A STUDY OF JEMA FOR INTONATION MODELING

Pablo Daniel Agüero, Juan Carlos Tulli

University of Mar del Plata
Engineering Faculty
Mar del Plata, Argentina

Antonio Bonafonte

Universitat Politècnica de Catalunya
TALP Research Center
Barcelona, Spain

ABSTRACT

In the literature many intonation models are trained using parameters extracted sentence-by-sentence on contours interpolated in the unvoiced segments. This may introduce a bias in the final parameters and a reduction of the generalization of the model due to the increased dispersion of them. Recently, we have proposed JEMA, a joint extraction and prediction approach for intonation modeling that avoids such assumption. The parameter extraction and model training are combined in a loop where *i)* the model is successively refined, and *ii)* the parameters are extracted using information provided by the model. In this papers we present experiments based on synthetic data to evaluate this approach in a controlled environment. Both, the results with synthetic data and with natural speech, show that the use of JEMA is clearly superior to the standard estimation approach. The parameters are correctly extracted using several degrees of missing data (0% to 80%) and gaussian noise. In fact, the study shows that including JEMA in the training algorithm is even more relevant than the selection of a particular representation of the intonation contours, as Fujisaki, Bèzier, Tilt, or others.

Index Terms— Speech synthesis, Intonation Modeling.

1. INTRODUCTION

Nowadays, most of the intonation models for text-to-speech synthesizers are generated using corpus-based approaches. Machine learning techniques are frequently applied to uncover the mapping between the linguistic features and the fundamental frequency contour. This process is named training of the intonation model.

In general we can distinguish three aspects: the mathematical formulation used to describe the frequency contours, the estimation of the parameters of the training data and the training of the model to map the linguistic features and the parameters.

Several mathematical formulations have been proposed to describe the fundamental frequency contour using a compact representation: exponential (Fujisaki [1]), polynomial (Tilt [2] and Bezier [3]), piecewise lineal (IPO [4]), etc.

For each training utterance, the parameters of the model are estimated to fit the real intonation contours. In some mathematical formulations, a close-form solution exist. However, in many cases, optimization algorithms as gradient descent or genetic algorithms need to be applied.

The training of the intonation model is the last step. It consists of finding the mapping function that generates a fundamental frequency contour (f_0) given a set of linguistic features (F) extracted from the text available in a text-to-speech system: $G(F) = \hat{f}_0$. In the case of data-driven approaches this task is done minimizing the energy of the prediction error ($e = f_0 - G(F)$) for the training data.

Traditionally, the parameter estimation is applied sentence-by-sentence to the whole training data. Afterwards, the mapping model is estimated [5, 6]. Many intonation modeling techniques made use of some avoidable assumptions they may harm the task of intonation model training:

- *Continuity of the fundamental frequency contour.* Some intonation models need continuous fundamental frequency contours to perform parameterization. Interpolation techniques are used to fill the unvoiced regions of speech. The main drawback is that the interpolated contours may bias the estimation of the parameters.
- *Removal of noise and microprosody.* The fundamental frequency extraction on a speech signal is a task prone to errors: pitch halving, pitch doubling, microprosody and measurement errors in voiced-unvoiced boundaries are sources of noise. Smoothing techniques are applied to remove such effects. However, this smoothing process may introduce new noise.
- *Parameter ambiguity.* In some optimization methods or even in some mathematical formulation (e.g.: Fujisaki), several values of the parameters can provide good approximations to the fundamental frequency contour. This makes the prediction task more difficult, because similar contours can have different parameterizations, increasing the dispersion of the parameters.

Recently we have proposed JEMA, a Joint Estimation and Modeling Approach. Once the mathematical formulation is chosen, the *parameter estimation* and the *training of the model* are combined in a loop. Section 2 describes the JEMA methodology and discuss why JEMA overcomes the previous assumptions, producing better intonation models. In Section 3 we show the results of several experiments applying the proposed approach to two mathematical formulations: Bezier and Fujisaki. In order to compare JEMA with the classic estimation procedure (*SEMA*, sentence-by-sentence parameter estimation and modeling approach), we propose to use not only real intonation contours, but also *simulated data* where the designed features are designed to be perfectly correlated with the contours and we can control the experimental conditions. Finally, in Section 4 we provide some conclusions.

