Comparing CART and Fujisaki Intonation Models for Synthesis of US-English Names

Marko Moberg & Kimmo Pärssinen

Audio-Visual Systems Laboratory
Nokia Research Center, Tampere, Finland
{marko.moberg; kimmo.parssinen}@nokia.com

Abstract

In this work two different speech synthesis intonation models were compared against a reference created with natural intonation. The models chosen were direct classification and regression tree (CART) based pitch estimation and simple implementation of Fujisaki model. The performance and the suitability of the models for low-footprint name synthesis were evaluated by carrying out a listening test. The results of the test indicated that the perceived quality of the intonation generated by the models was equal to the natural intonation reference. Despite the differences in the models they both offer a viable, high quality solution for intonation modeling of US-English names. The results may also apply to other languages and to the case of isolated word synthesis.

1. Introduction

Changes in fundamental frequency, *F0,* of human speech can be interpreted as changes in pitch or intonation. Pitch is sometimes regarded as a segment e.g. syllable level feature whereas intonation can be viewed as a suprasegmental feature of speech. Different intonation patterns are partially due to physiological properties of speech production and also due to linguistic factors such as lexical stress and semantics. Intonation is a significant contributor to naturalness of speech. It makes spoken utterances easier to understand, augments them and may even convey paralinguistic information. [1][2]

As in natural speech, intonation plays a very important role in speech synthesis. The perceived quality of synthetic speech is largely determined by the intonation generated during the synthesis. Even if the segmental speech sounds in synthesis were flawless replicas of human speech, insufficient modeling of intonation would make the perceived speech unnatural.

In text-to-speech (TTS) systems, intonation among other prosodic aspects must usually be generated from the plain textual input. The text may be analyzed automatically e.g. by parsing it linguistically and/or dividing it into tone groups [1]. The main challenge is to provide meaningful input for the particular intonation model that is used. The link between the text and the intonation can be obtained by a set of hand-written rules or by some data-driven methods, which rely on statistically predictable dependencies between linguistic (and phonological) features and intonation model parameters. Needless to say, rules and relationships between features are usually language dependent.

The limited domain TTS systems such as synthesis of names and isolated words cannot rely on complex linguistic analysis for intonation modeling. Simpler approaches are mandatory especially in low-footprint systems, which are targeted at embedded devices with limited memory resources.

This paper compares two different low-complexity methods for automatic intonation generation against the natural intonation extracted from the recordings. The aim of the study was to find out and quantify the perceptual differences between the two models and the natural intonation. The scope was limited to the synthesis of names (first name + last name) without evaluating the performance of intonation models in full sentence synthesis like in some earlier studies [3]. The text-to-speech system used in the tests was a low footprint system based on Klatt88 formant synthesizer.

2. Intonation modeling

There are many different approaches to model intonation in speech synthesis. One classification divides various models into acoustic, perceptual and linguistic models. Acoustic models aim to reproduce the intonation patterns in a compact way. The perceptual models concentrate on those intonation events, which are the most relevant perceptually. The linguistic approach treats pitch patterns as a part of the linguistic structure. Intonation events can be described by functional prosodic units and modeled by e.g. limited set of pitch contours or tone sequences. [4]

The main focus in this paper is on data-driven methods, which can be used together with various intonation models. Methods such as using of classification and regression trees (CART) [5] can capture a good amount of statistical variation into the model with small memory consumption. Significant features and dependencies for intonation modeling can be easily extracted and estimated without writing detailed linguistic rules by hand. CARTs have been used with many different intonation models including the tilt model [6] and PaIntE system [7]. In the above-mentioned systems, CARTs provide estimates for features e.g. accent location or type, which are used by the intonation model. CARTs can also be used directly to estimate the actual *F0* values like in the Festival synthesis system [8].

The models for our study were selected due to their compactness and assumed suitability for isolated word synthesis. The chosen models are the Fujisaki-model [9] and direct CART based *F0* estimation [10]. In both cases only the input phoneme sequence is provided together with the durations extracted from the natural recordings. The syllabification and stress assignment were carried out using CARTs trained with 520 annotated US-English utterances mostly consisting of short phrases and names.

