BASQUE INTONATION MODELLING FOR TEXT TO SPEECH CONVERSION

Eva Navas, Inmaculada Hernáez & Juan María Sánchez

Department of Electronic and Telecommunications
University of the Basque Country
{eva; inma; ion}@bips.bi.ehu.es

ABSTRACT

The present paper presents the modeling of standard Basque intonation to be used in text to speech conversion systems. The parameterization process of the Basque f0 curves made according to Fujisaki's intonation model is explained: experiments made in the placing of the phrase commands of the model are described and the results of these experiments are analyzed. The statistical analysis of the obtained parameters using classification and regression trees is described and the results obtained in this study are also explained.

1. INTRODUCTION

To get high quality text to speech (TTS) conversion, a good model of intonation is needed. Up to now, our TTS system AhoTTS [1] used a very simple model of intonation, which assigned peaks in the f0 curve to the stressed syllables and decreased lineally the value of f0 curve between them. The declination rules and the modeling of different types of sentences were also very basic.

To increase the quality of the synthesis, Fujisaki's model of intonation [2] was selected because it had already been used for the synthesis of intonation in many languages with successful results [3][4]. This model has been adapted for standard Basque language in this work and then introduced into AhoTTS.

The layout of the paper is as follows: next section describes the speech material utilized in the modeling of intonation. Then, in section 3 the way intonation labeling was made is explained. The statistical study of intonation parameters is described in section 4. Section 5 discusses the results and finally section 6 shows the conclusions of this work.

2. SPEECH MATERIAL

The validity of Fujisaki's intonation model for Basque was initially proven using a very small dialectal database [5]. However, to achieve the desired level of quality for text to speech synthesis in Basque, a more extended and complex database was needed. So, a new and more complete corpus was designed with this goal in mind.

This corpus was read by a native male Basque speaker in standard Basque in a laboratory environment. The resulting database was called *Jokin*, which stands for the name of the speaker. It comprises 344 isolated sentences with various syntactic structures, different lengths and diverse levels of complexity. The vocabulary used is very rich, with 1380 different words out of 2398. This database also includes special particles that could have a distinctive effect on intonation.

The main characteristics and the distribution of sentences in this database are shown in table 1.

Table 1: Main characteristics of Jokin database.

	Value
Size of recordings	64 Mb
# sentences	344
# declarative sentences	238
# question sentences	80
# exclamation sentences	26
# prosodic phrases	630
# words	2398
# different words	1380

The database was manually labeled at word level and then phone labels were automatically generated using speech synthesis and a dynamic time warping algorithm. Sentence and accent group labels were automatically generated using the information provided by the linguistic analysis module of the text to speech conversion system for Basque AhoTTS.

F0 curves were calculated with 1 ms precision, applying a method based in [6] and then decimated to have 5 ms precision, which was enough for the purpose of this work.

Breaks made by the speaker were also manually labeled and classified according to intonation criteria into two groups: breaks that should be modeled with a phrase command and breaks that should not. This was done by visually inspecting both synthetic and natural intonation curves, and classifying the break to achieve the best fit between them. The resulting distribution of breaks is shown in table 2 (where p.c. stands for phrase command). According to this distribution most of the breaks must be modeled with a phrase command (78%).

Table 2: Distribution of breaks in Jokin database.

	orthographic breaks	non-orthographic breaks
total #	169	117
with p.c.	140	83
without p.c.	29	34

3. LABELING INTONATION

The F0 curves were automatically parameterized according to Fujisaki's intonation model, with an algorithm based on analysis-by-synthesis. The best set of parameters is found