2. JOINT EXTRACTION AND MODELING APPROACH

In this paper we study an intonation model training technique that combines parameter extraction and the generation of the mapping function $G(F) = f_0$ in a loop, as shown in Figure 1.

As intonation model we use regression trees. This model not only provides the function $\hat{f}_0 = G(F)$ but also clusters the training

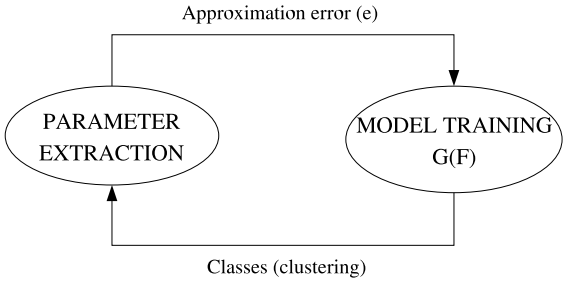

Fig. 1. Joint Extraction and Modeling Approach loop

data according to F. Other clustering algorithms can be applied into this scheme. An optimal clustering of the feature space (see figure 2) lets the contours of the f_0 space to be assigned to classes that will be represented by the same set of parameters, such as Fujisaki, Bèzier, Tilt, etc.

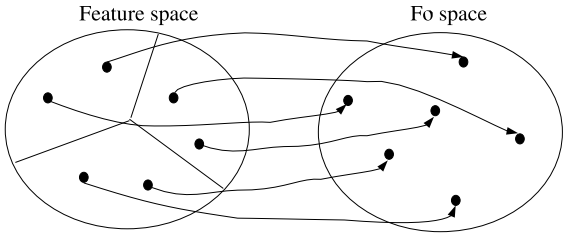

Fig. 2. Classes of contours given a clustering of the feature space

In our proposal, the parameters that represent each class of the cluster are calculated using a global optimization algorithm over all data available for training. One of the most important consequences of this configuration is that we do not need to interpolate the fundamental contours. The missing information of some contours that belong to a given class is compensated by the other contours of the class. In the same way, the use of a a global optimization minimizes the effect of the noise. And finally, minimizes the effect of ambiguities in the mathematical representation (multiple solutions with similar contours) or in the maximization algorithm (local minim).

On the other hand, all contours that belong to a given class will share the same set of parameters which will be optimal for the class. This global optimization leads the decisions of the clustering on the feature space to find out the optimal classes of contours.

In order to illustrate the process of intonation model training we show an example with just two sentences. In the example, we use a piecewise formulation: each contour is represented as a sequence of *intonation units* which are modeled independently.

- **Initialization.** Initially only one class exists, because the tree has only the root node. In this way, all prosodic units (accent groups, minor phrases) will be represented by the same set of parameters, as shown in Figure 3. These parameters are calculated using a global optimization algorithm over all training data.
- **Splitting.** The linguistic features are used to split the training data. The clustering splits the training data in two classes (figure 4).
- **Optimization.** When the new classes are obtained, a global optimization algorithm is used to find the new optimal parameters (Figure 5). Depending on the parameterization, if the optimal solution has not closed-form (e.g.: Fujisaki's intonation model), this optimization can be time consuming. In such cases hill-climbing algorithms are required to find the optimal solution.

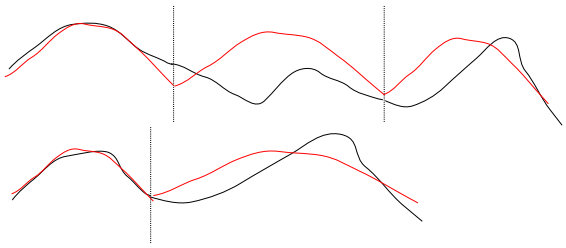

Fig. 3. Approximation with class 0 contour.

Prosodic unit contour	F1	F2	F3	F4	class
	B	...	...	...	1
	M	...	...	...	2
	E	...	...	...	2
	B	...	...	...	1
	E	...	...	...	2

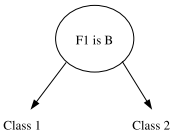

Fig. 4. Approximation with classes 1 and 2.

