SUBJECTIVE EVALUATION OF SYNTHETIC INTONATION

Eva Navas, Inmaculada Hernáez, Juan María Sánchez

Department of Electronics and Telecommunication
University of the Basque Country
eva, inma, ion@bips.bi.ehu.es

ABSTRACT

This paper describes a method for evaluating the quality of synthetic intonation using subjective techniques. This perceptual method of assessing intonation, not only evaluates the quality of synthetic intonation, but also allows us to compare different models of intonation to know which one is the most natural from a perceptual point of view. This procedure has been used to assess the quality of an implementation of Fujisaki's intonation model for Basque language. The evaluation involved 30 participants and results show that the intonation model developed has introduced a considerable improvement and that the overall quality achieved is good.

1. INTRODUCTION

The perceived naturalness of synthetic speech strongly depends on the prosodic quality of the text to speech (TTS) system. One of the main problems when developing a TTS system is precisely the evaluation of the quality of the signal. There is a great concern for obtaining reliable systems to evaluate this quality [1][2], but most of the work has dealt with segmental intelligibility [3] and by the moment the evaluation of prosody remains unstandardized.

Mainly two types of evaluation processes are distinguished: objective and subjective methods, each one with its advantages and drawbacks [4]. Objective techniques are usually based on the comparison with a reference utterance [5] and subjective techniques require the collaboration of human experts and are based on listeners' reports. Among the subjective measures, mean opinion square (MOS) on a five point scale has been proposed as a standardized method by ITU-T [6]. Since a difference in intonation with a reference utterance does not necessarily mean that the synthetic one is incorrect, the use of perception tests is considered necessary.

This work was mainly developed to evaluate the Fujisaki's model of intonation [7] that had previously been adapted to Basque language [8]. This model has been inserted into AhoTTS, a modular Basque TTS system [9] that up to that time had a very simple model of intonation driven by rule.

The layout of the paper is as follows: next section describes the goals pursued with the developed evaluation process. Subsequently, in section 3 the design of the assessment process is explained. Section 4 discusses the experimental results of the test and finally section 5 presents the main conclusions of this work.

2. EVALUATION MAIN GOALS

The main goal of this work was to assess the naturalness and acceptability of the new intonation model developed, as well as to compare it with the previous intonation model of AhoTTS. To achieve this general goal, three questions have to be answered:

- How similar are the intonation curves produced by the model and the natural f0 curves: the model was created to match as exactly as possible the intonation of the training database. Although a difference between both curves does not mean that the synthetic intonation is inappropriate, having a measure of this similarity may be a good reference.
- How important is the amelioration (if any) when introducing the new intonation model. Developing a new intonation model is a high cost process and it is important to know if this effort has really improved the results.
- Which is the overall degree of acceptability of the developed model. To assess the general quality of intonation, evaluation of longer passages than sentences is needed [10]. Evaluating isolated sentences is not sufficient to know if the general quality obtained is suitable, because intonation curves that are judged acceptable when evaluated in isolation, may be considered boring or no natural when heard on longer passages.

In this work, subjective techniques were selected to get a perceptual evaluation of the synthetic intonation.

3. EVALUATION DESIGN

The evaluation process designed had to give an answer to each of the presented questions. Thus, the complete test created consisted in three different experiments:

- Experiment I compares sentences having natural and synthetic intonation to evaluate how different they are.
- Experiment II compares fully synthetic signals differing only in the intonation model used to produce the f0 curve.
- Experiment III assesses the general degree of acceptance of the synthetic intonation.

3.1. Stimuli generation

To perform experiment I, that is, to compare natural and synthetic intonation 15 sentences (containing between 10 and 22 syllables) were selected from the database used to train Fujisaki's intonation model. These sentences were carefully selected to have the same mean squared error (MSE) distribution than the whole database, therefore avoiding using in the test only the sentences that have been better fitted with the model. Figures 1 and 2 show respectively the distribution of the MSE made when parameterizing the intonation with Fujisaki's model in the whole database and the same distribution among the sentences selected for the test.

All these stimuli were created replacing the natural intonation by the synthetic curve, in the LPC coded version of the natural signal. Both signals were then decoded in the same way, to avoid differences in signal quality. Three of the stimuli consisted in a pair having the same element (natural or synthetic) repeated. Those pairs were used to control the validity of the participants to carry out this task.

