Towards Intonation Control in Unit Selection Speech Synthesis

Cédric Boidin 1, Olivier Boeffard 2, Thierry Moudenc 1, Géraldine Damnati 1

1 Orange Labs, Lannion, France
2 IRISA / University of Rennes 1 - ENSSAT, Lannion, France
{cedric.boidin, thierry.moudenc, geraldine.damnati}@orange-ftgroup.com, olivier.boeffard@irisa.fr

Abstract

We propose to control intonation in unit selection speech synthesis with a mixed CART-HMM intonation model. The Finite State Machine (FSM) formulation is suited to incorporate the intonation model in the unit selection framework because it allows for combination of models with different unit types and handling competing intonative variants. Subjective experiments have been carried out to compare segmental and joint-prosodic-and-segmental unit selection.

Index Terms: Speech synthesis, statistical intonation model, joint prosodic and segmental unit selection, finite state machines

1. Introduction

In recent years, unit selection 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. Unit selection technology performs well on homogeneous corpora with neutral speaking styles.

Unfortunately, this technology is less successful for expressive corpora with more prosodic variability. In this case, a prosodic model is needed in order to better control unit selection. Ideally, this prosodic model should be data-driven and statistical, i.e. capable of adapting to different corpora, speakers, speaking styles, etc.

However, prosody depends on many complex factors that are difficult to identify, such as speaker attitude, intention, core accent of the sentence, etc., constituting the para-linguistic and non-linguistic information [1]. In the following work, we do not have enough clues to tag this information in a reliable manner, we consider it as variability intrinsic to speech, and model it as hidden information.

Previously [2], we described a mixed CART-HMM intonation model that was developed to model the intonation of expressive corpora. A CART classifies intonation patterns according to linguistic information, and a HMM models missing para-linguistic information. This model was used successfully to generate intonation contours for unseen utterances [3].

This paper describes how we integrate the proposed intonation model into the unit selection process, with the aim of giving coherent intonation to synthesized utterances. For this purpose, we use the Finite State Machine (FSM) framework, as done in [4][5], whose formulation provides clarity, consistency and flexibility.

Following the present introduction, Section 2 describes the intonation model and Section 3 presents how to include the intonation model in the unit selection framework. Section 4 presents the experimental process and the results of subjective evaluations, and Section 5 summarizes our conclusions and presents future work.

2. Overview of the intonation model

The intonation model consists of 3 consecutive steps: syllable-based intonation stylization and annotation, tree classification, and HMM modeling. These steps are described in detail in [2]; we will present here only the global structure of the model.

2.1. Intonation stylization and annotation

The proposed intonation model is linguistically anchored, considering syllables as the elementary prosodic units. Each utterance is associated with a sequence of feature vectors \( j_{10\%N} \), with each feature being represented by a feature vector \( V_n = (L_n, \Omega_n) \) extracted from the acoustic utterance.

The linguistic vectors \( L_n \) contain the linguistic tags obtained from automatic annotation, such as word parts of speech (POS), position of the syllable in the word, position of the word in the breath group and breath group type of the previous, current and following syllables.

The intonation vectors \( \Omega_n \) are 3-dimensional vectors. As in [6], for each syllable, an intonation vector is first composed of three \( F_0 \) points, obtained from the automatic \( F_0 \) extraction (located at 10%, 50% and 90% of the duration of the vocal nucleus, i.e. the central vowel of the syllable). For a given speaker and corpus, the intonation vectors are then mean and variance normalized, and mapped into another space by a Karhunen-Loeve transformation, giving the abovementioned normalized intonation vectors \( \Omega_n \).

2.2. Tree Classification

A classification tree (CART) is used to classify the intonation patterns according to linguistic information. The tree function \( T \) associates a class \( C_n \) to every syllable, given its linguistic vector \( L_n \). The tree is trained from a training corpus in order to make the classes \( C_n \) as uniform as possible according to the intonation vectors \( \Omega_n \).

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 each tree class: a tree class is split into several hidden states; each hidden state is associated with a single tree class. The hidden states are defined as \( Q_n j_{10\%N} \).

Therefore, to each syllable \( n \), described by its linguistic vector \( L_n \), corresponds a single class \( C_n \) and several competing hidden states \( \Omega_n \) (the hidden states associated with \( C_n \)). We then define a standard HMM:
$$P(O,Q) = P(O_1,...,O_N,Q_1,...,Q_N)$$

$$= P_{T(L)}(Q)P(O|Q)\prod_{n=2}^{N} P_{T(L_n),T_{n-1}}(Q_n|Q_{n-1}) P(O_n|Q_n)$$

Note that the authorized transitions depend on the classes \(\{C_{n}\}_{n\in\mathbb{N}}\) equal to \(\{T(L_n)\}_{n\in\mathbb{N}}\) obtained from the tree. The observation probabilities are modeled as Gaussian laws: 
\[
p(O_n=\text{on}|Q_n=q_j) = b_j(\text{on}) = N(\text{on} | \mu_j, \Sigma).
\]

The HMM parameters are trained with an EM algorithm.