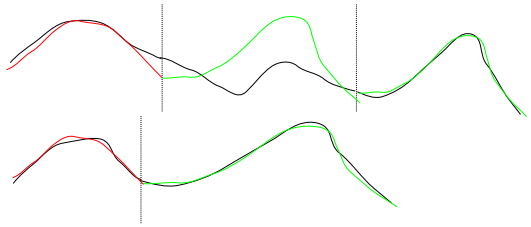

Fig. 5. Approximation with two classes.

- **Scoring of the splitting.** The new parameterization is used to measure the improvement of the goodness measure compared to its value previous to the splitting.
- **Selection of the highest improvement.** After all possible splittings are tried, the splitting with the highest improvement is chosen as the best split and the tree is updated for the next iteration.
- **Stopping condition.** The decision of another iteration for an additional splitting is performed taking into account a minimum number of elements on each leaf and a minimum improvement of the goodness score.

This approach can be applied to several parametric intonation models, because is a general technique to train intonation models, as was already shown for Bèzier [7], Fujisaki [8] and Tilt [9].

The clustering in the example is done using decision trees. However, this approach may be applied to other clustering techniques of the feature space, as shown in Agüero et al. [10].

3. EXPERIMENTS

In this section, we compare JEMA, the training methodology explained in the previous section, with the classic sentence-by-sentence parameterization approach (SEMA).

While the use of real data allows to draw the final conclusions, there are some characteristics of the intonation that make harder to compare different estimation methods. First of all, the initial intonation curves may include pitch estimation errors and these uncontrolled errors may influence our conclusions. Secondly, the mathematical formulation may not model exactly the intonation curves produced by humans. Therefore, this lack of accuracy can mask the effect of the estimation procedure. A third aspect is that the production of human contours is not a deterministic process. Therefore, using a real f_0 contour as a reference of the prediction capability of the model is just an approximation. And last but not least, the linguistic features used to predict the f_0 contour are not comparable with the information that humans use to talk. Therefore, the errors of each estimation methodology is contaminated by the fact that the features are not complete. For these reasons, this section starts using simulated date to analyse JEMA.

The artificial contours are generated using a set of eight classes with random parameters. Each class has a set of features that allow its complete separation of the others. The parameters for each class are selected so that the final fundamental frequency contours range from 100Hz to 200Hz.

Two parameterizations are evaluated. In the *Bèzier*, each contour is represented as a piecewise curve. Each piece of the curve is represented by a 3rd-order Bèzier polinomial, as shown in equation 1.

$$P(t) = \sum_{n=0}^{N} \alpha_n \begin{pmatrix} N \\ n \end{pmatrix} t^n \left(1-t\right)^{(N-n)} \tag{1}$$

The second parameterization is the superpositional model proposed by Fujisaki, and expressed in equations 2, 3 and 4).

$$\ln F_0(t) = \ln Fb + \sum_{i=1}^{I} Ap_i Gp(t - T_{0i}) + \tag{2}$$

$$+ \sum_{j=1}^{J} Aa_j \left[Ga(t - T_{1j}) - Ga(t - T_{2j})\right]$$

$$Gp(t) = \begin{bmatrix} \alpha^2 t e^{-\alpha t}, & t \geq 0 \\ 0, & t < 0 \end{bmatrix} \tag{3}$$

$$Ga(t) = \begin{bmatrix} \min[1 - (1+\beta t)e^{-\beta t}, \gamma], & t \geq 0 \\ 0, & t < 0 \end{bmatrix} \tag{4}$$

Forty training contours are generated using different percentages of artificially generated missing data (0% to 80%) and several levels of Gaussian noise with zero mean and standard deviation σ ($\sigma = 0Hz$, $\sigma = 1Hz$, $\sigma = 2Hz$ and $\sigma = 3Hz$). The missing data models the presence of unvoiced segments deleting segments of duration 50–100 msec. The duration of each contour is around 2–3 seconds and are composed by 4–8 minor phrases. The contours are sampled at 200Hz.

