CORRELATION BETWEEN PHONETIC FACTORS AND LINGUISTIC EVENTS REGARDING A PROSODIC PATTERN OF EUROPEAN PORTUGUESE: A PRACTICAL PROPOSAL

*Diamantino Freitas, Daniela Braga, Maria João Barros, Vagner Latsch, João Paulo Teixeira**

Faculdade de Engenharia da Universidade do Porto, Portugal, telephone +351 22 5081600
*ESTiG – Instituto Politécnico de Bragança, Portugal, telephone +351 273 303000

dfreitas@fe.up.pt, dbraga@fe.up.pt, mjbarros@fe.up.pt, vagner@fe.up.pt, joaopt@ipb.pt

Abstract

In this article a prosodic model for European Portuguese (henceforth EP) based on a linguistic approach is described. It was developed in the scope of the Antigona Project, an electronic-commerce system using a speech interface (Speech to Text plus Text To Speech, the latter based on a time concatenation technique) for EP language. The purpose of our work is to contribute with practical strategies in order to improve synthetic speech quality and naturalness, concerning prosodic processing. It is also our goal to show that syntactic structures strongly determine prosody patterns in EP. It is also important to emphasize the pragmatic commercial objective of this system, which is selling a product. Therefore, this type of application deals with a specific vocabulary choice, it is displayed in predictable syntactic constructions and sentences, making prosodic contours and focus become expected. This study was held in intimate articulation between the engineering experience and tools and the linguistic approach. We believe that this work represents an important achievement for future research on synthetic speech processing in particular for EP. Moreover, it can be applied to other Romanic languages, regarding their syntactic resemblances.

1. Introduction

There is a wide discussion around prosodic control of speech synthesis and the conditions and methods involved in order to process its modulations. Are prosodic events regulated and determined by linguistic facts or is prosody completely unpredictable and out of automatic control? This is the focus of present research and a most controversial issue in speech synthesis. Presently, research groups seem to spread their studies in linguistic, psycho-acoustic or stochastic approaches [1]. In our study, we try to associate both linguistic and stochastic approaches, using linguistic rules as far as possible.

It is widely accepted in the scientific community that prosody involves both physical, so-called suprasegmental, factors, on one hand, and linguistic events, on the other.

Despite the broad improvements on the prosodic field achieved by the scientific community concerning other languages [2], the studies on European Portuguese Prosody are still in progress. There is still scarce work on practical prosody applied to Speech Synthesis for EP. A few remarkable Portuguese researchers have already contributed with works on this field, namely Sónia Frota [3], Madalena Cruz-Ferreira [4], Martins[5], Mateus[6] and others. Electronic Commerce is now demanding more appealing multimedia strategies in order to reach more ambitious markets. Speech Synthesis for Portuguese is being required as an important contribution to this commercial area, as well as in others concerning Natural Language Processing.

In the present paper, we will describe the approach taken by our multidisciplinary team to develop prosodic control of our TTS. We will concentrate in f0 control, or intonation, that is a first priority in the way of achieving speech naturalness. Durations and rhythm control are very important as well, and studies in the topic are under way in our group. The method adopted for our work has started by surveying existing work in the field with selection of most viable techniques. It is our belief that linguistic contents of the text are intimately tied to the selection of the prosody pattern, so it was decided to design a special purpose text analyzer/parser. The output of this module, that comprehends morphological and syntactic analysis conveys the required information for selection of a prosodic pattern. Moreover, the prosodic f0 pattern is time anchored to the text's distinctive features extracted from the analysis, so that time warping is not a problem for production of the final result. The technical modeling of f0 patterns was performed using a Fujisaki model approach. The INTSINT system seems another possible way to model f0 contours, but TOBI is quite far from being practical.

Stochastic modeling is another possibility that was tested without much success, but, because rules exist for most aspects of f0 control in EP, it seems more promising to reserve stochastic modeling for the random part of the model that still has some positive influence in the final naturalness that can be achieved. The f0 modeling was done by means of a corpus, in written and recorded forms. The analysis produced nice modeling results that are reported in the following, using an analysis by synthesis technique.

