

PREDICTING SEGMENTAL DURATIONS FOR DUTCH USING THE SUMS-OF-PRODUCTS APPROACH

Esther Klabbers and Jan van Santen

IPO, Center for User-System Interaction, Eindhoven, The Netherlands

E.A.M.Klabbers@tue.nl

Center for Spoken Language Understanding, Oregon Graduate Institute, Portland OR

vansanten@ece.ogi.edu

ABSTRACT

This paper presents the results of a duration study performed for Dutch using the sums-of-products approach [5]. With a relatively small corpus of 297 sentences, a duration model could be constructed with an RMSE of 27 ms, which compares well to similar models for English, French and German. In an evaluation study the predicted durations of the duration model were compared to those predicted by a rule-based duration model.

1. INTRODUCTION

This paper reports on the development of a new duration model for Dutch, which replaces the traditional sequential rule-based model in IPO's diphone synthesis system Calipso [8]. A detailed description of this study can be found in [3]. The old model gave unsatisfactory results, probably because interactions among factors were not sufficiently modelled and some important higher-level prosodic and positional factors were not taken into account. Moreover, this system was not specifically modelled after the speaker of the diphones.

The sums-of-products approach was chosen because it takes advantage of the fact that most interactions are directionally invariant, which allows describing these interactions with equations consisting of sums and products. Phonemes that are affected similarly by the various factors are grouped into subclasses. The decisions concerning this grouping are based on exploratory data analysis and phonetic/phonological literature. For each subclass of segments, a separate sums-of-products model is trained. Additional reasons for using this approach are that it requires a fairly small amount of training data to construct a reliable model, the resulting model is generalisable to many text types, and it allows a thorough analysis of the durational characteristics of a speaker. The sums-of-products

The work by Klabbers is part of the Priority Programma Language and Speech Technology (TST), sponsored by NWO (The Netherlands Organization for Scientific Research).

approach has successfully been applied to many languages, including American English, French, German, Italian, Spanish, and Japanese [7].

2. THE DURATION MODEL

The duration model is created in a number of steps. First, a large text corpus is obtained. Our analysis of segmental durations in natural speech is based on a text corpus of approximately 27,000 sentences from public broadcast news and childrens' news transcriptions¹. From this text corpus all information relevant for duration prediction is computed. To this end, the corpus was phonetically transcribed. For all segments in the corpus, feature vectors were computed, including:

1. Identity of the current segment,
2. Phoneme class of the preceding and following segment,
3. Stress and accent,
4. Left position of the segment in the syllable, of the syllable in the word, of the word in the phrase and of the phrase in the sentence,
5. Right position of the segment in the syllable, of the syllable in the word, of the word in the phrase and of the phrase in the sentence.

Then, a greedy algorithm [6] was used to find the smallest subset that covers all combinations of feature vectors in the corpus. The program selected 297 sentences containing 16,775 segments. The corpus was then recorded and manually segmented, phoneme by phoneme.

The segments were then divided into sub-classes, such that within each sub-class, the cases are similarly affected. The grouping is based on phonetic/phonological knowledge and exploratory data analysis. The resulting category tree for Dutch is visualised in Figure 1.

¹We acknowledge the Institute for Dutch Lexicology for providing us with the text database.