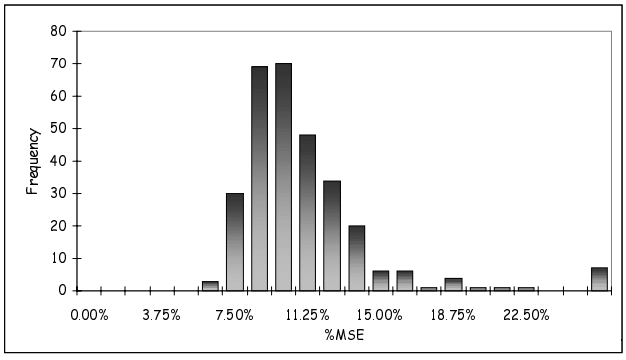

Figure 1: Distribution of %MSE over the mean value in the whole database.

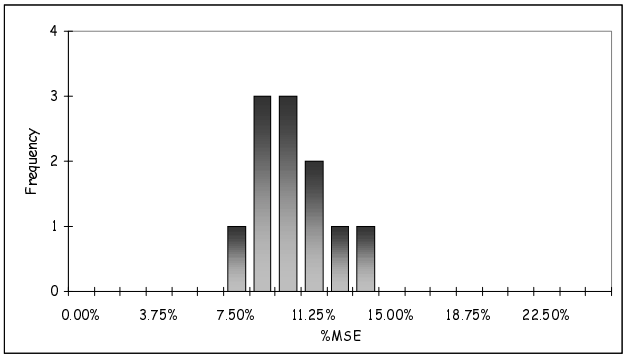

Figure 2: Distribution of %MSE over the mean value in the sentences selected for the test.

For experiment II and III both signals of the pair to be compared were created using AhoTTS varying only the intonation model applied. All the other prosodic characteristics (segmental duration, energy) were kept constant to allow the evaluation of intonation curve. Ten sentences from the same database were selected for experiment II (from 8 to 36 syllables). For experiment III four texts of different complexity (having between 45 and 83 syllables) were chosen from e-mails and a newspaper from Internet.

3.2. Evaluation process

The subjects taking part in the experiment were selected among the students and staff of the Electronics and Telecommunication Department of the University of the Basque Country. A total of 30 participants (17 males and 13 females with ages varying from 22 to 45 years) took part in the experiments. All of them were native of Basque, or at least fluent in standard Basque. None of them reported speech or hearing problems. Some of them were habituated to TTS systems, but none of them has a special phonetic training.

The tests were performed in the quasi silent environment of a research laboratory. Stimuli were presented to listeners over high quality headphones and reproduced with a standard Sound Blaster soundcard. They were created with a sampling rate of 8KHz for experiment I and 16KHz for experiment II and III and using 16 bits per sample in all cases.

The stimuli were presented to subjects by means of electronic forms that grouped five pairs of stimuli to be evaluated. Figure 3 shows a detail of the form used to evaluate the similarity between natural and synthetic intonation. Participants could hear the signals by clicking the adequate buttons and they had to score the pair in the corresponding combo box. Listeners could hear each stimulus as many times as they wanted and they had to score all the pairs presented in a form before starting evaluating the stimuli present in the following one. Once a form had been completed they could not return and modify

it. The order of the stimuli presented was randomized in all the experiments.

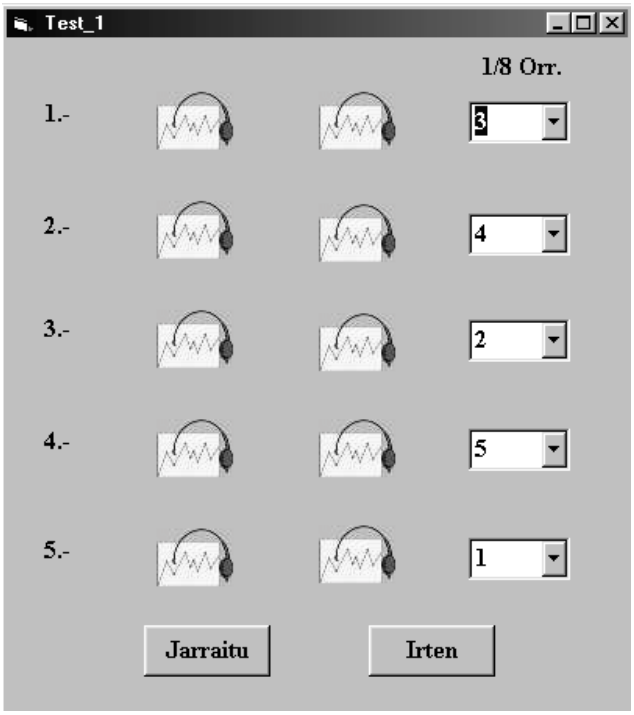

Figure 3: Interface program designed for intonation evaluation.

