# Modeling Intonation Variability with HMM for Speech Synthesis

*Cédric Boidin [1], Olivier Boeffard [2]*

[1] Orange Labs, Lannion, France
[2] IRISA / University of Rennes 1 - ENSSAT, Lannion, France
cedric.boidin@orange-ftgroup.com, olivier.boeffard@irisa.fr

## Abstract

This paper proposes a statistical intonation model designed to deal with intrinsic variability in speech. In combining the advantages of two well-known statistical algorithms, CART and HMM, the proposed model takes advantage of available linguistic information and successfully tackles the issue of missing para-linguistic information. Promising results of the training process are shown and analyzed.

## 1. Introduction

In recent years, corpus-based technology has brought major improvement in speech synthesis quality. The synthesized signal is the result of concatenating a string of acoustic units of variable length using as little signal processing as possible. Corpus-based technology performs well on homogeneous corpora with neutral speaking styles.

However, this technology is less successful for small or expressive corpora with more prosodic variability and less acoustic candidates for each contextual unit. In these cases, a prosody model is needed in order to better control unit selection or signal modification.

Ideally, a prosody model should be data-driven and statistical, i.e. capable to adapt to different corpora, speakers, speaking styles, etc. Such statistical, data-driven models have already been successfully used in unit selection for speech synthesis [1], speech recognition [2], and some work has already been done for prosody modeling [3].

However, prosody depends on many complex factors that are difficult to identify, such as speaker attitude, intention, and core accent of the sentence, etc. Even for the most advanced textual descriptions used in the Concept-To-Speech framework, there is no bijection between linguistics and prosody.

For instance Figure 1 shows two pitch curves of the French word "d'accord" from two different utterances. The pitch curves are significantly different even though the linguistic contexts are similar. The differences likely come from para-linguistic information, i.e. pragmatic or semantic information. Nonetheless, there are few elements to explicitly describe this additional para-linguistic information in a robust manner. Tagging this information, even manually, is currently out of reach.

In this paper, we include hidden information in order to model intrinsic variability in speech. Our proposed intonation model uses both linguistic information and an unsupervised HMM to deal with intrinsic variability in speech. Examining the example in Figure 1 again, under the same linguistic context, we expect that the HMM will accurately model both realizations.

Following the present introduction, section 2 describes the proposed intonation model and its subcomponents: intonation stylization, tree classification, and HMM modeling. Section 3 then presents the results of experiments evaluating the model. Section 4 summarizes our conclusions and presents future works.

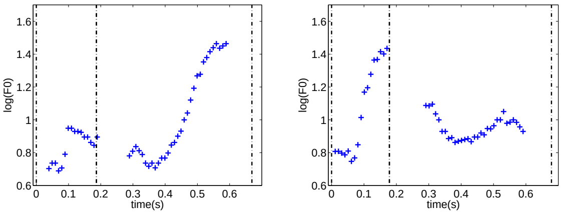

**Figure 1**: *Log fundamental frequency curves of the word "d'accord" (OK) in two different utterances: "D'accord, je le conserve." (OK, I keep it) on the left figure, "D'accord, à partir de maintenant vous avez des messages courts." (OK, from now on you will hear short messages) on the right figure. Dash lines correspond to syllable boundaries (d'a-ccord).*

## 2. Introducing variability into a statistical intonation model

Figure 2 shows a diagram of the intonation model proposed in this paper. It consists of 3 consecutive steps: syllable-based intonation stylization and annotation, tree classification, and HMM modeling.

### 2.1. Intonation stylization and annotation

The proposed intonation model is linguistically anchored, considering syllables as the elementary prosodic units. Each syllable is represented by a feature vector extracted from the acoustic utterance.

First, we consider the fundamental frequency ($F_0$). As in [4], we perform pre-processing in order to remove gross errors generated by the automatic $F_0$ extraction and also to remove microprosody. The result is a corrected $F_0$ contour, expressed on a logarithmic scale.

Next, each utterance is segmented into breath groups, syllables, and phonemes, with the nucleus identified for each syllable. As in [5], we fit the $F_0$ contour on each vocal nucleus with a second-order polynomial. We then create a 3-dimension syllable intonation vector comprised of the points located at 10%, 50% and 90% of the total contour duration.

A Karhunen-Loeve transformation is applied to all the intonation vectors in order to normalize and linearly de-correlate the components. The intonation vectors are first mean and variance normalized according to the speaker characteristics. They are then projected onto a new vector space in which the components are orthogonal. The Karhunen-Loeve transformation parameters are trained from a training corpus and then applied to any intonation vector.

