The Third ESCA/COCOSDA Workshop
(ETRW) on Speech Synthesis
Jenolan Caves House, Blue Mountains,
NSW, Australia, November 26-29, 1998

MODELING PHONE DURATION: APPLICATION TO CATALAN TTS

Albert Febrer, Jaume Padrell, Antonio Bonafonte
{febrer|jaume|antonio}@gps.tsc.upc.es

Universitat Politècnica de Catalunya
C/Jordi Girona 1-3 08034 Barcelona, SPAIN

ABSTRACT

There are many exhaustive works that deal with the use of models for segmental duration. The aim of this paper is to evaluate some of the properties mentioned in literature and evaluate factorial and sum-of-products models in front of a list-like approach for Catalan language as a base for a most exhaustive study on duration in this language. Sum-of-products models for vowels and subsystems of consonants seem to be more adequate to model phone duration. The parameters for the sum-of-products models are presented in the paper.

1. INTRODUCTION

Modeling prosody has become one of the biggest hazards in obtaining high quality text-to-speech (TTS) systems. The timing structure needs to be modeled in order to approximate as much as possible synthetic-speech characteristics to natural voice. An adequate timing structure in synthetic speech allows the perception of a rhythm and a natural evolution of prosody.

Segmental duration can be defined on phone boundaries on a first approach. However, there are many studies on the effects of considering larger entities on modeling duration. In this paper, segments are restricted to phones. The paper details the experimentation performed and the models considered. The models are particularized separately for vowels and different subsystems of consonants.

2. MODELING DURATION

2.1. Experimentation

The experimentation of this work is based on a registered female voice, in Catalan language, obtained from a professional radio speaker on a high-quality registration. The corpus contains 3600 short sentences with neutral intonation and constant speech speed. The number of phones is about 72000: around 54000 are used for estimate the models and 18000 are reserved to testing purpose. The restriction of the corpus to short sentences and non forced degrees of prominence can be seen as a limitation in general text coverage but an advantage for the study of other kind of parameters under similar situations.

2.2. Descriptor vector

The duration of a phone is predicted depending on different factors, which are considered to affect it. The definition of the possible factors is done taking into account the effects considered in other languages [1]. All this factors are joined in a descriptor vector.

A descriptor vector $\boldsymbol{d}$ is associated to each phone occurrence in the corpus. The components $\boldsymbol{d}_i$ of the vector are phone identity, stress, phrasal position, surrounding phones, syllable length and syllable position.

2.3. First analysis using CART

First of all, an analysis of the data with a Classification and Regression Trees (CART) system is used to evaluate the most relevant factors on phone duration. Vowels and consonants are studied independently and also different groups of consonants depending on manner of articulation.

This first analysis shows how different factors affect duration. Although trees obtained with this systems have to be read carefully when number of occurrences vary in each group, they show a first estimation of the main factors to be considered.

From the analysis, some already known considerations are observed, such as stress or prepausal position lengthening as of first importance, but also it is noted the importance of postvocalic vs. prevocalic phones. Comparing the mean duration of each of the groups obtained with the CART system with previous studies [2] a first evaluation of correctness of results is observed.

2.4. Sparsity and interactions

One of the main problems of modeling duration with a CART system, and all list-like approaches in general, is sparsity [3]. When dividing occurrences in different groups to find their mean values, not all the groups will contain enough occurrences to accept average measures as representatives, as variability in every group is often large enough.

CART do not generalize results benefiting from interactions of different parameters and abstracting properties which could help in estimating duration. This is why there is a need of a model that captures all the known properties that are observed in this specific domain. When evaluating the model used, different experimentation was done to evaluate those described advantages in front of sparsity and generalization.

As a first approximation, the effects of different parameters can be additive or multiplicative. The interactions of different parameters are difficult to model in independent terms. For example, the effect of syllable coda position in phone /s/ increases its duration, an effect that is hardly accentuated in prepausal position.

2.5. The models

The analysis with a CART system can be observed as a first model of segmental duration. However, other parametric models have been studied and tested in the corpus.

Models should predict duration of every occurrence from the descriptor vector associated. A measure of correctness of the model is necessary in the estimation procedure as well as in testing models. In this paper the mean squared error (MSE) is computed in order to evaluate the differences between observed duration occurrences (O_i) and predicted values (D_i).

$$E = \sum_i \frac{(D_i - O_i)^2}{N} \qquad (1)$$

List-like approach

The first approach consists on assigning a value to each possible combination of parameters of the descriptor vector. If the value is computed as the average duration of all occurrences of each group, the MSE will obviously be minimal for the training corpus.

This approach does not make use of the possibility of generalizing, and has no direct solution in front of sparsity.