2.1. Natural intonation

In this work real intonation was used as a reference to the other two intonation models. A simple block diagram of the algorithm used in extracting the natural intonation is shown in

Figure 1. Therefore, to obtain real pitch contours an implementation of the Robust Algorithm for Pitch Tracking (RAPT) was used to extract fundamental frequency from the 8kHz recorded NIST sound files [11]. The algorithm was set to output a pitch estimate in every 10ms, and after extraction a median filter of length seven was applied to smooth the *F0* contours removing some isolated peaks created by the algorithm.

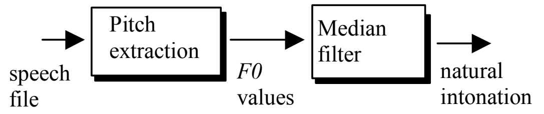

Figure 1. Natural intonation model.

For each name the process produced approximately 70 pitch points that were used in obtaining the pitch contour during synthesis. When the final parameter frames were created for the formant synthesizer, a simple linear interpolation method was used for filling the gaps between the fixed *F0* points.

2.2. CART based intonation

CARTs were used to encode and predict the location of the syllable, syllable stress and, in case of the CART based intonation also syllable pitch. The trees were trained using a variety of different level features extracted from the annotated training set. The training set consisted of 520 utterances including names, single words and short phrases. All the utterances were spoken by the same US-English male speaker.

The features used in the tree training were classified into different levels. The lowest level of features represented phonetic information about the individual phonemes. This included e.g. phoneme type (vowel or consonant), voicing and manner of articulation (stop, fricative, etc.). Second level represented segment level information such as phoneme label/id and position within a syllable. Finally, third level of features consisted of e.g. size of the onset and coda and syllable position within a word. When CARTs were used for predicting the intonation, a total of two pitch points per syllable were obtained from the trees. The first point represented the pitch value in the middle of the syllable and the second at the end of the syllable. Again, linear interpolation was used to create the final pitch contours that were given to the formant synthesizer.

In our implementation, the total size of the two intonation trees (mid-syllable and end syllable) was 328 bytes. The code overhead was quite minimal since the system already supported CART handling for syllable boundaries and accents.

2.3 Fujisaki model intonation

Fujisaki model is an analytical model for controlling the fundamental frequency variations [12]. This model has been successfully tested on many languages and it is capable of producing close approximations of the real *F0* contour. The model uses two kinds of inputs: phrase commands (impulses) and accent commands (stepwise functions). Hence, the final pitch contour can be regarded as a result of superposition of local, syllable level, and global, phrase level factors. Figure 2 shows the basic configuration of the Fujisaki model. In the model $g_p(t)$ is an impulse response of the phrase control mechanism and $g_a(t)$ is a step response function of the accent control mechanism. Both functions are assumed to be second-order linear systems and the final *F0* contour is calculated as a sum of their outputs[9][13][14].

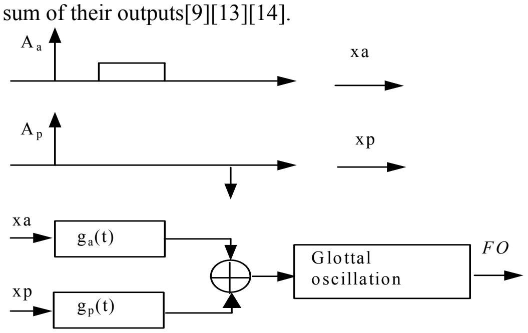

Figure 2. Fujisaki model. A_p is the magnitude of the phrase command, A_a is the magnitude of the accent command, xa is the accent command and xp is the phrase command input.