For experiment I, the participants were requested to evaluate the similarity between two intonations with grades varying between 1 (completely different) and 5 (completely similar). In experiment II they had to choose between two realizations of the same sentence the one that they preferred, paying attention only to their intonation or melody. They were asked to select the most natural utterance from each presented pair. For experiment III, they had to assess the quality of the stimuli with grades varying between 1 (very bad intonation) and 5 (very good intonation).

The complete test consisted in 8 forms and the duration of the test was about 10 minutes.

4. RESULTS

In the comparison between natural and synthetic intonation, the sentences were judged reasonably similar, having a scoring of 3.17 (variance 0.24). The distribution of scores is displayed in figure 4: the majority of the subjects (62%) selected a scoring of 3 and 4.

All the participants scored with 4 or 5 points the pairs that had the same sentences repeated, so none of them was excluded from the calculation of results.

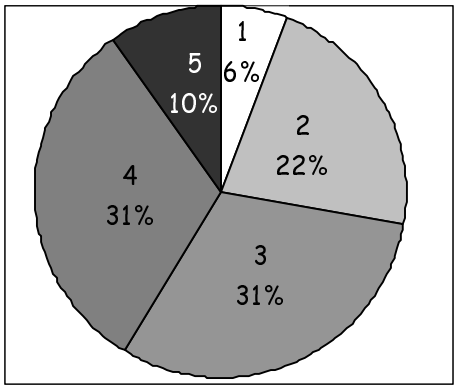

Figure 4: Distribution of the scoring in experiment I.

In experiment II, Fujisaki's model was preferred by large with 72.2% of selections. Figure 5 presents the percentages of each model separated by sentence: sentences 7 and 9 have the smallest difference between the two models, probably due to the fact that they were the shortest sentences in the test.

In experiment III, texts generated with Fujisaki's model were in all cases better scored than the ones created with the former intonation model of AhoTTS. Figure 6 shows the distribution of scores given by subjects to the texts: first part displays the results corresponding to the former model and second part to those of Fujisaki's model. None of the participants considered that the texts synthesized with the previous model had very good intonation, while 15.79% scored them with a 5 when synthesized with Fujisaki's model. The synthetic intonation created by rule received a score of non acceptable (1 or 2) in a 50% of the cases, quantity that reduces to only 13.3% when scoring the one created with Fujisaki's model.

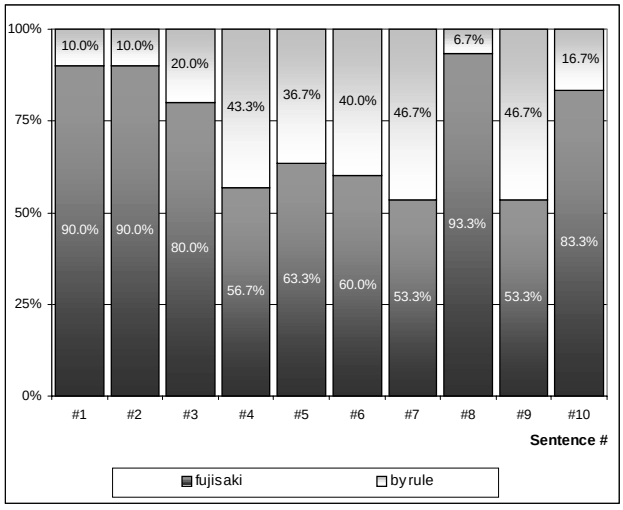

Figure 5: Preference of the two models of synthetic intonation separated by sentence.

