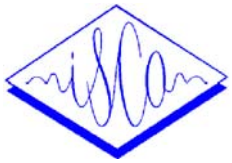

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

SEGMENTAL DURATION MODELLING IN A TEXT-TO-SPEECH SYSTEM FOR THE GALICIAN LANGUAGE

Xavier Fernández-Salgado and Eduardo R. Banga

Dpto. Tecnoloxías das Comunicacións. ETSE Telecomunicación.
Campus Universitario. Universidade de Vigo. E-36200. Vigo. SPAIN
xsalgado@tsc.uvigo.es, erbanga@tsc.uvigo.es

ABSTRACT

In this contribution we propose a segmental duration model for the Galician language. We have focused our work on the study of allophonic durations in their syllabic environment. Firstly, a study of the speech rate over a recorded corpus led us to consider different behaviours in certain types of sentences. Secondly, the corpus was analyzed in order to determine the main factors affecting duration (phonetic class, context, ...). Prosodic factors (stress and final lengthening) were found to be the most determinant, in quantitative terms, to predict timing. Finally, a model for assigning segmental durations is proposed.

Keywords: timing, duration modelling, prosody, text-to-speech.

1. INTRODUCTION

Galician is a Romance language spoken by three million people in the northwest of Spain. This language, which comes from the same root as Portuguese and resembles this language in lexicography and grammar, is closer to Spanish at the phonetic level, in fact, they share most of their phonemes. Furthermore, a great percentage of Galician speakers are bilingual, and this fact surely affects prosody.

Its being a Romance language suggests that its behaviour will be that of one syllable-timed language. In addition, our perception of the timing in this language suggests an *a priory* behaviour similar to Spanish.

The lack of previous works in this language leads us to consider the generalities of the phonetic behaviour of other languages, especially Spanish. An early work in this language is described in [1].

Our approach to the study of duration deals with the segments extracted from read utterances (connected read speech) and not with isolated words spoken in "laboratory conditions". No transformation has been applied to timing units in contrast with log-transformation in [2] and z-score transformation in [3], so all the references to durations will be given in milliseconds.

The outline of this paper is as follows: in section 2, we will give a brief description of the corpus; in section 3 a distinction among several types of sentences is done based on several measures of speech rate; section 4 deals with a previous analysis of several factors determining duration and, finally, a model for assigning durations is proposed in section 5.

2. CORPUS DESCRIPTION

Our work is based on the analysis of the durations of many allophonic segments which were obtained by the segmentation of a read corpus.

This corpus consists of 300 short sentences (14,000 allophones) and although it is mainly composed by subject + verb + predicate structures, it also considers some other types of constructions with enumerations, defining and non-defining clauses and so on. The whole corpus was recorded twice by the same speaker who had recorded the speech units for our concatenative speech synthetiser, and the sentences were read in a normal speech rate. In order to check the consistency in the speech-rate of the speaker, we have carried out some simple tests comparing the means and standard deviations of the allophones between the two recordings. Not only did the results confirm the consistency of the speaker rate but the consistency of the phonetic labellers as well.

During the labelling stage some difficulties came up, such as the contractions and ellipsions of some segments as well as the problem of tagging the closure of the plosive sounds at the beginning of an utterance. This fact obliged us to differentiate, or, at least, mark this segments.

The phonetic segments were automatically labelled according to the lexical accent of the word, syllabic structure (number of allophones in the syllable, position within the syllable), broad phonetic class (plosive, vowel, ...) and the syllable position within the breath group and the accentual group. Another important factor

is the position of the segment with respect to the next pause, either phrase internal or final.

3. SPEECH RATE

In unrestricted texts many different sentence structures may appear and any text-to-speech system must be prepared to deal with them.

The perceptual hearing of sentences suggests different speech rates depending on the sentence type, so our first task is to test whether this is true, because this consideration could drive us to regard sentence type as a new factor.