Factorial model

The UPC TTS system [4] used a factorial model for vowel duration. This model used factors estimated in previous works for Catalan language [2]. This model assumes that interactions between different components $\boldsymbol{d_i}$ of the descriptor vector $\boldsymbol{d}$ are multiplicative, i.e. the duration of a phone in msec. $\boldsymbol{D}$ can be modeled as a product of terms:

$$D(d) = \prod_i F_i(d_i) \qquad (2)$$

Sum-of-products model

Sum-of-products models capture the phenomenon of directional invariance [1][3]. Directional invariance was observed in the experimentation procedure: the effects of a factor, like stress or prepausal position, have always effects on the same direction. For example, observing the mean values of two vowels, average duration of non-prepausal /O/ is longer than /o/. Holding all else constant, the same vowels in prepausal position had longer duration values but holding /O/ longer than /o/.

But the effects of sentence position do not affect in the same percentage to all vowels. So the use of factorial models can not model properly this situation. Neither the use of an additive model can model the interactions between factors. A combination of sums and products is more capable of reflecting the properties of duration, as directional invariance and interactions, but their parameters are not easy to estimate.

$$D(d) = \sum_i \prod_j S_{i,j}(d_j) \qquad (3)$$

2.7. Parameter estimation

The estimation of the parameters of both the factorial and sum-of-products models is not obvious. A gradient algorithm has been used in the estimation with the goal of minimizing the error function $\boldsymbol{E}$ defined in (1).

The estimation of the parameters for a factorial or sum-of-products model is performed by the successive application of submodels considering the factors related to different components of the descriptor vector. Given a model, the estimation of the parameters begins with the estimation of the submodel formed only by the most important component (in vowels for example, phone identity). Once these parameters have been determined, the next component (in vowels, sentence position) is introduced and the gradient method determines the new set of parameters. This process obtains the global set of parameters when all components of the descriptor vector have been introduced and the error function $\boldsymbol{E}$ converges.

3. MODELS FOR VOWELS

The Catalan vowel system contains eight vowels divided in a stressed subsystem (/a, e, E, i, o, O, u/, SAMPA notation [5]) and an unstressed subsystem (/e, i, u, @/).

The observation of CART took to consider the next components of the descriptor vector as of first importance in the modeling of vowels duration:

· Vowel identity (v), 8 levels: /a, e, E, i, o, O, u, @/

· Stress (a), 2 levels: stressed, unstressed.

· Sentence position (p), 2 levels: prepausal, non-prepausal.

· Class of post-vocalic phone (c), 2 levels: voiced, voiceless.

· Manner of articulation of post-vocalic phone (t), 8 levels: silence, vowel, nasal, vibrant, plosive, approximant, fricative, lateral.

In the estimation of the parameters of the models the components of the descriptor vector were introduced gradually by its importance in the CART analysis. The analysis of the behavior of the models began with the consideration of vowel identity (v) and stress (a), and successively adding sentence position (p), class (c) and manner of articulation (t) of post-vocalic phone.

The resulting factorial model consists of 5 multiplicative terms as follows:

$$D = F_1(v)F_2(a)F_3(p)F_4(c)F_5(t) \qquad (4)$$

And the sum-of-products model used is based on the Klatt model [7] adding the effects of stress and manner of post-vocalic phone:

$$D = S_{1,1}(v) + S_{2,1}(v,a) + S_{3,1}(v)S_{3,2}(p)S_{3,3}(c)S_{3,4}(t) \qquad (5)$$

The same terms were tested in different sum-of-product models, varying the combination of parameters. Models with additional terms added to (5) (for example $S_{4,1}(v)$ $S_{4,2}(p)$) showed small improvement in MSE but an increase in the number of parameters to estimate.

The results of the application of the models for different set of components of the descriptor vector are shown in figure 1. Table 1 contains the number of parameters required for the submodels considering only vowel identity (v) and stress (a), adding sentence position (p), class (c) and manner of articulation (t) of post-vocalic phone.

The list-like model obtains the best MSE but increasing the number of parameters to estimate as the descriptor vector increases as shown in table 1. The sum-of-products model defined in (5) obtains lower MSE than the factorial model with few additional parameters.

Number of parameters

Model	v,a	v,a,p	v,a,c,p	v,a,c,p,t
List-like	11	21	41	152
Sum-of-products	11	20	21	29
Factorial	9	10	11	19

Table 1: Number of parameters of the models considering different set of components.