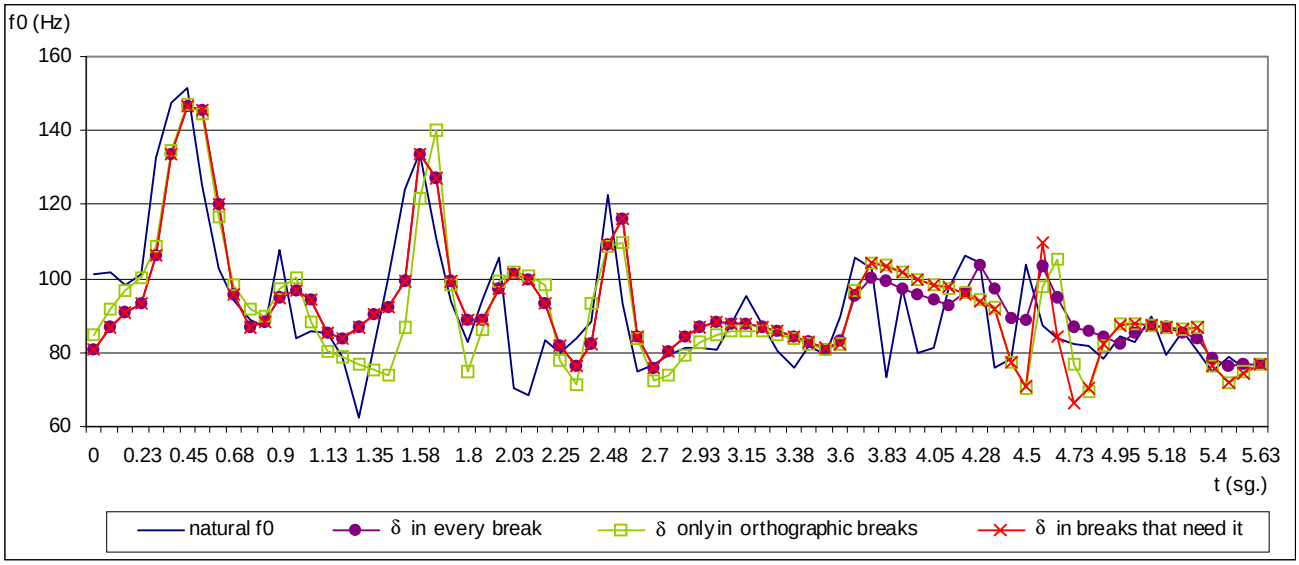

Figure 1: Result of the parameterization of the sentence "Erromako inperioa hain egin zen handi, ezen komunikazioa erabat zailtzen zuen", (The Roman empire grew so much, that the communications became very difficult) with three different placements for phrase commands.

performing an exhaustive search among the allowed combination of parameters. These allowed combination of parameters are selected according to certain linguistic constraints detailed in [5].

The accent commands were placed related with the position of the corresponding accent group and its duration was limited to vary in a certain range depending on the position of the stressed syllable within the accent group, as is also described in [5]. Only one accent command was used to model each accent group, except for the last accent group of questions and exclamation sentences, where an extra command has been added to model the final rise in the intonation curve.

3.1. Labeling experiments

To evaluate the importance of accurately placing the phrase commands in the Fujisaki's intonation model, the whole database was parameterized three times varying the locations and number of phrase commands:

- In the first experiment phrase commands were placed only in the breaks that were orthographically marked. This was the simplest experiment, because it implies that no model of break insertion is needed for intonation synthesis.
- The second experiment consisted in placing a phrase command at every break uttered by the speaker. This hypothesis means that a break insertion model is available at synthesis time.
- In the last experiment phrase commands were placed only in the breaks labeled as needing a phrase command. This indicates that both a break insertion model and a break classification algorithm are available.

Figure 1 shows the synthetic intonation curves corresponding to these three labeling experiments, related to the natural one for one of the sentences of the database. This sentence has only one orthographic break labeled as needing a phrase command and three non orthographic breaks, one of them labeled as not needing a phrase command. All the synthetic curves are very similar, with the best fit achieved by the case in which phrase commands are placed where needed.

3.2. Labeling results

After the whole database was parameterized under the three conditions, the root mean squared error (RMSE) obtained in each case was calculated and compared. Figure 2 shows the RMSE in all cases: the biggest error (11.09% or 10.5Hz at 95Hz) corresponds to the first experiment, where phrase commands are placed only in breaks indicated in text. The other two cases have smaller error, being the difference between them no meaningful (10.78% and 10.77%).

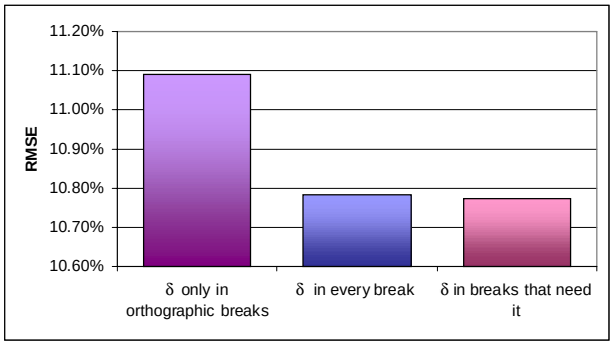

Figure 2: Comparison of RMSE in the three labeling experiments.

So, for the synthesis of intonation in Basque language using Fujisaki's model, a good model of insertion of breaks is

highly needed, but accurate classification of these breaks is not important. Considering these results, when synthesizing intonation one phrase command will be introduced for each prosodic phrase, not taking into account the type of the break.

4. PARAMETER ESTIMATION FROM TEXT

In order to get the appropriate synthetic intonation curve, the accent and phrase command parameters have to be related with characteristics extracted from input text.