As broad measures of the speech rate we have applied three different estimators for each phonic group in the corpus: simple speech rate, corrected speech rate, and allophone rate. The first one is the number of syllables per second. As Galician does not suffer from the phenomenon of ambisyillabicity, the use of this measure is not difficult. The purpose of the second estimation is to take into account the fact that a syllable can be composed by 1,2,3,4 or even five allophones. As in Galician the most frequent length is 2 allophones, the first estimation is multiplied by a correction factor of 2 and divided by the average of the number of allophones per syllable in the sentence. The third estimation represents the rate of allophones per second.

sentence type	sp. rate	corrected sp. rate	allophones/ second
statements	6,16	6,75	13,50
y/n questions	7,62	8,52	17,04
wh- questions	6.77	7,88	16,77
alternative questions	6,58	7.56	16,72
commands	6,01	6,86	14,91
exclamations	6,03	6,96	13,93

Table 1. Estimations of average speech rate for different sentence types.

In table 1 we show the values of these three estimators classified by the type of sentence. It can be observed that, apart from questions, where the speech rate is notably higher, there is no significant difference among the values that were obtained for the other types of sentences. Taking into account these previous statistics we decided to set questions apart from the other types of sentences in our computations.

Other suppositions, as the higher speech rate in long sentences versus short sentences, were not concluding looking at our database.

4. SOURCES OF VARIATION IN ALLOPHONE DURATION

The duration of allophones can be estimated by considering several groups of factors. The most important is intrinsic duration, because the duration of an allophone is mainly influenced by its manner and place of articulation. There are also contextual factors that reflect the influence of the adjacent phonetic segments and, finally, prosodic factors such as stress and the effect of final lengthening.

The main problem in modelling durations is that there are many possible factors, each one with many possible levels, which causes sparsity of data [2] and makes difficult the treatment of incomplete datasets. The consideration of the identity of each allophone is perhaps the factor that more expands the number of possible combinations of levels (the phonetic system of Galician comprises 7 vowels and 22 consonants). So as to avoid this problem, and considering well known facts such as the manner of articulation, we have clustered the phonetic identity into some broad phonetic classes. When testing this assumption by means of ANOVA methods and by maintaining constant the phonetic environment and conditions of stress and position within the utterance, the results did not reveal a significant relation between phonetic identity and duration. For example, in vowels the phonetic identity explained only (r^2) 14,56% of variability. Besides, looking at the difference between the mean of different vowel identities, Bonferroni's corrected t tests (testing every two vowel identities for equality) did not show any significance. The same procedure was also applied to the other phonetic classes and, finally, we have considered 7 broad classes: vowels, voiced stops, unvoiced stops, nasals, liquids, fricatives, and africates. The resulting classes follow the general trends of durations signalled by Coleman for Dutch [4] and Chang for English [5].

Some other effects such as the inverse relation between the duration of a vowel and its height could not be established with this small database, although a slight higher duration was observed in the phoneme /a/. Another effect also observed was a very little increase of duration (about 3-4 msec.) in vowels when followed by a voiced allophone, which we think is not relevant.

An important observation was the reduction of allophone durations depending on the number of allophones within the syllable, as it is shown for the /i/ vowel in the table 2.

	number of allophones per syllable				
	1	2	3	4	5
duration (msecs.)	96	89	65	58	62

Table 2. Compression of /i/ vowel in stressed non final syllable as function of the number of allophones in the syllable.

Nevertheless, the most important cuantitative factors affecting durations are lexical stress and the distance to the end in prepausal syllable which are studied in more detail in the following section.

4.1 Durations on stressed syllables

A great many studies have been done on what unit is lengthened when a word is stressed. For example in [6], it is suggested that stress lengthening in Scottish English affects not only the syllable bearing pitch accent, but it also extends to the following syllable. This is not the case for Galician, where it only affects the lexical stressed syllable.

As it can be observed in table 3, where the allophone durations for several types of syllabic structures are shown, the lengthening of a syllable does not always affect every allophone in the syllable. In onset position, we cannot observe a significant duration increase in voiced_obstruent within syllables with the structure voiced_obstruent+vowel. The same holds for liquids in onset position in 2-allophone-length syllables.