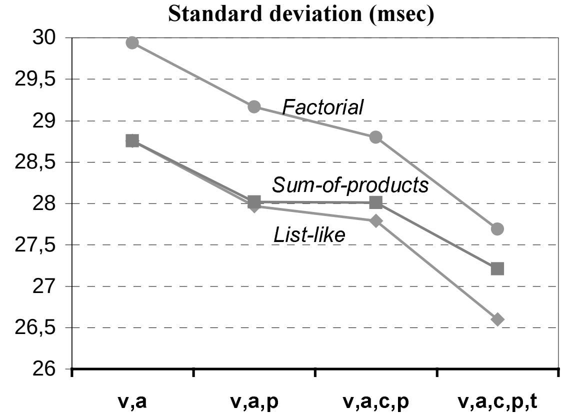

Figure 1: Evolution of the standard deviation of error ($\sqrt{\boldsymbol{E}}$) of list-like, factorial and sum-of-products models considering gradually the different components of the descriptor vector.

However, the importance of the models appears when there is sparsity in data. A simulation was designed to test these situations. Different groups of data were eliminated from the training corpus. For example, when all occurrences of vowel /o/, in non-prepausal position, followed by a voiceless stop were eliminated from the training text, the estimation of the parameters of the sum-of-products model was nearly the same, with an acceptable approximation of the actual average value of the group. This was accomplished due to the directional invariance property inherent in the model, which allows extrapolating the behavior of the parameters. On the other hand, the list-like model could not give any value for this descriptor vector when testing those occurrences. The experiment was repeated with different groups with similar results.

From these results, the sum-of-products model is selected for the application to the TTS system. The model has 29 parameters: $\mathbf{S}_{1,1}$ (8 vowels), $\mathbf{S}_{2,1}$ (3 vowels can be stressed/unstressed), $\mathbf{S}_{3,1}$ (8 vowels), $\mathbf{S}_{3,2}$ (prepausal factor), $\mathbf{S}_{3,3}$ (voiced phone factor), $\mathbf{S}_{3,4}$ (8 levels of **t**). The parameters estimated are shown in table 2.

v	$\mathbf{S}_{1,1}$	$\mathbf{S}_{2,1}$	$\mathbf{S}_{3,1}$
a	73.38	0	1.17
e	41.55	34.79	1.64
i	58.64	16.40	1.80
o	66.38	0	1.47
u	59.47	16.59	1.09
E	76.59	0	1.61
O	74.70	0	1.43
@	50.20	0	1.21

p	$\mathbf{S}_{3,2}$
Prepausal	4.25
Non-Prep	1.00

c	$\mathbf{S}_{3,3}$
Voiced	2.41
Voiceless	1.00

t	$\mathbf{S}_{3,4}$
sil	5.84
vow	6.17
nas	1.00
vib	2.23
plo	1.99
app	2.66
fri	6.89
lat	2.05

Table 2: Parameters of the sum-of-products model for vowels.

The comparison of the results obtained with the model with previous works shows a consistency between results obtaining more broad coverage. It is noticeable the effect of prepausality ($\mathbf{S}_{3,2}$) on the vowel lengthening (this effect is accentuated in final syllable position, i.e. followed by silence ($\mathbf{S}_{3,4}$)), and also the contrastive effect of voiced/voiceless fricatives vs. voiceless plosives ($\mathbf{S}_{3,4}$) on Catalan vowel duration as pointed in [2]. As an example of the use of the model, the predicted duration for /a/ (always stressed) in prepausal position followed by /p/ (voiceless plosive) is 83.28 msec. (73.38 + 1.17*4.25*1.99); and the duration for stressed /i/ in non-prepausal position followed by /s/ (voiced fricative) is 104.93 msec. (58.64 + 16.40 + 1.80*2.41*6.89).

From the observation of the parameter values of the model, it was noticeable that diphthongs were candidates to form a subsystem in a specific model. In the first evaluations, semivowels appeared as a parameter applied in the post-vocalic phone term. The difference of numerical order between semivowel terms and other phones confirmed the need of studying diphthong behavior independently.

Other models have been proposed to model the logarithm of duration. Some experimentation was done in a model similar to the one proposed in [6]. The model did not show an improvement in the error rate. Comparing all the sum-of-products models used and the factorial model, the one detailed in (5) seems the most adequate, but the need of a deeper study on models using the same number of parameters arises.

4. MODELS FOR CONSONANTS

The difference in the nature and properties of different groups of consonants suggests the division of the model for consonants in a set of subsystems based on manner of articulation.

The capability of a sum-of-products model in extrapolating and generalizing should be more effective when the set of possible descriptor vectors is restricted to similar contexts. The analyzed subsystems were nasals (m, n, N, J), voiceless plosives (p, t, k) fricatives (S, s, f, Z, z), liquids (l, r) and voiced plosives (b, d, g).

The process for consonants was similar to the described process for vowels. CART experiments with global set of consonants and independently in the subsystems resulted in the following descriptor vector:

- Consonant identity (v), 1 level each consonant:
- Stress of the syllable (a), 2 levels: stressed, unstressed.
- Sentence position (p), 2 levels: prepausal, non-prepausal.
- Syllable position (r), 2 levels: onset, coda.