In the experiments we compare the RMSE of the intonation model generated by two training approaches: the classic SEMA (Separate parameter Extraction and Modeling Approach) and our new proposal JEMA (Joint parameter Extraction and Modeling Approach). As already mentioned, the parameters are derived sentence-by-sentence for SEMA and globally calculated for JEMA, as shown in [7, 8]. Leave-one-out training in order to obtain results which are statistically reliable.

In the case of SEMA, the contours are pre-processed using linear interpolation in the *lost* segments and a median filter.

3.1. Experimental results.

In Figure 6 we show the experimental results using Bèzier parameterization, training with SEMA (solid lines) and JEMA (dotted lines) and adding different levels of noise: $\sigma = 0Hz$ (diamond), $\sigma = 1Hz$ (star), $\sigma = 2Hz$ (square) and $\sigma = 3Hz$ (x-mark). The horizontal axis represents different levels of missing data.

The models trained using JEMA have the same RMSE for any level of missing data, and the increase of the error is only due to the added noise. However, the models obtained using SEMA suffer a strong impact in their performance with the increase of missing data. The better performance is a direct consequence of the consistence of the parameterization using global optimization. JEMA avoids the bias the bias due to the *unvoiced* interpolation.

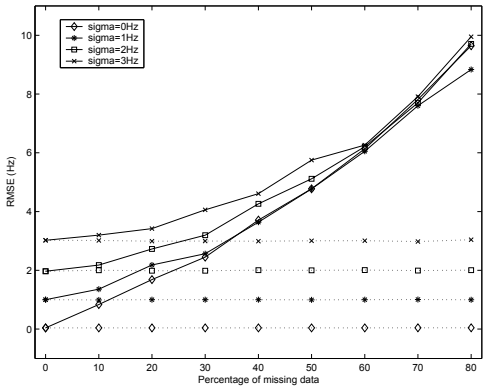

Fig. 6. Results using Bèzier parameterization (training data).

Figure 7 shows the experimental results using Fujisaki's parameterization. It can be observed that the parameters extracted without added noise have a higher RMSE than in the Bezier parameterization for the same condition. This is due to the sensitivity of the model on the time instants of phrase and accent commands. A small difference in time may introduce an error that varies depending on the choice of the constants α and β. This effect is less significant for higher values of σ due to a stronger influence of the Gaussian noise in the artificial contours.

Figure 7 shows how SEMA is also outperformed by JEMA at higher levels of missing data and added noise. Moreover, when missing data reaches 70% or 80%, the RMSE for SEMA is beyond the y-axis of the graphic. Nevertheless, JEMA has a flat performance at all percentages of missing data.

The simulations show that it is more important the estimation methodology (JEMA vs SEMA) than the method choosend to represent the intonation contour. For instance, for a noise level $\sigma = 2$Hz and 30% of missing data, the results for Bèzier and Fujisaki, using SEMA, are 3.1Hz and 2.5Hz, respectively. If JEMA is used, the error is reduced to 2.0 and 2.2Hz.

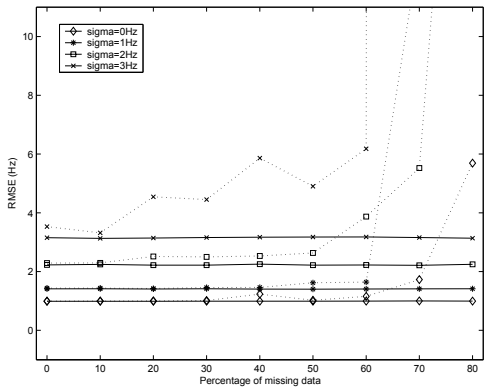

Fig. 7. Results using Fujisaki parameterization (training data).

3.2. Experimental results with real contours.

This section compares JEMA with SEMA using using real f_0 contours The models are estimated and evaluated using 220 paragraphs of the baseline Spanish male voice of the TC-STAR project. Figure 8 shows that the results with real speech are consistent with the ones obtained with simulated data. The figure shows the cumulative density distribution of RMSE (in this case, we use $\log f_0$). For each RMS value, the graph indicates the probability that the estimation gives an error smaller that this value. It can be shown how the errors are bigger for the SEMA estimation methodology (dashed line). This is true for the two mathematical formulations considered in the paper: Bézier (square) and Fujisaki (diamond). In fact the figure shows that the use of JEMA is more relevant that the mathematic formulation of the intonation contours: any models trained with JEMA is better than any model trained with SEMA.