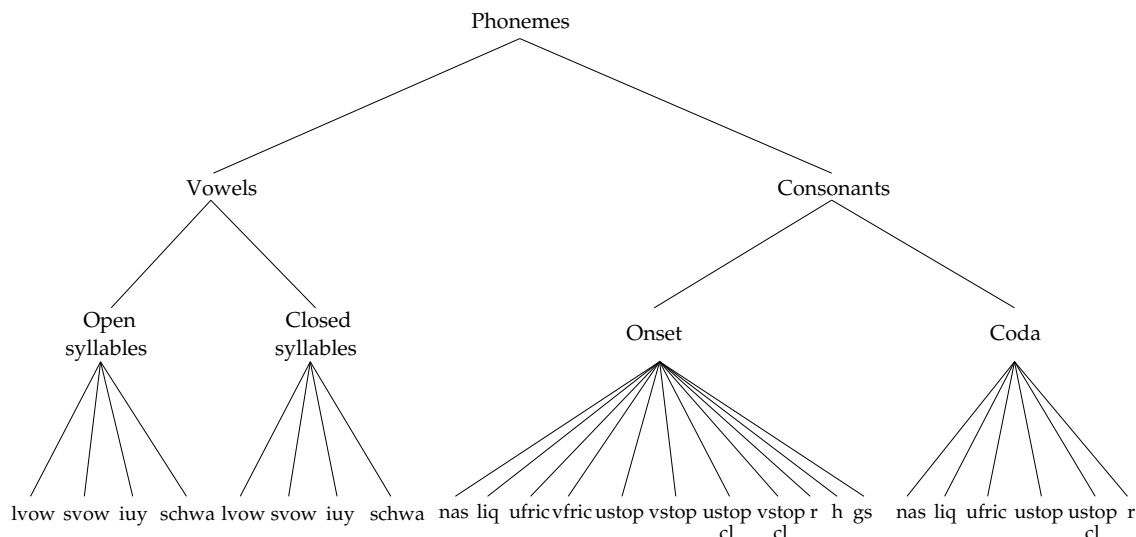


Figure 1: Category tree of the Dutch duration system; lvow/svow = long/short vowels, iuy = /i/, /u/ and /y/, nas = nasals, liq = liquids/glides, ufric/ustop = unvoiced fricatives/stop bursts, vfric/vstop = voiced fricatives/stop bursts, ustop-cl/vstop-cl = unvoiced/voiced stop closures, gs = glottal stop.

For each of the leaves in the tree, a sums-of-products model was trained that predicts the segmental durations in each sub-class in all possible contexts. We restricted ourselves to the use of purely multiplicative models. The resulting duration model has a RMSE (root-mean squared error between observed and predicted duration) of 27 ms, which compares well to similar duration models for French and German [7].

3. RESULTS

3.1. Vowels

Table 1 shows the estimated segmental durations for the different vowel classes in the database. It was not necessary to distinguish vowels within a class. Corrected means were used to account for the fact that the database was not balanced. As such, results are similar to the ones that would have been obtained if a balanced corpus was used.

Syllable type	diph	lvow	iuy	svow	schwa
Open	140	128	81	78	57
Closed	120	107	84	68	50

Table 1: Corrected means (in ms) for vowels in open and closed syllables; diph = diphthongs, lvow = long vowels, iuy = /i/, /u/, /y/, svow = short vowels.

The *previous segment* has a small effect on vowels. They are generally longer after sonorants than after obstruents. The *next segment* has a larger impact. The most obvi-

ous finding is that /i/, /u/, and /y/ are lengthened before /r/. This confirms findings by [4]. As can be seen in Figure 2, the effect persists not only in closed syllables, but also in open syllables, where the vowel is the last segment in the syllable and the /r/ is in the onset of the next syllable. In Dutch codas two allophones of /r/ are allowed, either the fricative-like /r/ or the velar /R/. Our speaker produced both, which makes comparison possible. Lengthening is more extreme before /r/ than /R/. Other vowels are also lengthened before /r/, but shortened before /R/. An interaction was found between the factors *stress* and *accent*, confirming results by [1]. This was dealt with by combining these two factors into a single one with four levels. The effect of this factor is consistent in all vowels. They are always shorter in unstressed, unaccented position than in unstressed accented position and also always shorter in stressed unaccented position than in stressed accented position. It is apparent that in each of these four conditions, long vowels are longer in duration than short vowels and /i/, /u/ and /y/. The effect of *left position* is small. A distinction between word-initial and non-initial vowels is maintained, since the first context consistently produces slightly longer vowels. The effect of *right position* is large. A three-way distinction remains between non-final, word-final and phrase-final syllables, with increasing durations. The effect is somewhat smaller in closed syllables, where a consonant follows the vowel in the same syllable.