The observation of the average duration values in each group of equal descriptor vectors showed again the properties of directional variance. Consonants in stressed syllables are longer than in unstressed syllables, prepausal consonants are longer than non-prepausal ones, and consonants on syllable coda are also longer than on syllable onset.

A reasonable sum-of-products model for consonant duration is:

$$D = S_{1,1}(v) + S_{2,1}(v,a) + S_{3,1}(v)S_{3,2}(p)S_{3,3}(r) \qquad (6)$$

The subsystems of nasals, voiced plosives and fricatives were modeled by (6), resulting in 14, 11 and 17 parameters to estimate respectively.

In table 3 the estimated parameters for the three subsystems are listed.

v	$S_{1,1}$	$S_{2,1}$	$S_{3,1}$
m	71.9	3.58	0.46
n	47.5	7.91	1.02
N	84.6	1.47	0.72
J	131	0.96	0.56

p	$S_{3,2}$
Prepausal	6.75
Non-Prep	1.00

r	$S_{3,3}$
Coda	11.17
Onset	1.00

v	$S_{1,1}$	$S_{2,1}$	$S_{3,1}$
p	77.2	1.72	0.71
t	69.3	1.58	0.75
k	85.0	0.65	0.53

p	$S_{3,2}$
Prepausal	6.66
Non-Prep	1.00

r	$S_{3,3}$
Coda	12.33
Onset	1.00

v	$S_{1,1}$	$S_{2,1}$	$S_{3,1}$
S	93.4	1.44	4.45
s	86.8	1.63	8.42
f	94.6	1.11	2.38
Z	69.7	1.85	2.70
z	67.4	1.68	4.75

p	$S_{3,2}$
Prepausal	5.87
Non-Prep	1.00

r	$S_{3,3}$
Coda	1.73
Onset	1.00

Table 3: Parameters of the sum-of-products model for nasals, voiceless plosives and fricatives.

From the observation of the parameters, a higher influence of accent (**$S_{2,1}$**) is manifested in nasals /m/ and /n/, while the other consonants are not so affected. The effects of sentence position (**$S_{3,1}$**, **$S_{3,2}$**) are highly accentuated in fricatives, especially in the phone /s/ prepausal and in coda (**$S_{3,3}$**).

The observation of the behavior of the subsystem of liquid consonants suggested a different treatment. Groups formed by a plosive consonant, a liquid consonant and a vowel should have a particular treatment. The coarticulation in these cases makes difficult to determinate the correct emplacement of phone boundaries, as these liquid consonants are highly contaminated by the following vowel.

The analysis of the subsystem of voiced plosives with CART did not make clear distinctions in which factors were determining for modeling duration.

5. APPLICATION TO TTS

The sum-of-products model has been integrated in the TTS of the Universitat Politecnica de Catalunya (UPC) [8]. A base of concatenative diphone units has been extracted from the same corpus. In the prosody assignment module the duration of every phone is predicted using the model with the adequate descriptor vector.

Various limitations can be observed in the use of this model. The first one is the lack of modeling of parameters related with stress levels, long sentences or intonation. Another question not solved by the experimentation is the alteration of values of duration when varying speech speed. Some experimentation should be done to determine which phones reduce duration when speaking quickly and which remain almost unaltered.

The model obtained reflects the behavior of the Catalan phonetic system, but its particular values have been estimated for only one speaker. A generalization of the model for other speakers should consider which parameters could be affected. However, the same set of values has been applied to the TTS system for all the speakers with similar results.

6. ACKNOWLEDGEMENTS

This work has been partially supported by CIRIT, Generalitat de Catalunya, through the Centre de Referència en Enginyeria Lingüística (CREL).

7. REFERENCES

[1] *Multilingual Text-To-Speech Synthesis: The Bell Labs Approach*, Richard Sproat, editor, Kluwer Academic Publishers, 1998.

[2] L. Aguilar et al. “Catalan vowel duration”, *Proceedings of EuroSpeech'97*, pp.771-774, Rhodes 1997.

[3] J. van Santen, “Prosodic modeling in text-to-Speech synthesis”, *Proceedings of EuroSpeech'97,* KN-19, Rhodes 1997.

[4] A. Bonafonte, I. Esquerra, A. Febrer, F. Vallverdu, "A bilingual text-to-speech system in Spanish and Catalan", *Proceedings of EuroSpeech'97*, pp. 2455-2458, Rhodes 1997.

[5] URL: http://www.phon.ucl.ac.uk/home/sampa/home.htm

[6] Klatt, D. “Interaction between two factor that influence vowel duration”, *Journal of the Acoustical Society of America*, 54: 1102-1114.

[7] J. van Santen, “Assignment of segmental duration in text-to-speech synthesis”, *Computer Speech and Language*, 0:95-128.

[8] URL: http://gps-tsc.upc.es/veu/demos.html