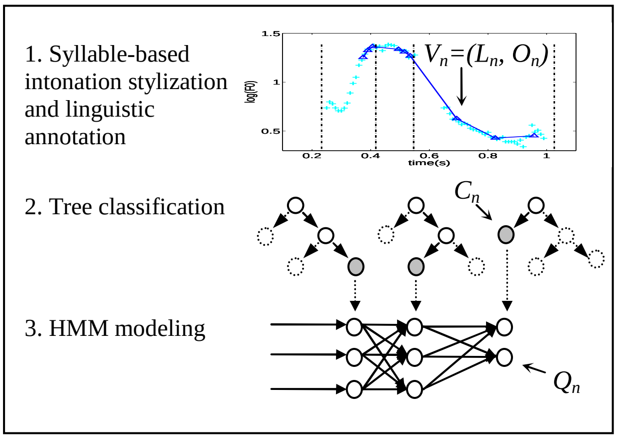

**Figure 2**: *Overview of the proposed intonation model for the French utterance "Je l'appelle." (I call him), composed of 3 syllables. Step 1 is the syllable-based intonation stylization from automatic F0 extraction and the linguistic annotation; step 2 is CART tree classification; step 3 is HMM modeling.*

For each utterance the result of the previous step is a sequence of 3-dimension normalized intonation vectors, $\{O_n\}_{1\leq n\leq N}$, gathered into breath groups.

The syllables are also automatically annotated with linguistic information. Linguistic tags deduced from the text include: word parts of speech (POS), position of the syllable in the word, position of the word in the breath group, and breath group type. These tags are gathered into linguistic feature vectors, described as $\{L_n\}_{1\leq n\leq N}$, which combine the tags of the previous, current and following syllables.

Each utterance is thus associated with a sequence of feature vectors $\{V_n\}_{1\leq n\leq N}$ combining intonation vectors and linguistic feature vectors: $V_n=(L_n, O_n)$.

### 2.2. Tree Classification

As done in many state-of-the-art prosody models [6], a CART is built to model intonation. The predicted variables are the intonation vectors $O_n$ and the possible splitting criteria are the components of the linguistic feature vector $L_n$.

The CART is described as a function $T: L_n \rightarrow T_{O_{train}}(L_n) = C_n$. The tree is trained on a training corpus $O_{train} = \{O_1 \ldots O_N\}$. The tree function assigns each feature vector $V_n$ to a corresponding class $C_n$.

Depending on a complexity threshold fixed manually, the tree size and number of syllable classes may vary. A low complexity tree corresponds to a small tree with few syllable classes whereas a high complexity tree corresponds to a larger tree with more syllable classes.

### 2.3. HMM modeling

The HMM modeling step is designed to deal with unexplained intonation variability. It introduces hidden states that allow several different prosodic realizations for a given linguistic context. The hidden states are linguistically dependent: a tree class is split into several hidden states; each hidden state is associated to one single tree class.

The hidden states are defined as $\{Q_n\}_{1\leq n\leq N}$. The class/state association is defined by the split function $S: C_n \rightarrow S(C_n)$, $S(C_n)$ being a subset of the hidden states $Q_n$.

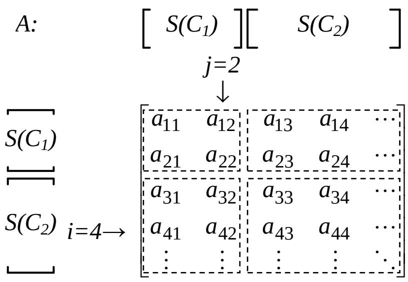
A:
S(C1)
S(C2)
j=2
S(C1)
S(C2)
i=4→
$a_{11}$ $a_{12}$ $a_{13}$ $a_{14}$ ⋯
$a_{21}$ $a_{22}$ $a_{23}$ $a_{24}$ ⋯
$a_{31}$ $a_{32}$ $a_{33}$ $a_{34}$ ⋯
$a_{41}$ $a_{42}$ $a_{43}$ $a_{44}$ ⋯
⋮ ⋮ ⋮ ⋮ ⋱

**Figure 3:** *Example of transition matrix A. The tree results in 2 classes $C_1$ and $C_2$. $C_1$ is splitted into 2 states ($Q_n \in S(C_1) = \{1,2\}$), the other states are associated to $C_2$. The term $a_{42}$ corresponds to $p_{\{C_2, C_1\}}(Q_n=2|Q_{n-1}=4)$.*