In nucleus position, the vowel tends to increase duration in all cases if we except the vowel+fricative structure. The limited amount of data did not allow us to test this behaviour more accurately. This increase in duration was observed to be about 20 to 30 milliseconds in the majority of cases.

In coda position we have only tested two different syllabic structures that exhibit different behaviours. Nasals tend to be lengthened but not fricatives. As the presence of a coda almost always implies that the length of the syllable is 3 allophones, in this case the tendency of allophone compression comes into play. When analysing the behaviour of fricatives in coda position, the results were contradictory. For example, preceded by voiced_obstruent+vowel the fricative did not seem to be lengthened, but it did when preceded by a liquid+vowel.

In other cases, (non tested with t-tests and not listed in the table) such as groups of 3-allophone-length syllables with tautosyllabic group, it was observed that there was lengthening both in the liquid (about 10 milliseconds) and in the vowel belonging to an unvoiced_plosive + liquid + vowel structure.

To summarise, it was found a clear relation between the increase in the duration and the consonant broad classes, in which voiced classes seem to have a general trend to increase in lesser quantity than unvoiced classes.

Syllable Structure	Position whitin the syllable	mean duration in non-stressed syllable	mean duration in stressed syllable	t student	critical value for 0.05
v_plovo	1	48,33	52,87	1,26	1,97
v_plovo	2	52,71	85,78	9,32	1,97
livo	1	49,72	47,21	0,52	1,98
livo	2	59,49	91,60	7,50	1,98
frivo	1	94,24	105,95	2,12	1,98
frivo	2	51,63	83	8,57	1,98
vofri	1	51	53,5	0,37	2,05
vofri	2	45	68,03	1,65	2,05
u_plovonas	1	77,11	89,64	1,31	2,01
u_plovonas	2	49,77	61,76	3,48	2,01
u_plovonas	3	56,62	90,29	5,73	2,01
u_plovo	1	46,54	66,47	9,42	1,97
u_plovo	2	83,103	80,62	0.80	1,97
nasvo	1	64,16	74,88	2,00	1,98
nasvo	2	55,61	79,81	5,36	1,98
u_plovofri	1	68,46	71,5	0,45	2,06
u_plovofri	2	43,92	62,91	2,87	2,06
u_plovofri	3	54,07	56,58	0,26	2,06

Table 3. Allophone duration (msecs.) in several syllabic structures (non final syllables). Key: u_plo=unvoiced plosive; v_plo=voiced plosive vo=vowel; li= liquid; fri=fricative; nas=nasal;

4.2 Lengthening in prepausal syllables

Final lengthening (before a pause) is a common feature for many languages and a lot of work has been done in this area [7].

The table 4 shows the final lengthening effect for different syllabic structures. It shows a great increase of duration as it can be observed in the value of t tests (all the values are significant) especially in the final segment. The segmental duration in this position can be from 2 to 3 times the duration in non-final non-stressed allophones. This large increment must be examined with more accuracy because the position of the mark just before a pause may fluctuate depending on the segmentation procedure.

syllable structure	position whitin the syllable	mean duration in non pre-pausal syllable	mean duration in pre-pausal syllable	t student	critical value for 0.05
v_plovo	1	48	69	4,85	1,97
v_plovo	2	52	125	13,78	1,97
v_plovofri	1	49	57	1,71	2,14
v_plovofri	2	52	91	5,54	2,14
v_plovofri	3	69	153	5,21	2,13
u_plovo	1	83	100	3,45	1,97
u_plovo	2	46	110	17,71	1,97
u_plovofri	1	75	94	2,27	2,20
u_plovofri	2	50	74	3,73	2,20
u_plovofri	3	54	149	6,66	2,20

Table 4. Allophone durations (msecs.) in several syllabic structures depending on prepausal or non prepausal position (non stressed syllables).

The final lengthening was also found to be noticeable and significant in the onset of a 3-allophone length syllable.

