Training the Tilt Intonation Model using the JEMA methodology

Matej Rojc[1], *Pablo Daniel Agüero*[2], *Antonio Bonafonte*[2], *Zdravko Kacic*[1]

[1]Faculty of Electrical Engineering And Computer Science, University of Maribor, Slovenia
[2]Technical University of Catalonia, Barcelona, Spain

Abstract

This paper focuses on the estimation of the Tilt intonation model [1]. Usually, Tilt events are detected using a first estimation which is improved using gradient descent techniques. To speed up the search we propose to use a closed form expression for some of the Tilt parameters. The gradient descent search is used only for the time related parameters because a close expression cannot be found. Furthermore, the original Tilt proposal estimates the Tilt events sentence by sentence. Here we propose to estimate the events of the whole training corpus at the same time, using what we call the *JEMA* methodology. This approach increases the consistency of the estimation producing better intonation models. It has been tested on two different languages: Slovenian and Spanish. The experimental results reveal that the Tilt model is appropriate for these languages and that the JEMA methodology produces better prosodic models.

1. Introduction

During last decade, text-to-speech systems have experienced a formidable quality improvement. The main reasons are the quality of the acoustic generation module when speech segments are selected from large speech databases and the improvement of the prosodic modules using data-driven models inferred from large databases.

However, when listening to a long speech fragment, there is still no doubt whether the voice is synthetic or natural. In order to achieve this final goal, prosodic models play a fundamental role.

Several intonation models have been proposed in the literature, such as Fujisaki [2], Tilt [1], Bézier [3] and INTSINT [4]. In general the training of those modules consists of two stages. First, a compact representation is obtained for each sentence, e.g.: Fujisaki commands or Tilt events (step 1). After that, machine learning techniques are used to infer *rules* that map the linguistic features (available during speech synthesis) to the parameters (e.g.: Tilt parameters) using the whole corpus (step 2). The complete process is shown in figure 1. Such models are named in this paper as two-stage algorithms. The application of these models is straightforward: given the text, the linguistic features are derived. Then, the rules are applied to these linguistic features in order to obtain the corresponding parameters. Finally, these parameters are used to synthesize the fundamental frequency contour.

In previous papers [5, 6] we presented *JEMA: Join feature extraction and modeling approach*, a new approach to train intonation models that combines the parameter extraction (step 1) and model generation (step 2) into a single loop. This approach avoids requirements of two-stage approaches such as continuity of fundamental frequency contours (unvoiced regions need to be interpolated) and increases the consistency of the extracted parameters.

This work has been partially sponsored by the European Union under grant FP6-506738 (TC-STAR project, http://www.tc-star.org) and the Spanish Government under grant TIC2002-04447-C02 (ALIADO project, http://gps-tsc.upc.es/veu/aliado

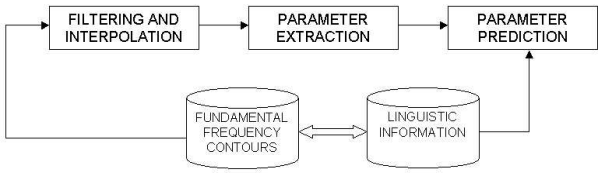

Figure 1: *Two-stage approach*

In this paper we propose the application of *JEMA* to the Tilt intonation model. Furthermore, we derive a closed-form solution to derive some of the Tilt parameters and speed-up the whole procedure.

The method is evaluated using two different languages: Slovenian and Spanish. The results show that for both languages the Tilt model and the proposed methodology perform well.

The paper is organized as follows. Section 2 introduces the *JEMA* and compares it with the two-stage approach. Section 3 reviews the baseline two-stage algorithm for training the Tilt intonation model. Then, section 4 shows the closed-form formulation and section 5 develops a method to apply *JEMA* to the Tilt model. In order to evaluate the method, section 6 contains the results of the intonation model. This evaluation has been done in two languages, Slovenian and Spanish, showing that the Tilt model is appropriate to describe intonation in both languages, selecting the appropriate Tilt events. Finally, section 7 summarizes the paper and provides the conclusions of this work.