The above system can therefore be described using the following equations[9][12][14]:

$$\ln(F0) = \ln(Fb) + \sum_{k=1}^{N_p} A_{p,k} g_p(t - T_{p,k}) + \sum_{k=1}^{N_a} A_{a,k}(g_a(t - T_{a,k}) - g_a(t - T^*_{a,k})) \qquad (1)$$

where

$$g_g(t) = \alpha^2 t e^{(-\alpha t)}, t \geq 0 \qquad (2)$$

and

$$g_a(t) = \min(1 - (1 + \beta t)e^{(-\beta t)}, \gamma), t \geq 0\ . \qquad (3)$$

The symbols used in equations (1), (2) and (3) are

Fb:	asymptotic value of the fundamental frequency,
N_p:	number of phrase commands,
N_a:	number of accent commands,
$A_{p,k}$:	amplitude of the kth phrase command,
$A_{a,k}$:	amplitude of the kth accent command,
$T_{p,k}$:	time instant of the kth phrase command,
$T_{a,k}$:	onset of the kth accent command,
$T^*_{a,k}$:	end of the kth accent command,
α:	natural angular frequency of the phrase control mechanism,
β:	natural angular frequency of the accent command mechanism,
γ:	relative ceiling level of accent components (= 0.9).

Recently different methods have been proposed for automatic extraction of Fujisaki model parameters from the speech data. During synthesis, it is then possible to use e.g. CARTs in estimating model parameters from the linguistic information extracted from the input text. [15][16][17]

In this work magnitudes of phrase and accent pulses and also both α and β were hand-tuned and assumed to be constant within all synthesized utterances. In synthesis, a positive accent pulse was created for every accented syllable.

The pulse was summed to the phrase pulse and to the asymptotic value of fundamental frequency resulting in the necessary variations to the pitch contour. Moreover, at the end of each name a small negative phrase pulse was also applied to bring down the *F0* contour.

The implementation of the Fujisaki intonation model did not require any significant memory for constant data. However, the model is computationally more complex than the CART based solution due to the usage of exponential functions. The complexity is mainly an issue in fixed point implementation of the model.

3. Test set-up

A set of 20 arbitrary English names was selected for the test. Each name was recorded by a native US-English male speaker and the speech data was annotated. The annotation process included only the creation of the phonetic transcription of the utterances and marking of the phoneme boundaries. The transcribed sequences of phonemes were fed into a formant synthesis system, which was used in comparing different intonation models. The syllabification and stress assignment for each name was carried out automatically by the synthesis system using CART based prediction. The 20 names used in this test were not included in the larger data set used in training the CARTs. The segment durations i.e. durations of each phoneme were taken from the hand-written annotations of the recordings and they were used in all the three test conditions: 1) natural intonation, 2) direct CART based intonation and 3) Fujisaki model intonation. Figure 3 presents the different test conditions used in this work.

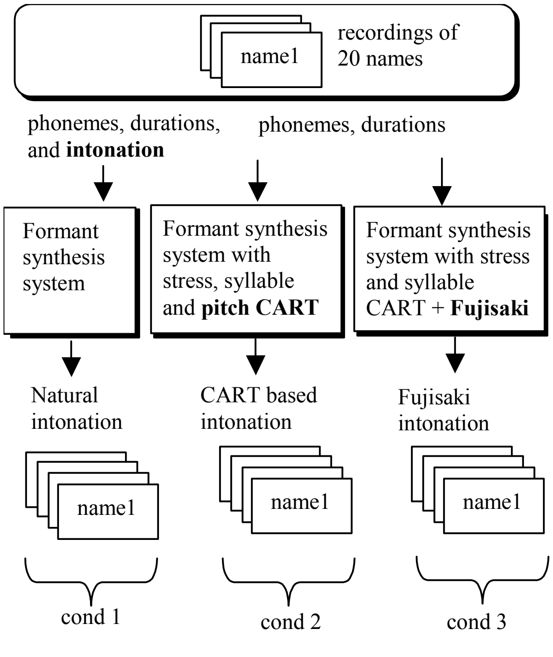
name1
recordings of 20 names
phonemes, durations, and **intonation**
phonemes, durations
Formant synthesis system
Formant synthesis system with stress, syllable and **pitch CART**
Formant synthesis system with stress and syllable CART + **Fujisaki**
Natural intonation
CART based intonation
Fujisaki intonation
name1
name1
name1
cond 1
cond 2
cond 3

Figure 3. Test arrangement for comparison of two different intonation models against natural intonation.