We then define a standard HMM:

$$P(O,Q) = P(O_1,\ldots O_N,Q_1,\ldots Q_N)$$

$$= P_{T(L_1)}(Q_1)P(O_1)\prod_{n=2}^{N} P_{\{T(L_n),T(L_{n-1})\}}(Q_n|Q_{n-1})P(O_n|Q_n)$$

We define the transition matrix $A=\{a_{ij}\}$ *as:*

$$p_{\{C_k, C_l\}}(Q_n=j|Q_{n-1}=i) = a_{ij} \text{ such that } (i, j) \in (S(C_k), S(C_l))$$

with the property: $\forall (i, l), \quad \sum_{j\in S(C_l)} a_{ij}=1.$

Figure 3 illustrates an example of such a transition matrix $A$. It is a full matrix composed of independent blocks representing the transition submatrices from one class to another. Each block is normalized as a usual transition matrix, i.e. the sum of the coefficients over a line is 1.

Similarly we define the initial probabilities $\Pi =\{\pi_i\}_{1\leq i\leq N}$ as:

$$p_{C_k}(Q_1=i) = \pi_i \text{ such that } i \in S(C_k)$$

with the property: $\forall k, \quad \sum_{i\in S(C_k)} \pi_i=1.$

We define the observation probabilities $B=\{b_j\}_{1\leq j\leq N}$, the probability of a particular observation vector at a particular instant $n$ for state $j$ being modeled as a Gaussian probability law described by: $b_j(o_n) = p(O_n=o_n|Q_n=j) = \mathcal{N}(o_n |\mu_j;\Sigma_j)$.

The parameters of the model $\lambda=\{B;A;\Pi\}$ are trained similarly to standard HMM with an Expectation-Maximization algorithm. We wish to find the optimal model parameters $\lambda^*$ which maximizes the likelihood $P(O_{train}|\lambda)$. The equations were implemented to take into account the new form of the initial and transition probabilities.

As the training data were decorrelated in a previous step, we suppose that the data are approximately decorrelated for each state. We thus choose diagonal variance matrices to reduce the number of parameters. The initial parameters are obtained from a Vector Quantification computed for each tree class.

## 3. Experiments

### 3.1. Material

We tested our model on a corpus of French utterances designed for France Telecom Interactive Vocal Response (IVR) servers. They were recorded by a professional female speaker for the operational server; the prosody is thus natural and expressive. The corpus consists of 582 sentences, 1033 breath groups and 6439 syllables.

The utterances were manually segmented into phonemes. Syllables were automatically deduced from phoneme segmentation and were automatically tagged with linguistic tags. Fundamental frequency is automatically extracted every 10 ms and stylized according to 2.1.

### 3.2. Experiments

We conducted experiments in order to measure the appropriateness of the model to fit on real data and to determine the optimal parameters for the different steps.

The experimental process consisted in training the parameters (normalization, tree, HMM) on a training corpus and test the likelihood on a separated test corpus. For each experiment we measure the log-likelihood of the training corpus, the log-likelihood of the test corpus, the overall number of states and the Root Mean Square Error (RMSE) between the natural utterance and the estimated solution obtained by the Viterbi algorithm.

The model parameters to be adjusted are the complexity (cp) of the CART tree (HMM -only 1 leave in the tree-; low complexity -few leaves in the tree-; middle complexity; high complexity -high number of leaves-) and the number of states associated to each class (for a class gathering *M* syllables we limit the number of states associated to the class to a maximum of *int(M/100) + 1*). Figure 4 and Figure 6 correspond to 1 to 8 states per class, whereas Figure 5 and Figure 7 also display experiments with a higher number of states per class.

In order to get significant results, each experiment was reproduced 10 times with a random split of the overall corpus into the training corpus (80% utterances) and the test corpus (20%).

### 3.3. Results: log-likelihood comparison

Figure 4 and Figure 5 show the evolution of the syllable average log-likelihood with the different parameter sets. The plotted curves show the mean values of the log-likelihood of the training corpus (plain line) and the test corpus (dash line) with their 95% confidence intervals. The confidence intervals might not be visible when they are smaller than the symbols.

Figure 4 shows the evolution of the log-likelihood for both the training and test corpora for several tree complexities and several numbers of states per class. The blue curve with square symbols corresponds to standard ergodic HMM, with all transitions between states allowed. In our formalism it corresponds to the case where there is one single class $C_n$. The left points of each curve correspond to one state per class; they are the results obtained with only the tree classification, i.e. without any split into hidden states. The average number of classes obtained with the tree classification is 4.0 for low complexity trees, 7.8 for middle complexity trees and 12.3 for high complexity trees.