2. JEMA: Joint feature extraction and modeling

As stated in the introduction, most of the data-driven intonation models are estimated using two stages. This procedure presents some characteristics that can cause some training problems:

- **Interpolation of fundamental frequency contour.** An initial interpolation of f0 in the unvoiced regions is required. As this interpolation is somehow arbitrary, this may introduce *noise* in the extracted parameters: contours with the same F0 contour in the voiced parts may be represented by different parameters. This introduces dispersion in parameters reducing the accuracy of the machine learning techniques.
- **Multiple solutions.** In some intonation models, different sets of parameters can represent the $f0$ contour with

the same accuracy. Again, this increases the variance of the parameters and reduces the accuracy of the machine learning techniques.

- **Sentence by sentence extraction.** Sentence by sentence parameter extraction lacks general information about the intonation of the language. The f0 contour is noisy, in the sense that it is affected by micro-melody, errors in the measure, etc. To solve this the f0 contours are usually filtered before computing the parameters of the model. But this filtering is somehow arbitrary. The knowledge of all the other sentences could be used as *a priori* probability to derive the underlying parameters.

The *JEMA* combines parameter extraction and model generation into a single loop. The model generation is performed using machine learning techniques that cluster segments of F0 contours from the training databases. Each cluster is considered as a class that is approximated by a set of parameters given the intonation model, e.g.: Tilt parameters. The parameter extraction is performed using an optimization algorithm that finds the global parameters that best represent all the contours of the cluster. As the same parameters are used for all the contours, stylization or unvoiced interpolation is not required. The extracted parameters are more consistent and the prediction capabilities of machine learning techniques that are used to generate a model are improved. Figure 2 shows the scheme of joint approach. Further details of the application of *JEMA* to the Tilt intonation model are given in section 5.

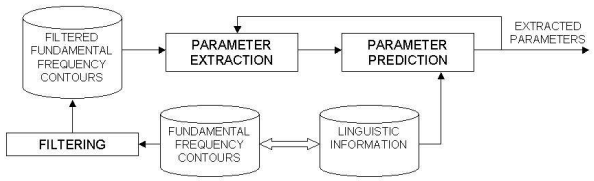

Figure 2: *Joint approach*

3. Review of the Tilt intonation model

The Tilt intonation model defines intonational events which are represented by a set of curves. Intonational events are phrase breaks, accents, etc. The curves are piecewise and they are composed of a rise and a fall component. Each curve is connected with the adjacent curves by a line. The way the model approximates the fundamental frequency contour gives the name to Tilt parameters: rise, fall and connection (*RFC* parameters). In particular, the *RFC* parameters for a Tilt event are (see figure 3): rise amplitude (A_r), rise duration (t_r), fall amplitude (A_f), fall duration (t_f), position (t_e) and F0 height ($f0_{offset}$).

The intonation model requires detecting the Tilt events in the database, e.g.: phrase breaks events and accent events. This task can be performed using knowledge based rules or automatically derived rules. For instance, in [7], HMM are used to detect Tilt events.

After this preliminary process, the Tilt acoustic parameters are extracted from the sentences of the database (step 1). This task can be performed using gradient descent techniques or the method proposed in section 4.

In the second step, the linguistic information present in the sentences of the database is transformed into linguistic features and used for construction of the model. In this work, binary

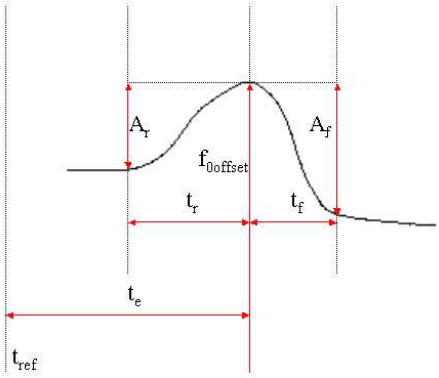

Figure 3: *Tilt parameters*

regression trees (CART) are used to predict the parameters of Tilt intonation events. The Tilt intonation model provides the mapping between linguistic features of the sentences and the parameterization of the fundamental frequency contour. Therefore it is possible to get linguistic features of new sentences and predict a suitable fundamental frequency contour.