Statistical analysis of the values of intonation parameters has been made using classification and regression trees (CARTs) [7], because they are able to manage variables of discrete and continuous nature, they automatically select the factors that have the greatest influence in the prediction of the target variable and they produce easy to read diagrams.

4.1. Prediction variables for accent commands

The information used for predicting accent command parameters is mainly related to the accent group whose intonation this accent command has to model. In particular, the variables provided to the trees are:

- Initial position, final position and duration of each accent group, measured in ms. and normalized to the duration of accent group and prosodic phrase.
- Order position of the accent group in the prosodic phrase, expressed both as an absolute quantity and relative to the total number of accent groups in the prosodic phrase, sentence and utterance.
- Total number of accent groups in the prosodic phrase, sentence and in the whole utterance.
- Total number of syllables in the current prosodic phrase, sentence and utterance.
- Position of the stressed syllable of each accent group, measured in ms. from the beginning of the prosodic phrase and from the start of the accent group. These variables are also given normalized to the duration of the prosodic phrase and accent group.
- Type of sentence which in the case of *Jokin* database can have the values of declarative, question, exclamation or pause sentence.
- Type of accent, which depends on the position of the stressed syllable within the accent group. This variable can have two different values: normal, if the stressed syllable is the second one and lexically marked if it is the first one.
- Type of accent command. Accent commands have been classified into three groups: last command of a question or exclamation sentence, penultimate command of a question or exclamation sentence and any other command. Two accent commands correspond to the last accent group of questions and exclamations, due to the introduction of an extra command to model the final rise in the intonation, so these commands could be different from the others.
- Index of accent command, which indicates the number of accent commands that are left until the end of the utterance.

Another aspect that has to be considered is the variable the trees are predicting: depending on the distribution of this variable, efficiency of the prediction varies. For the pulse parameters, amplitude is predicted without any transformation, but duration is normalized to the duration of accent group, and position is given relative to the beginning of the accent group and normalized to its duration.

4.2. Prediction variables for phrase commands

For phrase command amplitude prediction, the information used is related with the prosodic phrase whose intonation has to be modeled by this command. Specifically, the variables provided to the tree are:

- Order position of the prosodic phrase in the sentence, given both as an absolute number and relative to the number of prosodic phrases in the sentence.
- Total number of prosodic phrases in the corresponding sentence and utterance.
- Duration of the prosodic phrase, measured in ms. both as an absolute quantity and relative to the duration of the sentence and utterance.
- Duration of the utterance measured in ms.
- Total number of accent groups and syllables in the prosodic phrase, the sentence and in the whole utterance. The number of accent groups in the prosodic phrase is also given normalized to the total number of accent groups in the sentence and the utterance.
- Position of the first stressed syllable of the prosodic phrase, indicated from the beginning of the corresponding prosodic phrase, expressed as an absolute quantity and relative to the duration of the prosodic phrase.
- Type of sentence and utterance.

In this case, the variable predicted by the tree is directly the amplitude of the phrase command.

5. DISCUSSION

For each parameter of accent commands (amplitude, position and duration) a binary regression tree has been built. For the phrase commands only one tree was built to predict their amplitude, because their positions were directly related with the positions of the breaks: phrase commands were placed 323 ms before the beginning of the corresponding prosodic phrase.

The importance given by the trees to the different predicting variables are displayed in the figures contained in this section. The most important variable is given a 100% importance and the others are given values relative to that one. In all figures PP stands for prosodic phrase and AG for accent group.

Figure 3 shows the importance given by the tree to the different variables provided for accent command parameter prediction, when estimating accent command amplitudes. In this case, the type of sentence is the most influential factor, being the amplitude of commands bigger in questions and exclamations than in declaratives and pause sentences. Then type and index of accent command are considered with greater amplitudes assigned to ultimate and penultimate commands of questions and exclamations than to the others.

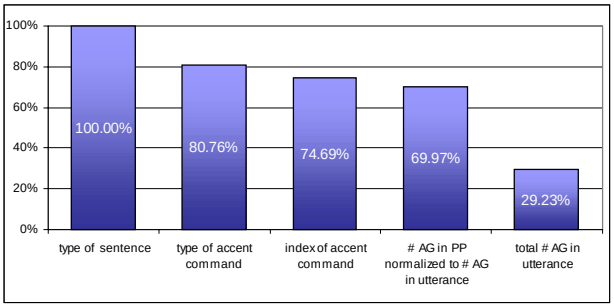

Figure 3: Variable importance in the prediction of accent command amplitude.

The importance of the variables when predicting accent command duration is shown in figure 4. The most important variable is number of syllables in the prosodic phrase and the number of syllables in the sentence has also great influence, being the predicted command longer for long phrases. Besides, the accent commands located towards the beginning of the prosodic phrase and the last commands of questions and exclamations are shorter than the others.