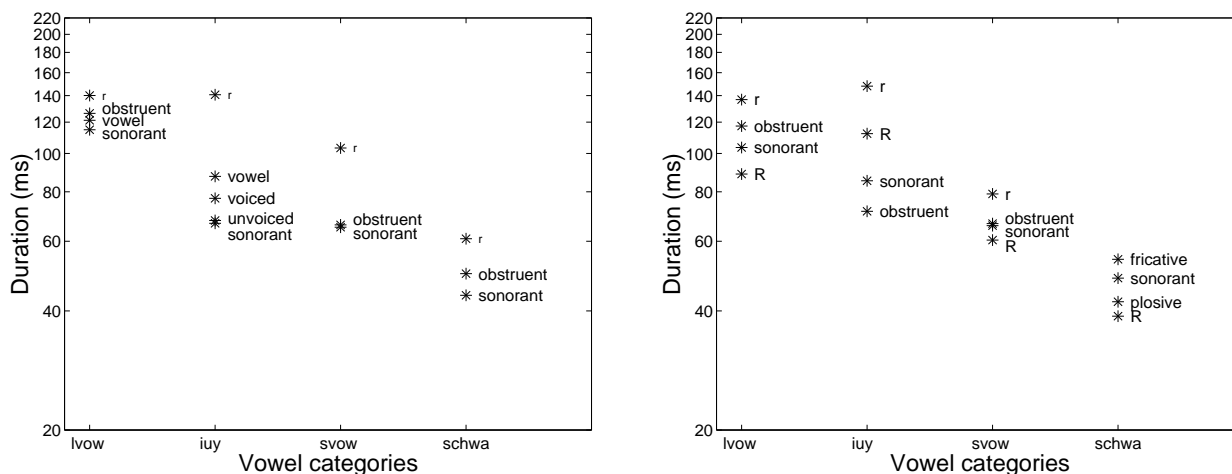


Figure 2: Effect of next segment on vowels in open (left) and closed (right) syllables.

Consonant position	m	n	N	J	j	l	w	f	s	x	S	c	v	z	Z	k	p	t	b	d	g	r	Q	h
Onset	60	47	61	109	65	52	55	93	104	91	131	123	90	80	102	82	81	79	52	43	59	36	40	67
Coda	86	64	80	89	68	66	39	86	105	85	130	-	-	-	-	72	81	81	-	-	-	40	-	-

Table 2: Corrected means (in ms) for consonants in onsets and codas.

3.2. Consonants

Table 2 shows the durations for the consonants in onsets and codas in the database. In contrast to vowels, they are not grouped together according to their phoneme class. In general, obstruents are longer than nasals, which in turn are longer than glides and liquids. Voiceless obstruents are longer than voiced ones. These findings are in line with [9], who found an increasing duration as sonority decreases.

For consonants it is more difficult to generalise the effects of the *previous* and *next segment* to all consonant categories. There is much variation due to the different distributional properties of the consonants. In general, consonants in onsets are longest when preceded by a sonorant or vowel and shortest when preceded by an obstruent. When they occur in a cluster of two or three consonants shortening applies. In Dutch codas, all obstruents are devoiced. The effect of the next segment is larger than that of the previous segment. The consonants are longest before vowels or sonorants. Here too, shortening is applied when the consonants occur in a complex cluster.

The effect of *stress and accent* varies greatly among consonant categories. The largest effect can be observed in /h/, fricatives and stop closures. On stop bursts this factor has only a small effect. For onsets, the effect of *left position* is small but consistent. Consonants are slightly longer in word-initial syllables than in non-initial syllables.

For codas this effect is negligible. The effect of *right position* is small for consonants in onsets, although they are slightly longer in word-final syllables than in non-final syllables. This effect is of course considerable for consonants in codas. We see not only phrase-final lengthening, but also word-final lengthening.

4. EVALUATION

In this section the performance of the new sums-of-products duration model (NDM) will be compared to the old rule-based duration model.

4.1. Subjective evaluation

To compare the performance of the NDM to the ODM, a paired comparison listening experiment was performed. The material for the experiment consisted of 20 sentences (1493 phonemes) taken from news reports on Teletext. Each sentence was generated twice with the same diphone system, once using the NDM and once using the ODM. The experiment was presented over the internet. Listeners had to push a button to listen to a sentence pair and had to express their preference on a five-point scale ranging from -2 (strong preference for the first sentence over the second) to 2 (strong preference for the second sentence over the first). The sentences and conditions were presented in random or-