The two-stage approach serves as a baseline for the evaluation of the novel algorithm based on *JEMA*, which is presented in section 5.

4. Closed-form determination of amplitude parameters

In Tilt intonation model it is not possible to obtain a closed-form solution for all the parameters of the model. However, it is possible to obtain the optimal solution for the amplitude parameters and $f0_{offset}$ assuming that the time instants are known.

The optimal values of the time instants can be found using grid search or gradient descent techniques. The time instants remain constant during closed-form amplitude optimization, and the amplitude values are kept constant during time instant optimization using gradient descent techniques. The update loop is shown in figure 4.

We must point out that the loop that combines closed-form determination of some parameters and gradient descent of the other parameters has a better convergence rate. This optimization procedure is used both in the two-stage intonation model and in the joint approach presented in next section.

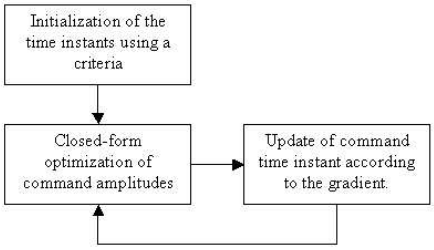

Figure 4: *Update loop*

The closed-form formulation is obtained by minimizing the mean square error:

$$e^2 = (f0 - \hat{f0})^T (f0 - \hat{f0}) \quad (1)$$

where $\hat{f0}$ is a function of all RFC parameters of each event:

$$\hat{f0} = f(A_r, t_r, A_f, t_f, t_e, f0_{offset}) \quad (2)$$

In order to obtain the set of linear equations we take derivatives of the error (t_r^i, t_f^i and t_e^i are kept constant):

$$\frac{\partial e^2}{\partial A_r^i} = 0 \quad \frac{\partial e^2}{\partial A_f^i} = 0 \quad \frac{\partial e^2}{\partial f0_{offset}^i} = 0 \quad (3)$$

If this formulation is applied to the two-stage method, there is one set $(A_r^i, A_f^i, f0_{offset}^i)$ for each event in the sentence. In the next section the system of linear equations simultaneously finds the optimal solution for all the RFC parameters (A_r^i, A_f^i and $f0_{offset}^i$) of the training corpus. Then the event i is used for all the contours which belong to the cluster i, defined by the machine learning technique (see section 5).

The gradient descent algorithm consists of an update equation (6) and (7). t_r, t_f and t_e are updated while A_r, A_f and $f0_{offset}$ are kept constant.

$$p_{n+1} = p_n - diag(\mu_n)\nabla e(p_n) \quad (4)$$

$$\nabla e(p_n) = [\frac{\partial e}{\partial t_r}, \frac{\partial e}{\partial t_f}, \frac{\partial e}{\partial t_e}] \quad (5)$$

5. Joint parameter extraction and prediction algorithm

In this section we propose a novel algorithm to apply the *JEMA* to the Tilt intonation model. The goal is to estimate simultaneously the Tilt parameters and the prediction model. As stated above, the global optimization avoids the interpolation step of the stylization process and produces more consistent parameters, improving their predictability from linguistic features.

Classification and regression trees (CART) are selected to estimate the model. The advantage is that they can use both discrete and continuous features. Furthermore, the representation provides useful information to increase the knowledge about the task. This information can be used for future improvements of the system. The classification tree is used to cluster the Tilt intonation events using questions concerning the prosodic and phonetic context of the events. Each leaf of the tree collects a set of fundamental frequency contours from the training corpus that must be approximated by the Tilt intonation parameters. The optimal parameterization is obtained using a combination of closed-form solution for the amplitudes and f0 offset and a hill-climbing procedure for time instants, as explained above. These optimizations provide a global optimal approximation to all the fundamental frequency contours in the training database.

