table 4, should be considered as being inappropriate. Given an appropriate general ordering of factors, certain pairs of factors can still be switched, which may have small consequences on the explained variation. This effect was also reported by Sun and Deng (1995).

From the individual cells in Fig. 9 it can furthermore be seen that, at that level of one phone of one speaker, the number of observations is already very small. This also complicates our ANOVA analysis, because of single observations or empty cells. Those, of course, cannot contribute to the *SS*.

For the moment we suppose that the upper row of percentage numbers given in Table 4, properly reflects the importance of various factors, in terms of

Fig. 9. Duration histograms of 8 randomly selected speakers (one from each dialect region) and of 8 phones (one from each phonetic class). The right-most column shows the duration histogram for each speaker over all 50 phones. The bottom row shows the histogram of the duration *means* across all 462 speakers, for that phone, whereas the right-most histogram on the bottom row gives the histogram of the means of all the speakers, each averaged over "all" his phones. For maximum visibility, each histogram in the right column and in the bottom row is scaled to its maximum height, whereas all other histograms are scaled to one unique maximum height across all of them.

variation explained. However, this puts us in a somewhat uneasy position, since several factors that appeared to have a consistent effect in the distributions presented in Section 2, such as speaking rate R and stress S , here are only responsible for 2.3% and 0.39%, respectively! Position within the word (Lw) and within the utterance (Lu) seem to have equal (although small) importance in the present analysis, whereas earlier on in Section 2 Lw seemed to be much less important than Lu . A possible explanation

for these discrepancies is that any ANOVA can only present overall effects per factor, whereas for instance in Fig. 3 the effect of stress is presented for long and short vowels separately. In a regular ANOVA such effects shows up in interaction terms, which are not available in the present analysis.

In terms of phone duration the variability over male and female speakers (G) and over dialect regions (Dr) actually appears to be rather small, only 0.30% and 2.16%, respectively. What the numbers in

the upper part of Table 4 do show us, is the amount of variability caused by the 8 broad phonetic classes (*Cl*) and by the phones themselves (*Ph*), this adds up to 41.24%. Despite the small number of average observations (less than 1) at the level of the phone segment (*Sg*), these few observations themselves apparently vary a lot, since 34.93% of the total variation goes there. The only fully hand-labelled speech database available to us (TIMIT), is apparently limited in size at that level.

5. Discussion

The analysis and description of the durational characteristics of the speech material in the TIMIT database offers several interesting perspectives. Both the duration distribution curves, as well as the analysis of variance, show different characteristics of segmental duration. But the general picture is not very clear and consistent yet, not to speak about the potential usefulness of the present data for improving recognition performance.

It strikes us that many of the very consistent facts that Van Santen (1992) could demonstrate in his large speech database for a single male speaker (enabling him to design a 7-terms sums-of-products model) did not show up in the TIMIT database. It is not straightforward to develop a similar model based on the TIMIT-data. It made us aware once again that the combination of “real speech” (although still read from paper), and “non-professional speakers” under semi-controlled conditions introduce a lot of variation. As a consequence, only certain, perhaps very global duration features may be beneficial for improving recognition performance. We continue the study how to use duration knowledge in HMM-based recognition, and we hope that the methodology that we have applied so far, may show to be productive for other “knowledge sources”, such as the pitch contour, as well.

Meanwhile, in our research project about duration modelling, duration models are designed based on a subset of all the contextual factors discussed in this work. These models are integrated into a post-processing part of the recogniser using *N*-best sentence alternatives. This post-processing is based on a

rescoring of these *N*-nest hypotheses (cf. Wang, 1996).

References

- H. Bourlard, H. Hermansky and N. Morgan (1996), “Towards increasing speech recognition error rates”, *Speech Communication*, Vol. 18, No. 3, pp. 205–231.
- F. Brugnara, D. Falavigna and M. Omologo (1993), “Automatic segmentation and labeling of speech based on Hidden Markov Models”, *Speech Communication*, Vol. 12, No. 4, pp. 357–370.
- D. Byrd (1992), “Sex, dialects, and reduction”, *Proc. ICSLP’92*, Banff, Vol. 1, pp. 827–830.
- M. Chen (1970), “Vowel length variation as a function of the voicing of the consonant environment”, *Phonetica*, Vol. 22, pp. 129–159.
- T.H. Crystal and A.S. House (1982), “Segmental durations in connected-speech signals: Preliminary results”, *J. Acoust. Soc. Amer.*, Vol. 72, pp. 705–716.
- T.H. Crystal and A.S. House (1988a), “Segmental durations in connected-speech signals: Current results”, *J. Acoust. Soc. Amer.*, Vol. 83, pp. 1553–1573.
- T.H. Crystal and A.S. House (1988b), “Segmental durations in connected-speech signals: Syllabic stress”, *J. Acoust. Soc. Amer.*, Vol. 83, pp. 1574–1585.
- T.H. Crystal and A.S. House (1990), “Articulation rate and the duration of syllables and stress groups in connected speech”, *J. Acoust. Soc. Amer.*, Vol. 88, pp. 101–112.
- W. Eefting (1991), Timing in talking. Tempo variation in production and perception, Ph.D. Thesis, University of Utrecht.
- J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallatt and N.L. Dahlgren (1993), The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM, NTIS order number PB01-100354, now available from LDC.
- J.L. Gauvain, L.F. Lamel, G. Adda and M. Adda-Decker (1994), “Speaker-independent continuous speech dictation”, *Speech Communication*, Vol. 15, Nos. 1–2, pp. 21–37.
- M.M. Hochberg and H.F. Silverman (1993), “Constraining the duration variance in HMM-based connected-speech recognition”, *Proc. Eurospeech’93*, Berlin, Vol. 1, pp. 323–326.
- M. Jones and P.C. Woodland (1993), “Using relative duration in large vocabulary speech recognition”, *Proc. Eurospeech’93*, Berlin, Vol. 1, pp. 311–314.
- P.A. Keating, D. Byrd, E. Flemming and Y. Todaka (1994), “Phonetic analyses of word and segment variation using the TIMIT corpus of American English”, *Speech Communication*, Vol. 14, No. 2, pp. 131–142.
- D.H. Klatt (1987), “Review of text-to-speech conversion for English”, *J. Acoust. Soc. Amer.*, Vol. 82, pp. 737–793.
- L.F. Lamel and J.L. Gauvain (1993a), “Identifying non-linguistic speech features”, *Proc. Eurospeech’93*, Berlin, Vol. 1, pp. 23–30.