2. Linguistic corpus

2.1. Text

The analyzed text was selected from several electronic-commerce Portuguese sites. Part of it represents the linguistic corpus to be used in our e-commerce application. Obviously, the lexical field is reduced, as well as the syntactic structures involved. The ilocutory objective, in Searle's[7] theory, is also very specific: seducing and convincing the potential buyer. Therefore, informative sentences (Declaratives) are expected, in order to report daily promotions, to tell phone numbers, addresses, payment ways, and so on. Imperative sentences are expected to be used for orientations or instructions. Finally, Interrogative sentences are necessary to allow the successful performance of both the system and the sale. Exclamatory phrases (we use sentence as a synonym of phrase) are also possible, specially to advertise promotions or new products in general.

The corpus was initially organized in the above mentioned classic phrase division, and each of the four types displayed according to their extension. We began by analyzing simple sentences, but soon the requirements of our application forced us to extend the scope of the study to more complex sentences, with two or three clauses, and enumerations, frequently used in lists of products. This will have repercussions in syntactic structures and, as a result of it, in f0 contours.

2.2. Speakers and recording

For our prosodic data basis, 4 graduated speakers were chosen, 2 male and 2 female, with ages between 20-40 years. As our study is done on EP normative pattern, it explains the preference for speakers' higher education in order to avoid dialectal interferences that have specific prosodic displays. The choice of more than one speaker has the purpose of trying to eliminate what we call "prosodic individual creativity" and find an average prosodic pattern.

The speech signal was sampled at 11 kHz, 16 bits, mono. The sound recording was done in the Cool Edit Pro® environment. The entire database was recorded in homogeneous conditions, the same microphone was used and the same quiet room conditions. The phoneme labeling and f0 extraction was performed in the Pratt [8] signal analyzer.

3. The Antigona Model

3.1. A matrix of linguistic events

To perform the apparently easy action of reading a text, supposing that the reader is physiologically normal and a native speaker, complex psycho-cognitive processes are involved, besides the phonetic and anatomic aspects related to speech production itself. It is quite difficult to follow a reader that does not sufficiently understand the contents he/she is reading. In our study, we tried to find out how does a Portuguese native speaker organize and interpret a written text in such a way that he/she is able to read it and be understood by an audience. This intelligibility is mainly reflected in the natural prosody produced.

It is widely accepted that prosody is not only essential to synthetic speech naturalness, but is also required to give sense to the textual information.

Therefore, what linguistic information available from the text is determinant to perform prosodic processing? This was the question that ruled our research and led us to structure linguistic information in the following 6 levels, which we are going to describe. They can be displayed in a matrix where the different rows and columns are set in correlation.

3.2. Graphic level

The graphic level, the text itself, is the input in a TTS system. This is the most relevant level of the matrix. It is very important to select and extract as much information as you can get from the text. After a grapheme-phoneme conversion based on linguistic rules and after an accurate numerals and acronyms conversion, punctuation is the best and the first prosodic parser in EP language. Differently from English, it is not expected to have very long sentences in EP without any punctuation mark at all. Punctuation rules are quite strict and these marks are usually enough to map the text in terms of breathing pauses. Of course we are referring to written texts, which immediately involve careful sentence structuring. In this e-commerce application, this issue is even more specific and restricted: we are dealing with short sentences where punctuation marks are abundant.

The type of information given by punctuation is, in a first view, the boundaries between sentences, and then between sentence internal sequences. This is relevant to decide what we call the "Prosodic Groups", in other words, phrasal sequences to which a certain f0 contour is set in relation with.

Other profitable cues given by the text are capital letters. They are useful sentence beginning markers, but can also be used to emphasize a certain word or sequence as

if it is a "broad focus" [9], when converting the whole word in capitals. Obviously, this is part of the pre-processing of the text that is used as input to the synthesizer.

3.3. Syllabic level

Syllable was considered the minimal unit where important prosodic phenomena can be detected in terms of f0, intensity and durations. Although it is agreed that phones carry prosodic information, which made us consider phone labels in this level, from our observation it seemed that we could extrapolate an average f0 value for each syllable. It is not our purpose to present here the syllabic structure in Portuguese. We followed an approach based on an orthographic syllabic division rather than a phonetic one, since TTS systems start from written text. Nevertheless, phonetic aspects related to natural speech realization such as vowel reductions or neutralizations, phoneme's assimilations and other transformations, were also considered in the recorded linguistic corpus and are dealt with elsewhere [15]

Stress is another important aspect that is considered. There is a *word accent* and a *sentence accent*, the latest closer to what some authors call rhythm.