The names were synthesized using a male voice and the data was presented in 16-bit format using a sampling rate of 16 kHz. The mean pitch of the "synthetic speaker" was adjusted to match the one of the natural speaker used in the test condition one (natural intonation).

The comparison of the different intonation models was performed using a standard MOS (Mean Opinion Score) test where test subjects rated each audio sample on a scale from 1 to 5 (1=bad, 2=poor, 3=fair, 4=good, 5=excellent). Each of the twenty names were synthesized with all three intonations and played in a random order. The 60 unique test samples were also repeated once to make the test results more reliable. The total number of samples rated by each test subject was thus 120.

There were a total of seven native US-English speakers who took part in the test. The test subjects were asked to concentrate on the quality of intonation rather than on the quality of the synthesis itself. The actual test was carried out using dedicated MOS listening test software running on a personal computer. After the test, the results were analyzed statistically thus obtaining an average MOS scores for each intonation model. The validity of the results was verified by calculating 95% confidence intervals.

4. Results

The average MOS ratings for each intonation scheme are shown in Figure 4. The top region of each bar represents the 95% confidence interval. That means that with 95% confidence the average of each MOS rating lies in that region.

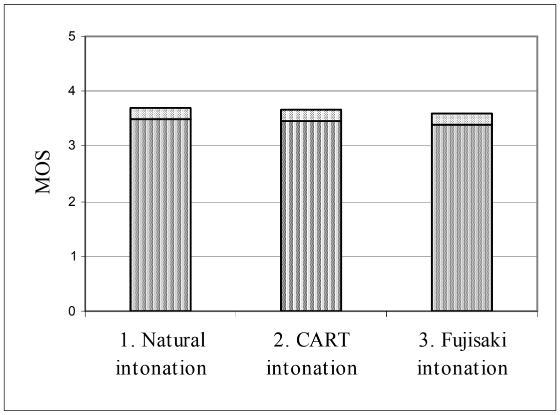

Figure 4. Listening test results comparing 20 English names using Fujisaki intonation, CART based intonation and natural intonation.

The results of the listening test showed that no significant differences were perceived between the intonation models in context of name synthesis. Each model received a rating of about 3.5 on a MOS scale. Although the intonation contours of different test conditions were audibly different, the differences did not result in a degradation of perceived quality. According to some listeners the Fujisaki samples were the only ones, which were in some cases consistently different from samples generated with other models. The Fujisaki model produced a very good quality intonation for the first name but the second name (family name) remained quite flat in some cases. However, the ratings of the Fujisaki model were still comparable with the natural intonation although further tuning of the Fujisaki parameters could have improved the intonation even further.

5. Discussion

This study was limited only to intonation models for synthesis of US-English names. The recordings used in training the CART intonation model included also isolated words and short phrases in addition to two part names. Because of that, it is fair to assume that the same results would be valid for any isolated word synthesis. Since the CART based method is data-driven, the results of this test can be extended to other languages as well. The parameters of the Fujisaki model were hand-tuned for the US-English names but they performed well also for single, isolated words. It is well known that Fujisaki model has been used in modeling intonation of many different languages so the results of this test concerning Fujisaki intonation should also quite well apply to other languages.

The test results showed that no significant differences between the models were detected in a random order MOS test. One reason for this might be that the low footprint synthesis system applied in this test is relatively low quality making it difficult to differentiate the intonation models. However, in the future, some additional information about the subtle differences might be obtained by carrying out a pair comparison test. The comparison of other aspects of prosody such as duration could be easily implemented by replacing the natural duration used in this test with the duration estimated using dedicated duration CARTs.

6. Conclusions

The results presented in this paper show that the intonation of synthesized names in US-English can be successfully generated by direct CART based method or by a Fujisaki intonation model. In both cases, the quality of the intonation was perceived as equal to the natural intonation extracted from the recordings of a native US-English speaker.

The advantages of the CART based method are, for example, low complexity implementation and automated data-driven training without the extensive hand tuning. But on the other hand, the CART based method requires large amount of training data (recordings) and the actual synthesizer must allocate some memory for storing the trees. Another drawback of the method is the statistical nature, which might in some cases produce inconsistent intonation contours. Also the irregularities in intonation of some words might be problematic.