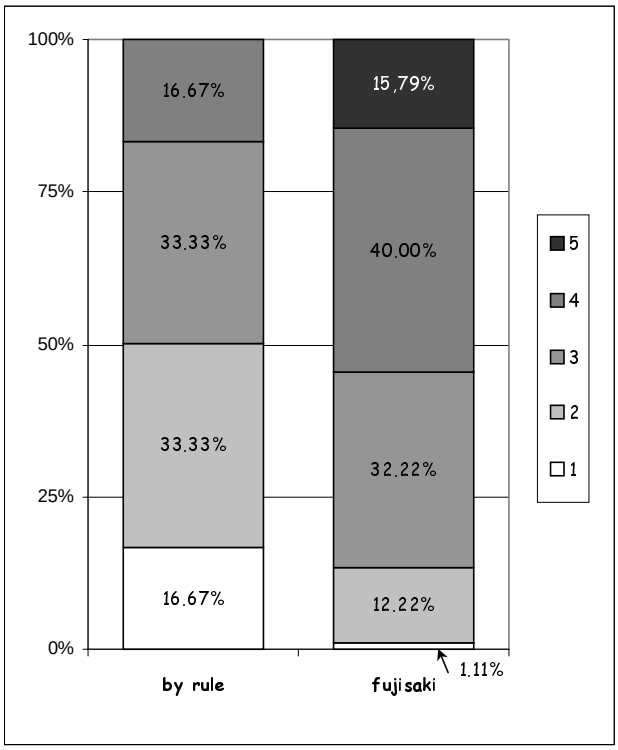

Figure 6: Distribution of scores in experiment III.

Figure 7 displays the results obtained for each text: Fujisaki's model is always scored over the mean value of 3, which is not the case of the former model judged inadequate in all cases.

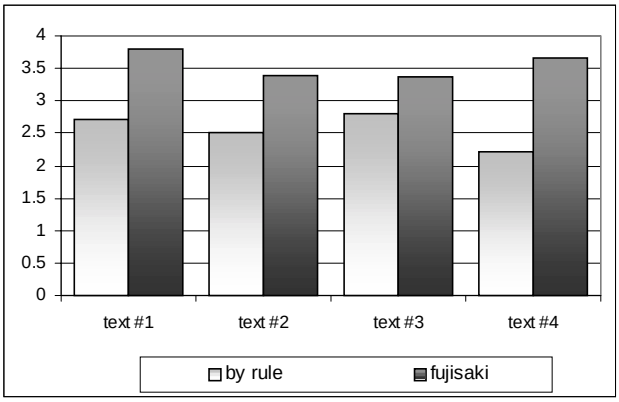

Figure 7: Scoring of the two intonation models separated by text.

5. CONCLUSIONS

A subjective evaluation of synthetic intonation has been carried out and all the evaluation goals have been met. Fujisaki's model has been judged as sufficiently well fitted to natural curves and has been preferred in all cases to the former intonation model used in AhoTTS, achieving a 72.2% of selections. In the general acceptance test Fujisaki's model has also attained the best scorings with a mean of 3.5 points against the 2.5 given to the former model.

The evaluation process developed does not depend on the model of intonation used, and can be extended to assess the quality of any synthetic model of intonation.

6. ACKNOWLEDGEMENTS

The authors would like to thank all the subjects for their participation in the experiments.

We also acknowledge the financial support of the Ministry of Science and Technology (TIC2000-1005-C03-03 and TIC2000-1669-C04-03).

7. REFERENCES

[1] K. Morton, "Expectations for Assessment Techniques Applied to Speech Synthesis", *Proceedings of the Institute of Acoustics*, Vol. 13, 1991.

[2] A. Di Cristo, P. Di Cristo, J. Véronis, and E. Campione, "A model of prosody for French text-to-speech synthesis", *Intonation: Models and Theories*, Kluwer Academic Publishers, Berlin, pp. 321-355, 2000.

[3] A. Fourcin, "Assessment of synthetic speech", *Talking Machines: Theories, Models and Designs*. Elsevier Science, Amsterdam, pp. 431-434, 1992.

[4] D. Hirst, A. Rilliard, and V. Aubergé, "Comparison of subjective evaluation and an objective evaluation metric for prosody in text-to-speech synthesis", *3rd ESCA/COCOSDA Workshop on Speech Synthesis*, Blue Mountain, Australia, 1-4, 1998.

[5] R. A. J. Clark, and K. E. Dusterhoff, "Objective methods for evaluating synthetic intonation", *Eurospeech'99*, Budapest, pp. 1623-1626, 1999.

[6] ITU-T Recommendation P.85, "A method for subjective performance assessment of the quality of speech voice output devices", study group 12, 1994.

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

[8] E. Navas, I. Hernáez, A. Armenta, B. Etxebarria, and J. Salaberria, "Modelling Basque intonation using Fujisaki's models and CARTs", *State of the art in Speech Synthesis digest*, London, 3/1-3/6, 2000.

[9] I. Hernáez, E. Navas, J. L. Murugarren, and B. Etxebrria, "Description of the AhoTTS System for Basque Language", *4th ISCA Tutorial & Research Workshop on Speech Synthesis*, Edinburgh, pp. 151-154, 2001.

[10] H. Mixdorff, and D. Mehnert, "Exploring the Naturalness of Several German High-Quality-Text-to-Speech Systems", *Eurospeech'99*, Budapest, pp. 1859-1862, 1999.