The steps of the algorithm are:

1. The tree has an initial root node, which has to represent all the events. The initial optimal solution is found that approximates all contours with the same set of Tilt acoustic parameters.
2. All possible questions are examined in the leaves. For each question, the optimal Tilt acoustic parameters are determined and the approximation error is calculated. The optimization is performed using a combination of closed-form solution and hill-climbing algorithm.
3. The splitting linguistic question for the tree is chosen next. The criterion for selection of the best linguistic question splitting the node is the minimization of the approximation error.
4. The process is iterated (from 2) until a minimum number of elements in the leaves is reached or the differential gain on accuracy is lower than a threshold.

6. Experimental results

In this section we present experimental results for two languages: Slovenian and Spanish, including objective and subjective evaluation of the intonation model.

6.1. Slovenian language

The Slovenian language [8] is a pitch accentuated language. The pitch and stress are namely strongly related. Stress is marked by a pitch rise within a stressed syllable, followed by a fall which depends on the syllable (baritone or ocsitone) and the accent (acute or circumflex).

In Slovenian language tonemic or stressed accentuation are used. Stress can influence also the meaning of words. Approximately half of Slovenians use tonemic accentuation: divide double tone shape of long stressed vowels. In tonemic speech they are able to separate the meaning of words or their parts with different tone position of stressed vowels. The other half of Slovenians use stressed accentuation (also known as dynamic accentuation). In this case the stressed vowels are distinguished from non-stressed by strength: the energy of the speech material. Stressed vowels are distinguished from non-stressed also on the tone bases: usually they are higher. Using just tone they are not able to distinguish different meaning of words or their parts (morphems).

The experiments for Slovenian language were performed using the male voice of the Slovenian PLATTOS corpus (1205 sentences) [9]. The speaker uses the stressed accentuation. The utterances for Slovenian database were manually segmented in phones and the fundamental frequency was obtained automatically.

The automatic detection of Tilt events was performed using HMM recognizer for the whole database [9]. Different sets of Tilt event sets were analysed. The best performance was achieved when using full label set defined by Paul Taylor [10], consisting of rising boundaries, falling boundaries, pitch accents, minor and normal pitch accents.

A comparison of objective measures for test data using two-stages and *JEMA* is shown in figure 1. The results support our proposal that *JEMA* improves the prediction accuracy. Global optimization of parameters reduces dispersion and increases the capabilities of machine learning techniques. However, correlation and RMSE measures show that we are still not able to model the Slovenian speaker intonation perfectly. Further work must be done in order to improve these results.

Method	**RMSE [Hz]**	ρ
Two-stages	23.22	0.46
Joint	18.24	0.60

Table 1: *Global results for test data for Slovenian Language.*

6.2. Spanish language

In the previous experiments we showed that *JEMA* can improve significantly the performance of the Tilt intonation model. This approach has already been used for other intonation models: superpositional-Bézier [5] and Fujisaki [6]. This section com-

pares the Tilt model and these other models. We have used a Spanish corpus with 750 sentences of a female voice. It includes enuntiative sentences of our previous work [6] and also prompts recorded for a dialog system. The utterances were manually segmented in phones, and the fundamental frequency contour was obtained from the laryngograph channel.

The Tilt events are initialized using two simple rules: one event is assigned to each stressed syllable and one event is assigned to the syllable before each phrase break. The linguistic features are the same that the ones used in [5, 6]: features related to the accent groups, the intonation group, position of the accented syllable in the accent group, etc.

The results obtained for the three methods (estimated by *JEMA*) are shown in Table 2. We observe that Bézier and Fujisaki intonation models have a similar performance which is better than the Tilt performance. One reason for this may be that Tilt intonation model performs a local approximation around the intonational event without achieving an accurate approximation of the rest of the contour. It should be noted that the three methods achieve higher performance (RMSE and ρ) when *JEMA* is used.