The Fujisaki model is very flexible but it requires the adjusting of the parameters. Automated methods exist but they were not used in this study. Fujisaki intonation guarantees smooth, deterministic and well controlled intonation contours. Fixed-point implementation of the Fujisaki model is somewhat complex due to the usage of exponential functions but there is no constant data to increase the memory consumption.

If only the perceived quality is considered, the direct CART based intonation model and Fujisaki model are both viable solutions for modeling the intonation in a low-footprint, limited domain text-to-speech system. It is fair to assume that the results also apply to any isolated word synthesis, not just names. Furthermore, the intonation models should perform in the same manner in other languages as well provided that the CARTs are trained using proper data and Fujisaki parameters are tuned accordingly.

7. References

[1] Ball, M. J.; Rahilly J., 1999. *Phonetics, the science of speech*. New York, USA: Oxford University Press Inc, 101-122.

[2] Keller, E. (ed.), 1994. *Fundamentals of speech synthesis and speech recognition*. West Sussex, England: John Wiley & Sons Ltd, 23-40.

[3] Syrdal, A.; Möhler, G.; Dusterhoff, K.; Conkie, A.; Black, A., 1998. Three Methods of Intonation Modeling, In *Proceedings of the 3rd ESCA Workshop on Speech Synthesis, Jenolan Caves, Australia.* 305-310.

[4] Dutoit, T., 1997. *An introduction to text-to-speech synthesis*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 129-174.

[5] Breiman, L. et al., 1984. *Classification and Regression Trees*. Belmont, CA, USA: Wadsworth Inc.

[6] Dusterhoff, K. E.; Black, A.; Taylor, P., 1999. Using Decision Trees within the Tilt Intonation Model to Predict F0 Contours, In *Proceedings of the 6th European Conference on Speech Communication and Technology, Eurospeech 1999, Budapest, Hungary*. Vol 4, 1627-1630.

[7] Möhler, G.; Conkie, A., 1998. Parametric modeling of intonation using vector quantization. In *Proceedings of 3rd ESCA Workshop on Speech Synthesis, Jenolan Caves, Australia*. 311-314.

[8] Taylor, P.; Black, A.; Caley, R., 1998. The Architecture of the Festival Speech Synthesis System, In *Proceedings of the 3rd ESCA Workshop on Speech Synthesis, Jenolan Caves, Australia.* 147-151.

[9] Fujisaki, H.; Hirose, K., 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of Acoustical Society of Japan*, 5(4), 233-241.

[10] Black, A.; Taylor, P.; Caley, R., 1999. *The festival speech synthesis system, system documentation*, Centre for Speech Technology Research, University of Edinburgh..

[11] Kleijn, W.; Paliwall, K., 1995. *Speech Coding and Synthesis*, Lausanne: Elsevier

[12] Fujisaki, H., 1993. From information to intonation In *Proceedings of the 1993 International Symposium on Spoken Dialogue*, 7-18.

[13] Mixdorff, H.; Fujisaki, H., 1995. A Scheme for a Model-based Synthesis by Rule of F0 Contours of German Utterances, In *Proceedings of the 4th European conference on speech communication and technology, Eurospeech 1995, Madrid, Spain*. Vol 3, 1823-1826.

[14] Fujisaki, H; Ohno, S., 1995. Analysis and modeling of Fundamental Frequency Contours of English Utterances, In *Proceedings of the 4th European Conference on Speech Communication and Technology, Eurospeech 1995, Madrid, Spain*. Vol 2, 985-988.

[15] Mixdorff, H., 2000. A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters, In *Proceedings of ICASSP 2000, Istanbul, Turkey*. Vol 3, 1281-1284.

[16] Navas, E.; Hernaez, I.; Armenta, A.; Etxebarria, B.; Salaberria, J., 2000. Modelling Basque intonation using Fujisaki's model and CARTs, *State of the Art in Speech Synthesis*, London, UK.

[17] Rossi, P.;Palmieri, F.; Cutugno, F., 2002. A Method for Automatic Extraction of Fujisaki-Model Parameters, In *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France, April.