5. THE PROPOSED MODEL

In order to assign durations, our approach has been to classify each allophone according to the number of allophones in the syllable, its position within the syllable, its broad phonetic class and its stressed/non-stressed and prepausal/non-prepausal features. With this information, we estimate the syllable duration for every distinct sequence of broad classes, taking into account other features such as stress and prepausal position. Then, the syllable duration is splitted into the allophone durations according to its average percentage of duration within the syllable.

Our model is based on a look-up table, where we have stored the mean value for each cell. At least five cases were employed to estimate the mean value of the different cells.

	mean absolute error	r.m.s. error
allophone	16,32	19,60
syllable	25,40	34,8

Table 5. Prediction error (msecs.) over the training database (look-up table method).

In order to validate our model, we have compared the original segmental durations of a new set of sentences (recorded in the same session) with the values predicted by our model. Looking at the largest prediction errors, we concluded that they mainly occur at final segments, because it is difficult to determine the end of the allophones.

Some other experiments have been carried out. For instance, we have used a CART model to predict allophone duration from the set of discrete features. A tree was generated for each broad class. The first branches in each CART were obtained by questions involving the distance to the pause and lexical stress. This fact evidences that these features are the most important from a quantitative point of view. Although this method is automatic, the performance did not improve the mean prediction error (ranging from 21 to 26 milliseconds for the different broad classes).

Just a few methods on duration modelling report figures of their performance. One is the model of Riley [8] based on CART. For a database with a standard deviation of 65 msec, he reports a standard deviation of the prediction error of 23 milliseconds, which is comparable to our error figures (see table 5).

Our model has also been perceptually validated by the resynthesis of original sentences with the new predicted durations by means of a sinusoidal synthesis algorithm [9]. This algorithm allows accurate modifications of the segmental durations. Informal tests did not reveal relevant perceptual errors and the duration model has been included in our Galician text-to-speech system.

6. CONCLUSIONS

A simple, but efficient method that combines exploratory statistics and prediction has been proposed to study allophone durations. This work has shown that phonetic segments can be clustered into broad phonetic classes without affecting largely the prediction of the allophone durations. Moreover, the most important factors affecting duration have been found to be the prepausal position and the lexical stress, which are prosodic factors.

It is important to emphasize that the predicted durations are affected by two main sources of “noise” that make difficult to obtain an accurate model: the natural fluctuations of the speaker and the inaccuracies and uncertainties in the segmentation.

7. ACKNOWLEDGEMENTS

This work has been partially supported by the “Centro Ramón Piñeiro para a Investigación en Humanidades” and the projects 1FD97-0077-C02-C01 and TIC96-0956-C04-02.

REFERENCES

[1] Pointon G. E. (1980), Is Spanish really syllable-timed? Journal of Phonetics 8, pp. 293-304 .

[2] Jan P. H. van Santen (1994),. Segmental duration in text-to-speech synthesis. Computer Speech and Language 8, pp. 95-128 .

[3] Campbell, W. N. and D. Isard (1991), Segment Durations in a Syllable Frame. Journal of Phonetics, 19, pag 37-47.

[4] John Coleman, Arthur Dirksen, Sarmad Hussain, Juliette Waals (1997), Multilingual phonological analysis and speech synthesis. Proc. of the 2nd meeting of ACL SIGPHON, Santa Cruz.

[5] Grace Chung (1991), Hierarchical Duration Modelling for a Speech Recognition System. Master Thesis . MIT.

[6] Alice Turk and Laurence White (1997), The Domain Of Accentual Lengthening In Scottish English. Proc. EUROSPEECH’97. pp. 795 – 798.

[7] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, P.J. Price (1992), Segmental durations in the vicinity of prosodic phrase boundaries. JASA, March 1992 Volume 91, Issue 3, pp. 1707-1717.

[8] Riley M.D. (1992), Tree-Based Modelling of segmental durations. Talking machines. G. Bailly & C. Benoit Editors .

[9] Banga, E. R. and García-Mateo C. and Fernández-Salgado X. (1997), Shape-Invariant Prosodic Modification Algorithm for Concatenative Text-to-Speech Synthesis. Proc. EUROSPEECH’97. pp. 545-548.