Word accent is a slight, but perceptible, increase of f0, intensity and duration in a certain syllable. Portuguese word accentuation is not fixed. There are tonic syllables in the last syllable, e.g. «*sofá*», «*Canadá*» (sofa, Canada); in the last syllable but one (penultimate), e.g. «*cidade*», «*conferência*» (city, conference); and in the antepenultimate syllable, e.g. «ridículo», «vitória» (ridiculous, victory).

Table 1

Punctuation signal	Prosodic Group/ meaning
[.]	Expected in a declarative contour. Sentence boundary.
[?]	It has several possibilities of f0 contour according to the type of sentence: Wh-question, Alternative question, Yes/no question, Eco question or Question-tag. Sentence boundary.
[!]	Present in emphatic f0 contours, in colloquial speech to express orders, likes, dislikes and emotions. Sentence boundary.
( )	Both sentence and intra-phrasal boundary.
:	Sentence boundary. Suspensive contour.
;	Intra-phrasal boundary.
,	Intra-phrasal boundary.

We labeled all syllables under a simplistic tonic versus non-tonic definition. There are acoustical studies, relating the relative positions of the non-tonic syllables and the tonic syllable, that report the influence of proximity over non tonic syllabic behavior. Nevertheless, this aspect was not treated yet in our work.

Moreover, there is also the rhythm, also said *phrasal accent* sub-level, to be considered. Sentences have a *focus*, usually corresponding to a tonic syllable, more clearly perceptible, because they have the highest intensity values and because it is longer than the neighbour syllables. F0 is not necessarily higher. But there are secondary phrasal accents which are responsible for the rhythm of the sentence, and that obviously coincide with tonic syllables. Rhythm has already been studied by some authors [10], but there is still scarce work in this subject for Portuguese.

The position of the focus in the sentence is something that is still not very clear. It must happen in a content-word, that is to say a noun, an adjective, a verb or an adverb. Since these words shape the meaning of the text and are enough to fulfil the communication objective, they are likely to be the focus, while function words are responsible for establishing connections between these content words. Anyhow, in a sentence that has not any topicalization phenomenon, the last content word is expected to be the focus of the sentence. This is more evident in interrogative sentences. The focus in Portuguese, contrarily to English, in which the focus is in the entire word, affects mainly the tonic syllable of the content word.

3.4. Word level

In this level we started by classifying words manually, according to their morphological class. The word labels are inscribed into a traditional grammar theoretical frame. The classes considered were the following content words: nouns, verbs, adjectives, adverbs and some pronouns; and the following function words: prepositions, determiners and some personal pronouns in Portuguese (personal pronouns which replace noun phrases with direct and indirect object syntactic function are phonologically subordinated to the verb: e.g. «*Disse-o*», «*Falei-lhe*»). Conjunction class is a special case with very different behaviors. Each conjunction or sentence connector requires individual analysis. Other classes were also considered like contractions between some prepositions and articles or some demonstrative determiners. A few sub-classes were also created: proper nouns, auxiliar verbs, full verbs, articles, and numerals.

After this manual first phase approach, the morphological analysis task was then directed to a commercial analyzer produced in Portugal.

The graphic level informs us of the number of words in a text, just by processing the spaces between them. Anyway we need to extract more information than that. This level of linguistic analysis is responsible for the grouping of words into Phrasal Groups. It is also important to help in Phrasal Groups' boundaries