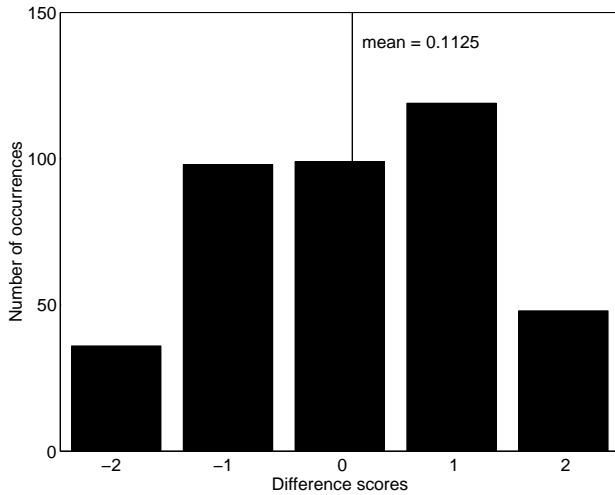


Figure 3: Results from the preferential choice experiment; -2 = strong preference for ODM over NDM, 2 = strong preference for NDM over ODM.

der to 20 participants. The scores were processed such that negative scores indicated a preference for the ODM and positive scores indicated a preference for the NDM. With 20 participants rating 20 sentence pairs, 400 data points were obtained. Figure 3 displays the distribution of scores in the experiment. The mean score was 0.1125, which reveals a slight preference for the new model over the old one. The standard deviation was 1.17. A one-sample t-test was conducted ($t(399) = 1.920$, $p = 0.056$), which revealed that the difference of the mean from zero was not significant. The 95% confidence interval for the mean ranged from -0.003 to 0.23.

4.2. Objective evaluation

To make some statements about the performance of the NDM and ODM, correlations and RMS errors were computed on the observed durations in the training corpus and on the test set of 20 sentences from the subjective evaluation. Table 3 gives the results. As can be seen, the NDM performs better (RMSE is approx. 5 ms smaller) than the ODM, both for the training and for the test data.

Data set	N	NDM		ODM	
		Corr.	RMSE	Corr.	RMSE
Training data	12,918	0.76	26.96	0.62	31.88
Test data	1493	0.77	23.40	0.65	29.48

Table 3: Quantitative evaluation of new duration model (NDM) and old duration model (ODM); N = number of phonemes, Corr. = Pearson’s r correlation.

5. CONCLUSION AND DISCUSSION

This paper has shown that a duration model using the sums-of-products approach can be produced in a very straightforward manner, resulting in a quality that is as good as the old duration model, and in some cases even better.

The fact that the subjective evaluation did not show a significant preference for the NDM may have been caused by the poor segmental quality of the diphone synthesis obscuring subtle differences in segmental duration. It may make sense to repeat the experiment with natural speech onto which the predicted durations are transplanted.

It is not certain whether the RMSE is an adequate predictor of perceived quality. It does not take into account the fact that errors in prediction may be less or more acceptable dependent on the phoneme class, the stress or accent status, or the position of the syllable in the word, etc. In a study by [2], it was investigated what the listeners’ acceptability for temporal modification was of single vowel segments in isolated words. It was shown that the acceptable range of modification depended on the position of the vowel in the word, on the vowel identity and on the voicedness of the following consonant. This suggests that the RMSE may not be a suitable predictor of acceptability because no weighting of factors takes place.

REFERENCES

- [1] W. Eefting. The effect of information value and accentuation on the duration of Dutch words, syllables and segments. *JASA* 89(1):412-424, 1991.
- [2] H. Kato, M. Tsuzaki and Y. Sagisaka. Acceptability for temporal modifications of single vowel segments in isolated words. *JASA* 104(1):540-549, 1998.
- [3] E. Klabbers. *Segmental and Prosodic Improvements to Speech Generation*. PhD Thesis, Eindhoven University of Technology, 2000.
- [4] S. Nooteboom. *Production and perception of vowel duration: A study of the durational properties of vowels in Dutch*. PhD Thesis, Utrecht University, 1972.
- [5] J. van Santen. Contextual effects on vowel duration. *Speech Communication*, 11:513-546, 1992.
- [6] J. van Santen and A. Buchsbaum. Methods for optimal text selection. In *Proceedings of EUROSPEECH 1997*, pages 553-556, 1997.
- [7] R. Sproat *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, Kluwer Academic Publishers, Boston, 1998.
- [8] J. Terken. Spoken Language Interfaces: Developments. *IPO Annual Progress Report*, 31:61-65, 1996.
- [9] J. Waals. *An experimental view of the Dutch syllable*. PhD Thesis, Utrecht University, 1999.