For all experiments the difference between the scores obtained on the training corpus and the test corpus are quite low. This is a sign of good training behavior.

For both corpora the log-likelihood increases with the number of states per class. This shows that the HMM model captures well speech variability, and that this variability is reproducible.

The higher complexity models tend to give better results than the lower complexity models. The tree models (low-cp, middle-cp and high-cp) perform much better than the standard HMM for the same number of states per class. However the comparison is not really fair since the overall numbers of states is lower for the low complexity models than for the high complexity models, for a given number of states per class.

Figure 5 plots the evolution of the syllable average log-likelihood as a function of the overall number of states in the HMM on the test corpus. The overall numbers of states were grouped into 5-state large intervals in order to simplify the plots and get 95% confidence intervals.

The general trend is that less complex trees get higher likelihood for similar overall number of states. However the

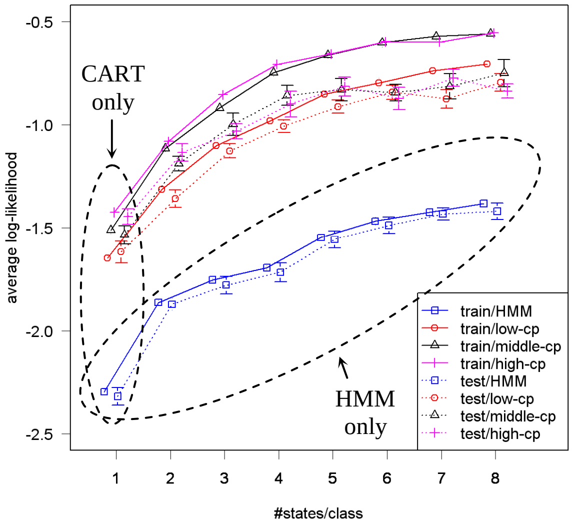

**Figure 4:** *Evolution of the log-likelihood of the training and test corpora with the complexity (cp) and the number of states per class.*

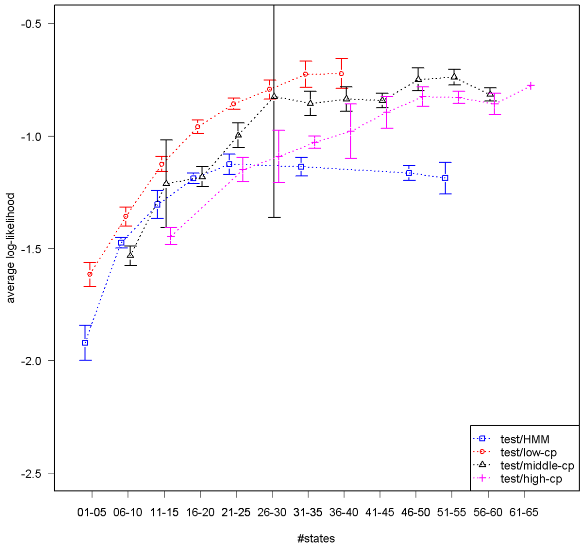

**Figure 5:** *Evolution of the log-likelihood of the test corpus with the complexity (cp) and the overall number of states.*

standard HMM (the lowest complex tree) overtrains and does not reach the log-likelihood of the other models. The low-cp model appears the best trade-off, reaching high likelihood with fewer states than the other more complex models.

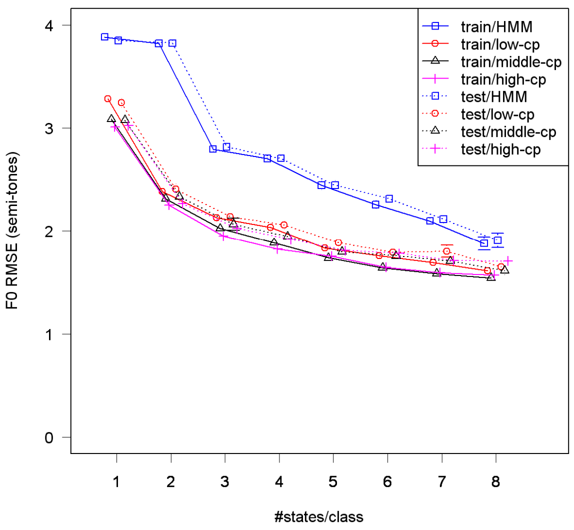

**Figure 6:** *Evolution of the $F_0$ RMSE of the training and test corpora with the complexity (cp) and the number of states per class.*