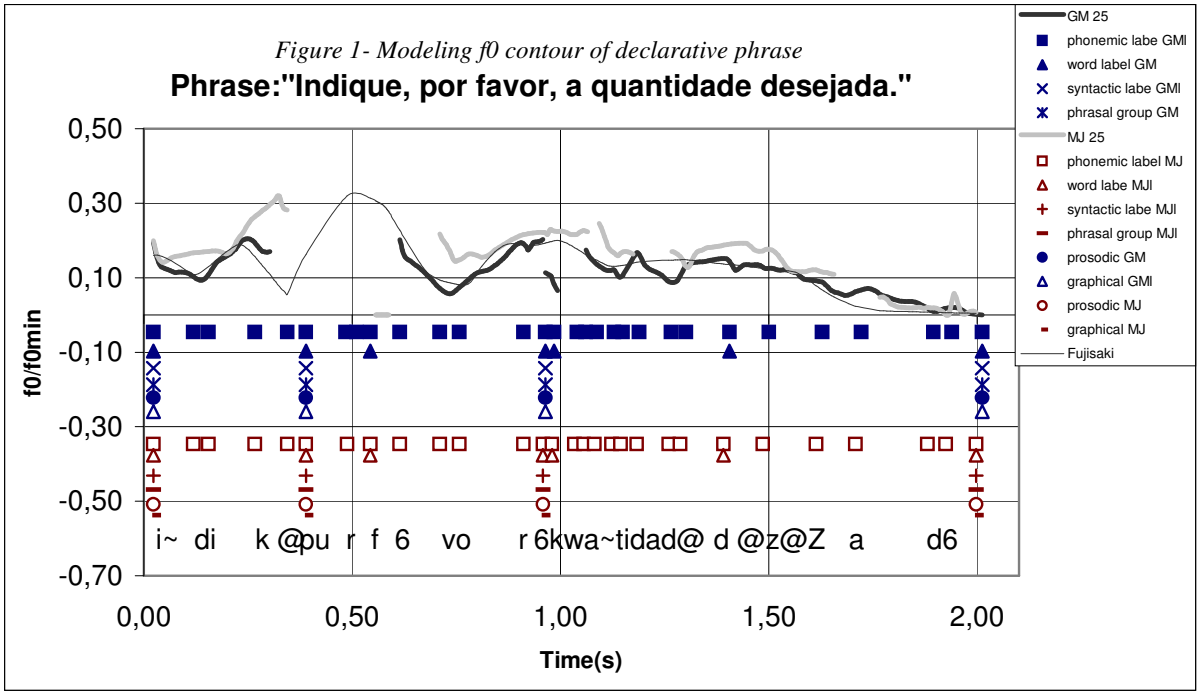

Figure 1- Modeling f0 contour of declarative phrase

decision. It does the mapping of the sentence rhythm, since there are monosyllabic content words that are obviously more stressed than disyllabic function words. It is a great factor for determination of the prosodic focus, since we hardly have focus over function words.

3.5. Syntactic level

Words are arranged and related according to syntactic restrictions and rules. Verbs are the nucleus of sentences in Portuguese. They behave as a gravity center attracting and conditioning words around them. This scope of influence is their "valence". This is the central concept in the Valence Grammar, a linguistic theoretical approach funded by a French linguist, Tesnière, in the sixties. It represents a step forward with regard to Chomsky's internationally known Generative Grammar, because it is more concerned with semantics. A Valence Grammar approach was applied to Portuguese grammar by M. Vilela [11]. Our syntactic module is partly based on the Valence Grammar framework. It analyses which words or expressions are supposed to exist at both sides of the verb. Those words are grouped under the names of "Actants" and "Circumstants". Actants are the obligatory "arguments" in Generative Grammar. Nevertheless, their syntactic and semantic roles are more complex and so were extended. Circumstants are the circumstances, the phrasal sequences that are accessories in the verb argumental framework.
When there are no verbs, nouns replace verbs' position and project their valence. Content words such as verbs, nouns and adjectives are nucleus of valence.

The Generative Grammar is also present in our syntactic analysis. The concept of "noun phrase", "verb phrase", "prepositional phrase" and so on, was brought back under the name of *Syntagma*. Distributional rules and predicates for these syntagmas were programmed in PROLOG. The morphologic analyzer was integrated by PROLOG as well, resulting in a syntactic analyzer/parser.
In short, this level of linguistic analysis works under an hybrid grammar, a theoretic frame that combines the generative syntactic approach and the semantic valence approach.

3.6. Phrasal Group level

In this level previous linguistic information is organized in longer groups than *syntagmas*. They may include more than one syntagma. Boundaries are usually coincident with eventual punctuation marks. Nevertheless, there are sentences without enough punctuation. Based on the syntactic analysis, we systematized this level in the following types of Phrasal Groups:

- Subject
- Simple Predicate
- Double Predicate
- Vocative
- Qualitative sentence
- Discursive Markers
- Non-restrictive relative clause
- Circumstant
- Coordinate clauses

- Subordinate clauses
- Enumeration sequence
- Last enumeration term
- Syntagma
- Lexeme

There are no specific Phrasal Groups for e-commerce texts. But some are obviously more frequent, such as predicates, used for instructions and enumeration sequences common in lists of products. It is not the purpose of this paper to describe each of these Phrasal Groups in detail. Their classification derives from the morphological and syntactic structure. This level of linguistic analysis is decisive to organize phrasal units or segments of meaning, which is important to extract prosodic boundaries. Inside each type of these Phrasal Groups, f0 presents a specific configuration, which is apparently constant in non-final Phrasal Groups. The major f0 variations occur in final Phrasal Groups and this is related with type of sentence and communication objective.