JEMA (solid line) has a better cumulative density distribution of RMSE over the 220 paragraphs than SEMA (dashed line) for both intonation models taken into account in this paper: Bézier (square) and Fujisaki (diamond). Moreover, the results show that both models trained with JEMA are better than the best model trained with SEMA.

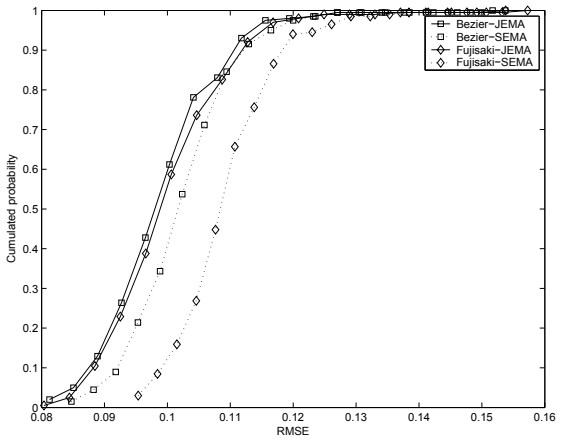

Fig. 8. Results using real contours.

4. CONCLUSIONS

In this paper we have analyzed a new approach for training intonation models: JEMA (Joint parameter Extraction and Modeling Approach). It overcomes some limitations in the extraction of parameters of classic proposals in the literature: requirements of continuity of the fundamental frequency contour and sentence-by-sentence parameter extraction. We have proposed to used *simulated* contours in order to avoid uncontroled effects (as for instance speaker variability) when comparing the modeling method.

The proposed training algorithm has shown the same RMSE for different percentages of missing data in the artificial contours used in the experiments. In the same conditions, SEMA (Separate parameter Extraction and Modeling Approach) has shown a degrading performance.

JEMA has shown in the experiments robustness to missing data. As a consequence, the extracted parameters are more consistent and have better generalization properties. In fact, we have found that not only JEMA provides better estimation but that the use of JEMA is most crucial than selecting the intonation representation.

The results with simulated data have been validated using real intonation contours derived from a Spanish database.

5. REFERENCES

[1] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of the Acoustical Society of Japan*, vol. 5, pp. 233–242, 1984.

[2] P. Taylor, "Analysis and synthesis of intonation using the Tilt model," *Journal of the Acoustical Society of America*, vol. 107, pp. 1697–1714, 2000.

[3] D. Escudero, "Modelado estadístico de entonación con funciones de Bézier: Aplicaciones a la conversión texto-voz en Español," *PhD Thesis, Universidad de Valladolid*, 2002.

[4] J. Hart, R. Collier, and A. Cohen, "A perceptual study of intonation. An experimental approach to speech melody," *Cambridge University Press*, 1990.

[5] H. Fujisaki, S. Narusawa, and M. Maruno, "Pre-processing of fundamental frequency contours of speech for automatic parameter extraction," *Proceedings of the International Conference on Signal Processing*, pp. 722–725, 2000.

[6] H. Mixdorff, "A novel approach to the fully automatic extraction of Fujisaki model parameters," *Proceedings of ICASSP, Istambul, Turkey*, pp. 1281–1284, 2000.

[7] 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, Pittsburgh, USA*, pp. 67–72, 2004.

[8] P. D. Agüero, K. Wimmer, and A. Bonafonte, "Joint extraction and prediction of Fujisaki's intonation model parameters," *Proceedings of ICSLP, Jeju Island, South Korea*, pp. 757–760, 2004.

[9] M. Rojc, P. D. Agüero, A. Bonafonte, and Z. Kacic, "Training the Tilt intonation model using the JEMA methodology.," *Proceedings of Eurospeech 2005, Lisboa, Portugal*, pp. 3273–3276, 2005.

[10] P. D. Agüero and A. Bonafonte, "Facing data scarcity using variable feature vector dimension," *Speech Prosody 2006, Dresden, Germany*, pp. 1–4, 2006.