- L.F. Lamel and J.L. Gauvain (1993b), "High performance speaker-independent phone recognition using CDHMM", *Proc. Eurospeech'93*, Berlin, Vol. 1, pp. 121–124.
- L. Lamel, R. Kassel and S. Seneff (1986), "Speech database development: Design and analysis of the acoustic-phonetic corpus", *Proc. DARPA Speech Recognition Workshop*, pp. 100–109.
- K.-F. Lee and H.-W. Hon (1989), "Speaker-independent phone recognition using Hidden Markov Models", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-37, pp. 1641–1648.
- S.E. Levinson (1986), "Continuously variable duration hidden Markov models for automatic speech recognition", *Computer Speech and Language*, Vol. 1, pp. 29–45.
- A. Ljolje (1994), "High accuracy phone recognition using context clustering and quasi-triphone models", *Computer Speech and Language*, Vol. 8, pp. 129–151.
- A. Ljolje and M.D. Riley (1991), "Automatic segmentation and labeling of speech", *Proc. Internat. Conf. Acoust. Speech Signal Process.-91*, Toronto, pp. 473–476.
- S.G. Nooteboom (1970), Production and perception of vowel duration, Ph.D. Thesis, University of Utrecht.
- O.K. Oller (1973), "The effect of position in utterance on speech segment duration in English", *J. Acoust. Soc. Amer.*, Vol. 54, pp. 1235–1247.
- D. Pallett (1990), "Speech corpora and performance assessment in the DARPA SLS program", *Proc. ICSLP'90*, Kobe, pp. 24.3.1–24.3.4.
- D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, B.A. Lund, A. Martin and M.A. Przybocki (1995), "1994 Benchmark tests for the ARPA Spoken Language Program", *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 5–36.
- G.E. Peterson and I. Lehiste (1960), "Duration of syllable nuclei in English", *J. Acoust. Soc. Amer.*, Vol. 32, pp. 693–703.
- T. Robinson (1991), Several improvements to a recurrent error propagation network phone recognition system, Technical report, Cambridge Univ., CUED/F-INFENG/TR.82.
- D.X. Sun and L. Deng (1995), "Analysis of acoustic-phonetic variations in fluent speech using TIMIT", *Proc. Internat. Conf. Acoust. Speech Signal Process.-95*, Detroit, pp. 201–204.
- N. Umeda (1975), "Vowel duration in American English", *J. Acoust. Soc. Amer.*, Vol. 58, pp. 434–445.
- V.J. Van Heuven and L.C.W. Pols, Eds. (1993), *Analysis and Synthesis of Speech. Strategic Research towards high-quality Text-to-speech Generation* (Mouton/De Gruyter, Berlin).
- J.P.H. Van Santen (1992), "Contextual effects on vowel duration", *Speech Communication*, Vol. 11, No. 6, pp. 513–546.
- R.J.J.H. Van Son (1993), *Spectro-temporal features of vowel segments*, Studies in Language and Language Use 3, Ph.D. thesis, University of Amsterdam.
- A. Vorstermans, J.P. Martens and B. Van Coile (1995), "Fast automatic segmentation and labeling: Results on TIMIT and EUROM0", *Proc. Eurospeech'95*, Madrid, Vol. 2, pp. 1397–1400.
- X. Wang (1994), "Durationally constrained training of HMM without explicit state durational pdf", *Proc. Inst. Phonetic Sciences Univ. of Amsterdam*, Vol. 18, pp. 111–130.
- X. Wang (1995), "Durational modelling in HMM-based speech recognition: Towards a justified measure", in: A.J. Rubio and J.M. López, Eds., *Speech Recognition and Coding. New Advances and Trends* (Springer, Berlin), pp. 128–131.
- X. Wang (1996), Duration modelling in HMM-based speech recognition, Ph.D. thesis, University of Amsterdam.
- P.C. Woodland and S.J. Young (1993), "The HTK tied-state continuous speech recogniser", *Proc. Eurospeech'93*, Berlin, Vol. 3, pp. 2207–2210.
- S.J. Young (1992), HTK Version 1.4: User, reference and programmer manual, Cambridge University Engineering Department, Speech Group, September 1992.
- S.J. Young and P.C. Woodland (1993), "The use of state tying in continuous speech recognition", *Proc. Eurospeech'93*, Berlin, Vol. 3, pp. 2203–2206.
- S.J. Young and P.C. Woodland (1994), "State clustering in hidden Markov model-based continuous speech recognition", *Computer Speech and Language*, Vol. 8, pp. 369–383.
- V. Zue, S. Seneff and J. Glass (1990), "Speech database development at MIT: TIMIT and beyond", *Speech Communication*, Vol. 9, No. 4, pp. 351–356.