3.7. Prosodic level

This level is the highest and the one to which the entire linguistic information is oriented. It contains f0, intensity and durations behaviors along the phrases. In spite of the importance of all prosodic aspects, we have concentrated our study on f0 in this paper. Anyhow, durations are the parameter that deserve more attention, after f0. Intensity is the least of the three aspects in importance, and the less treated one as well.

Prosody derives directly from two levels in the matrix: the graphic level, in one way, and the Phrasal Group level, in the other. Word level brings useful account to the focus of the sentence, by distinguishing content words, the best candidates, from function words. Finally, syllabic level reports tonic syllables and phrasal accents, contributing with two levels of prosody: word prosody and phrasal prosody.

Returning to the graphic level, punctuation is, most of the times, an efficient boundary marker. It indicates where sentences finish and divides internal phrasal sequences. In Table 1, above, we tried to represent the correlation between graphic and prosodic levels, for the most usual punctuation marks in e-commerce.

Phrasal Groups confirm or fulfill punctuation boundaries. Each of the Phrasal Groups is associated with a specific prosodic pattern.

Prosodic patterns may be in a non-final position and/or in a final position. Non-final Prosodic Groups present typical f0 modulations. Every non-final Phrasal Group has a similar prosodic contour, no matter the sentence where it is located.

Prosodic creativity for EP can clearly be observed in final Phrasal Groups. Based on Cruz-Ferreira's [5] contribution and in f0 observations, we came to the following prosodic groups in emphatic sentences:

a) Final Prosodic Groups
   - Declarative
   - Suspensive declarative
   - Last term of Enumeration
   - Wh-question
   - Yes/no question
   - Alternative question
   - Eco-question
   - Question-tag

b) Non-final Prosodic Groups
   - intra-phrasal sequence
   - enumeration sequence

In natural speech, emphatic and modalized utterances may equally occur. We have studied some, more concerned with exclamatory sentences. Anyhow, we decided to choose one pattern among the non-emphatic prosodic groups that seemed to us more usual, not only in e-commerce, but also in Portuguese language in general. In chapter five we present an example with a declarative phrase. For this example a full syntactic analysis was produced. Recording of two speakers reading of the phrases was analyzed and the f0 pattern displayed with superimposed syntactic labels and synthesized f0 contour.

4. Mathematical model for f0 pattern

The approach used for derivation of the technical model for f0 control in EP for our TTS system was based on the analyzes by synthesis method with Fujisaki (see appendix) elements for phrase components and word accent components [12]. Consideration was taken of other techniques, e.g TOBI based [13], and target-point modeling, but the present approach seemed less prone to be influenced by the variability of the corpus and is definitely more connected to the linguistic framework we've developed.

A quantitative study of the main aspects only of the f0 patterns in simple phrases and syntactic minimal-pairs of various types allowed the collection of a rich set of values for the parameters of the Fujisaki mathematical model for EP. Modeling by means of identification techniques is also a trend that is gaining momentum in the research community [14], but was not considered by us in this wok up to now, although it is very promising due to its intrinsic parametric nature. The Fujisaki model has been applied with success to several languages. It employs two functions with parameters that produce logarithmic guaranties that can be added to produce thee desired logarithmic f0 patterns. The parameters must be found from experimental studies of the real sounds. Base f0 value and timing commands (inputs to the two Fujisaki functions) with precise time and amplitude values, are required. The other parameters, namely α, β and γ, can take the suggested values in [12], but refinements are possible, to cater

with varying phrase lengths and accent styles, for instance.

5. Results

In our work the location in time of the commands were derived from text analysis/parsing results, in face of the quantitative studies, above referred for EP. The resulting f0 patterns were superimposed over the example recorded sentence f0 pattern, for 2 speakers, to show the reasonably good agreement. Small micro-prosody aspects remain to be tackled, but the present results already bring a perceptually quite natural f0 modulation, as shown in informal listening tests done in our lab. Figure 1, above, presents a plot of the 3 f0 patterns, being 2 for the real speech waveform f0 contour, after some time warping, needed for superimposition and the third plot for a practical approach of use of the Fujisaki f0 model. The model uses 3 phrase components plus 6 accent components. The phrase components are referenced to the first 3 major sentence boundaries indicated by the phrasal group labels and are given durations values equal to the respective sentence's. The accent components are placed at the most prominent tonic syllables with identical durations and in the present example take decreasing amplitudes from beginning to end of the phrase.