Method	**RMSE [Hz]**	ρ
Bézier	20.9	0.764
Fujisaki	21.2	0.761
Tilt	23.1	0.685

Table 2: *Global results for test data.*

Objective measures are the first indicators about the performance of an intonation model. However, in order to have a measure of acceptance by final users a listening test has been performed by twelve subjects. They were asked to judge the naturalness of the intonation of several sentences using a five point scale (1:unnatural, 5:natural). Each intonation model predicts the F0 contour of the test sentences. This contour is imposed to the test sentences by resynthesis using Praat.

Table 3 shows the results of perceptual evaluation of naturalness for all methods. The natural intonation is included in the test as a reference for the evaluators and also to ensure the competence of the evaluators.

Method	**MOS**
Natural	4.6
Bézier	3.4
Fujisaki	3.5
Tilt	3.4

Table 3: *MOS for 3 different intonation models trained JEMA.*

The table shows that the three intonation models are performing similar with MOS scores around 3.5. It also shows that we are far from obtaining the quality of natural contours. We believe that these results show that the main limitation is due to the prediction. More effort should be devoted to derive new features from the input text (including syntactic and semantic features) that may influence the intonation.

7. Conclusions

In this paper we have reviewed the Tilt algorithm. We have proposed a new formulation to estimate the Tilt events: instead of finding all the parameters using gradient descent techniques, some parameters (amplitudes and $f0_{offset}$) are derived using a closed-form formulation getting faster and better parameter estimation.

We have proposed to apply *JEMA* to train the Tilt model. In *JEMA*, the optimization is jointly performed for all the fundamental frequency contours of the corpus. This avoids stylization steps which are needed by two-stage approaches and that increase the variability of the Tilt events. *JEMA* had already been successfully applied to Fujisaki's intonation model and Sup-Bézier intonation model. All three models achieve similar objective and subjective results.

The experiments support the theoretical advantages of this approach (consistency and avoidance of some assumptions regarding continuity and component separation). The comparative results with two-stage approach reveal a significant improvement in the objective measures that support our hypothesis about higher consistency of the events. The model has been trained for two different languages, Slovenian and Spanish. In both cases the Tilt model seems to be appropriate.

8. References

[1] P. Taylor, "Analysis and synthesis of intonation using the Tilt model," *Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1697–1714, 2000.

[2] H. Fujisaki, S. Ohno, and S. Narusawa, "Physiological mechanisms and biomechanical modeling of fundamental frequency control for the common Japanese and the standard Chinese," *Proceedings of the 5th Seminar on Speech Production*, pp. 145–148, 2000, bavaria, Germany.

[3] D. Escudero and V. Cardeñoso, "Corpus based extraction of quantitative prosodic parameters of stress groups in Spanish," *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pp. 481–484, 2002.

[4] D. Hirst, A. D. Cristo, and R. Espesser, "Levels of representation and analysis for intonation," *Intonation : Theory and Experiment. Kluwer Academic Press, Dordrecht*, 2000.

[5] P. D. Agüero and A. Bonafonte, "Intonation modeling for TTS using a joint extraction and prediction approach," *Proceedings of the International Workshop on Speech Synthesis*, 2004.

[6] P. D. Agüero, K. Wimmer, and A. Bonafonte, "Joint extraction and prediction of Fujisaki's intonation model parameters," *Proceedings of International Conference on Spoken Language Processing*, 2004.

[7] K. E. Dusterhoff, "Synthesizing fundamental frequency using models automatically trained from data," Ph.D. dissertation, University of Edinburgh, 2000.

[8] J. Toporisic, *Slovenska slovnica*. Maribor: Zalozba Obzorja, 1984, in Slovenian.

[9] M. Rojc, "Time and space efficient structure of multilingual and polyglot TTS system - architecture with finite-state machines," *PhD Thesis*, 2003.

[10] P. Taylor, "The rise/fall/connection model of intonation," *Speech Communication*, vol. 15, pp. 169–186, 1995.