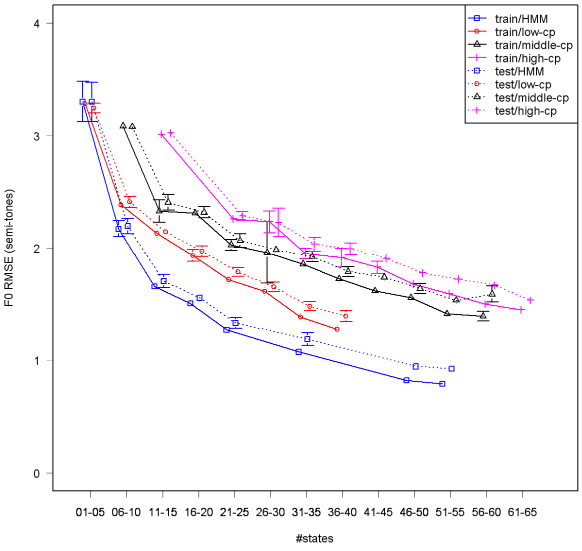

**Figure 7:** *Evolution of the $F_0$ RMSE of the training and test corpora with the complexity (cp) and overall number of states.*

### 3.4. $F_0$ Root Mean Square Error comparison (RMSE)

Figure 6 and Figure 7 show the F0 RMSE between the natural utterance and the Viterbi solution. The Viterbi solution corresponds here to the sequence of the means $\mu_j$ of the states $Q_n^*$ that maximize the probability $P(O, Q| \lambda)$.

Figure 6 shows the $F_0$ RMSE curves as a function of the number of states per class for both the training and test corpora. Figure 7 shows the $F_0$ RMSE curves as a function of the overall number of states for both the training and test corpora. As the speaker average $F_0$ is 211Hz a semi-tone corresponds roughly to 13Hz.

On both figures training and test curves are close: no overtraining appears yet for the RMSE criterion.

The ranking of the models changes if the comparison is done for the same number of states per class or for the same overall number of states. For instance the HMM model has the worst performance in the former case and the best in the latter case. However, neither of these scales allows clear insight into our problem. Figure 6 compares models with different overall complexity whereas Figure 7 compares models in which the number of states available for the decoding of each syllable differs. Indeed, for each syllable, the Viterbi decoding selects one state among the available states for the syllable, whose number corresponds to the number of states per class. Due to the different results obtained on the different scales, there is no clear evidence on which model outperforms the others for the $F_0$ RMSE.

## 4. Conclusion

In this paper, we propose a statistical intonation model that captures unexplained variability in speech. Initial results demonstrate good training behavior of the model on "real-world" data. The model significantly outperforms its two subcomponents -CART and HMM- taken independently. Experimental results also suggest that a low-complexity tree with a high number of hidden states for each class (8 or more) is the best compromise between maximizing likelihood and minimizing complexity.

The results are promising but work remains in order to use such a model for speech synthesis. Indeed the scores shown here are only likelihood and RMSE, measured on natural utterances. Future work will focus on using the model to synthesize speech, either to generate an intonation curve or to assist in unit selection, and also will involve subjective evaluation.

Additionally, future work could combine the two components that are currently handled sequentially -HMM and CART- into a single statistical framework, with each component taking into account the other's behavior. Such a unified framework is appealing, though much care is needed to avoid a excessive increase in the number of parameters.

## 5. References

[1] A. Hunt and A. Black, Unit selection in a concatenative speech synthesis system using a large speech database, ICASSP-96.

[2] L. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE, Vol. 77, No. 2, February 1989

[3] C. Traber, Talking machines: theories, models and designs, chapter F0 generation with a database of natural F0 patterns and with a neural network, pages 287-304, 1992.

[4] S. Narusawa, H. Fujisaki and S. Ohno, A method for automatic extraction of parameters of the fundamental frequency contour, ICSLP 2000.

[5] G. Bailly and B. Holm, SFC: a trainable prosodic model, Speech Communication 46, 2005.

[6] E. Eide, A. Aaron, R. Bakis, P. Cohen, R. Donovan, W. Hamza, T. Mathes, M. Picheny, M. Polkosky, M. Smith, M. Viswanathan, Recent improvements to the IBM trainable speech synthesis system, ICASSP 2003.