6. Conclusions

In the course of the presented work it was demonstrated that a full syntactic analysis of text is capable of supplying enough information to drive a parametric rule-based model for artificial f0 generation in replica of natural patterns. The use of a syntactic analyzer/passer built in PROLOG, together with a mathematical modeling produces f0 patterns similar to the real ones giving natural f0 contours. In our present phase of work an increasing interaction between the engineering approach and linguistic rules is needed for future achievements. For durations and intensity we planned to have practical results in the near future.

7. Acknowledgements

The work reported in this paper is largely inspired in the activities of COST 258-Naturalness of Synthetic Speech and is partly funded by the EUREKA/ IC-PME Project Antigona. We also acknowledge PRIBERAM, PRAAT, FEUP and IPB for the support given.

8. References

[1] Monaghan, Alex 1999 "State-of-the-Art Summary of European Synthetic Prosody R&D", Cost 258.

[2] Lopez, Eduardo 1993, *Estudio de Técnicas de Processado Linguístico y Acústico para Sistemas de Conversión Texto-Voz en Español basados en Conactenación de Unidades*, Tesis Doctoral, Universidad Politécnica de Madrid

[3] Frota, Sónia 2000; *Prosody and Focus in European Portuguese,*New York, Garland Publishing ,Inc.

[4] Cruz-Ferreira, Madalena 1998; "Intonation in European Portuguese" , in Hirst,D.; Di-Cristo, A.; *Intonational Systems,* Cambrige University Press.

[5] Martins, M.R. 1988; *Ouvir Falar*, Caminho, Lisboa.

[6] Mateus et al 1990; *Fonética e Fonologia do Português*, Universidade Aberta, Lisboa.

[7] Searle, J. 1969; Speech Acts. An Essay in the Philosophy of Language, Cambridge University Press.

[8] PRAAT, Copyright 1992-2001, by Paul Boersman and David Weenink, www.praat.org.

[9] Monaghan, Alex 1993; "What determines Accentuation? A Reply to Cruttenden & Faber", *Journal of Pragmatics* 19, pp. 559-584

[10] A. Monaghan: "Rhythm & Stress Shift in Speech Synthesis." *Computer Speech and Language* 4 (1), pp. 71-78.

[11] Vilela, Mário 1999; *Gramática da Língua Portuguesa*, Almedina, Coimbra.

[12] Sagisaka, Y. Et al 1997; *Computing Prosody*, Spring New York, USA, ISBN 0-387-94804-X, ch 3, pp. 27-40

[13] Bruce, E. et al 1995, Speech Synthsis in Spoken Dialogues Research, proc. Of Eurospeech95, Madrid, September95, pp. 1169-1172.

[14] Vich. R. et al 2000;Homomorphic Decomposition of the Fundamental Frequency Contours of Utterances, Konvens, Sprach Kommunikation, Tech Unive. Limeneau, Germany, Verlag, Berlim, pp.179-184.

[15] Barros, M. J. et al, Backclose nonsyllabic vowel [u] behavior in European Portuguese: reduction or suppression, paper to be presentes at the ICSP'2001, Taejon, Korea, 22-24 August, 2001.

9. Appendix

Brief reference to the Fujisaki model equations:

$$\ln f0 = \ln fb + \sum(Api\ Gp(t-t0i)) + \sum(Aaj\ \{Ga(t-t1j) - Ga(t-t2j)\})$$

$$Gp(t) = \alpha^2 . t . \exp(-\alpha . t);\ \text{when } t>=0 \text{ or } 0 \text{ when } t<0.$$

$$Ga(t) = \min[1-(1+\beta t)\exp(-\beta t);\ \text{when } t>=0 \text{ or } 0 \text{ when } t<0.$$

α=3.s^-1; β=3.s^-1; α=3.s^-1;

Api and Aaj are values of phrase command and accent command amplitudes, T0i, T1j e T2j are onsets of phrase and word accent components, i and j, respectively and T2j is the offset of the word accent component j.