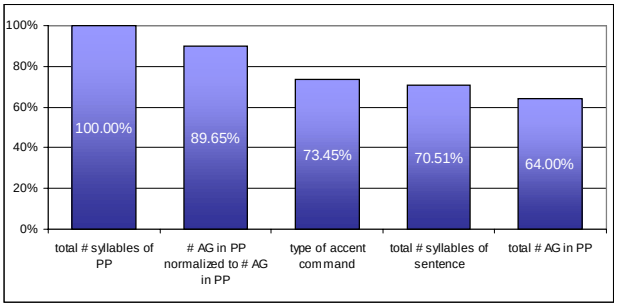

Figure 4: Variable importance in the prediction of accent command duration.

Figure 5 details the importance of the variables provided to predict accent command position. The most important variable is type of accent command: last commands of questions and exclamations are predicted farther from the beginning of the accent group. The rest of variables have little influence compared to this one.

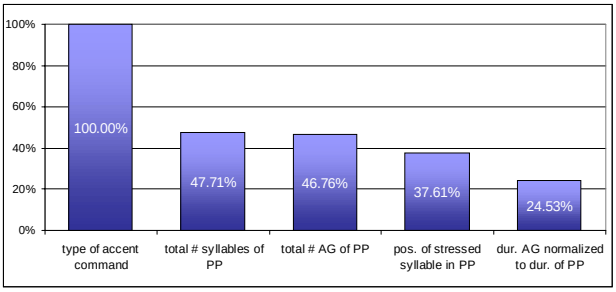

Figure 5: Variable importance in the prediction of accent command position.

The importance of the variables when predicting phrase command amplitude is displayed in figure 6. In this case, the most important variable is the order position of the prosodic phrase, being the command bigger for the first and second prosodic phrases than for the rest. The following three variables are related with the length of the prosodic phrase and indicate that the shorter the prosodic phrase is, the smaller the amplitude of the phrase command will be.

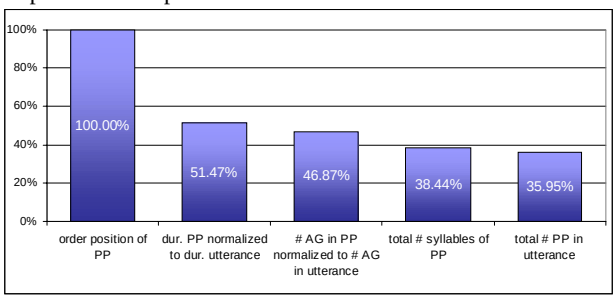

Figure 6: Variable importance in the prediction of phrase command amplitude.

6. CONCLUSIONS

Standard Basque intonation has been modeled according to Fujisaki's intonation model. Parameters of the model have been automatically calculated for a new corpus specifically designed with the purpose of modeling intonation. These intonation parameters have been related with linguistic characteristics of the corpus by means of binary regression trees and the model has been introduced in our TTS system AhoTTS, which is available for a demo at: http://bips.bi.ehu.es/tts/tts_en.html.

7. ACKNOWLEDGEMENTS

We acknowledge the financial support of Ministry of Science and Technology (grant TIC2000-1005-C03-03).

8. REFERENCES

[1] Hernáez, I.; Navas, E.; Murugarren, J.L.; Etxebarria, B., 2001. Description of the AhoTTS System for Basque Language. *4th ISCA Tutorial & Research Workshop on Speech Synthesis*.

[2] Fujisaki, H.; Hirose, K., 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of Acoustic Society. Jpn.* (E) 5, 4.

[3] Mixdorff, H.,1998. *Intonation patterns of German-model-based quantitative analysis and synthesis of F0 contours.* PhD Thesis. Technische Universitat Dresden.

[4] Wang, Ch.; Fujisaki, H.; Ohno, S.; Kodama, T., 1999. Analysis and synthesis of the four tones in connected speech of the standard Chinese based on a command-response model. In *Proceedings of Eurospeech'99*, Budapest, pp. 1655-1658.

[5] Navas, E.; Hernáez, I.; Armenta, A.; Etxebarria, B.; Salaberria, J., 2000. Modelling Basque intonation using Fujisaki's model and CARTs. *State of the art in speech synthesis digest, 3/1-3/6.*

[6] Griffin, D.; Lim, J. S., 1988. Multiband excitation vocoder. *IEEE Trans. ASSP.* Vol 36, N 8.

[7] Breiman, L.; Friedman, J.H.; Olsen, R.A.; Stone, C. J., 1984. Classification and Regression Trees. *Chapman & Hall.*