For any utterance, the intonation model thus provides a lattice of several competing HMM states per syllable, with transition probabilities coming from the HMM states of the preceding syllable and towards those of the following syllable. The lattice can evaluate the likelihood of any intonation sequence \(O\).

3. Unit selection

We propose to integrate the HMM state lattice produced by the intonation model into the unit selection. The lattice cannot be included into the standard target and concatenation costs, because the lattice allows several possible \(F_0\) targets for each syllable - a \(F_0\) target per HMM state corresponding to the maximum of the observation probability - and evaluates the coherence of the \(F_0\) evolution by consecutive syllables - through the transition probabilities.

We used the FSM framework in order to combine the two lattices: the segmental lattice including the target and concatenations costs and the intonation lattice produced by the intonation model. We therefore performed a joint segmental and prosodic unit selection, by the composition of 3 automata:
- an acceptor \(S\) representing the standard diphone-based unit selection lattice (also referred as segmental unit selection),
- a transducer \(V\) from units to vocal nuclei allowing to combine the diphone-based unit selection and the vocal-nucleus-based intonation model,
- a transducer \(I\) from vocal nuclei to HMM states representing the lattice given by the intonation model.

The composition results in the joint automaton \(J\): 
\[J = S \circ V \circ I\].

The following parts describe the automata \(S\), \(V\) and \(I\).

3.1. Standard Unit Selection

The standard unit selection consists of finding the best path over a lattice of target and concatenation costs. It can be represented in the form of a "segmental" automaton [7].

The "segmental" automaton \(S\) (\(S: \{u\}\)), represented in Figure 1, combines the target and concatenation costs. Considering that the states represent the units, each arc carries the symbol of the unit it leads to, and its cost is the sum of the target cost of the unit it leads to and the concatenation cost between the unit it comes from and the unit it leads to.

3.2. From diphones to vocal nuclei

A transducer \(V\) (\(V: \{u\} \rightarrow \{n\}\)) is built to allow for the future composition of the segmental and intonation automata, which consider different types of units, respectively diphone-based acoustic units and vocal nuclei. The role of \(V\) is to convert the units into vocal nuclei.

Note that, although the intonation model is syllable-based, the intonation vectors \(O_n\) depend only on the syllable vocal nuclei. Therefore it is possible to consider only the vocal nuclei in the following automata, thus reducing the size of \(V\).

\(V\) is built by considering all the possible successions of units that lead to vocal nuclei. The units that are not part of any vocal nucleus are converted into the null symbol \(\varepsilon\).

An example of the automaton \(V\) is illustrated on Figure 2. As the diphones \(<s>\text{v} \text{f} \text{a} \text{r} \text{f} \text{o} \text{r} \text{a} \text{l} \text{e} \text{r} \text{a} \text{n} \text{g} \text{a} \text{p} \text{p} \text{e} >\text{v} \text{i} \text{c} \text{k} \text{a} \text{p} \text{e} \text{r} \text{a} \text{n} \text{g} \text{a} \text{p} \text{p} \text{e} >\text{v}\) contain no vowel, the associated units \((u^{11}, u^{12}, u^{61}, u^{62})\) are converted into \(\varepsilon\). The other units, containing a vowel, are combined to produce vocal nuclei \((n^{11}, n^{12}, n^{13})\) to \(u\). The costs associated with the arcs of \(V\) are all equal to \(0\).

This automaton could also be obtained by the composition of two sub-automata, a first automaton realizing the transition from units to half-units (representing half-phones) and a second automaton realizing the transition from half-units to vocal nuclei.

3.3. Intonation model

For any new utterance, the intonation model determines, in chronological order, the tree class \(C_n\) for each syllable, the possible states for each syllable, the transition probabilities between the states, and the observation probabilities of the intonation of any vocal nucleus given any possible state. From this data we build the automaton \(I\) (\(I: \{n\} \rightarrow \{q\}\)), measuring the quality of the intonation of any succession of vocal nuclei.

The automaton \(I\) is built as the composition of two automata: \(I = O \circ T\):
- an acceptor \(T\) representing the transition probabilities between the possible states,
- a transducer \(O\) representing the observation probabilities of any vocal nucleus \((n)\) given any possible HMM state \((q)\), the associated costs are \(P(O_q|q)\).

Figures 3, 4 and 5 illustrate these automata when two HMM states are possible for each syllable.
3.4. Implementation issues

For an utterance composed of $M$ diphones and $K$ syllables, having $N$ candidate units per diphone and $O$ HMM states per syllable, the number of transitions of the automaton $S$ is $O(MN^2)$, the number of transitions of the automaton $I$ is $O(KN^2Q^2)$ and the number of transitions of the automaton $J$ is $O(KN^2Q^2+(M-K)N^2Q)$, which leads to computational cost issues. It has therefore been chosen to prune the acoustic transducer $S$ before composition with the other automata, in order to reduce the complexity of $J$.

The pruning step consists in suppressing all of the paths with cost superior to the best path cost plus a manually fixed threshold. On average, for our parameters, the pruned $S$ is composed of 9% of the states and 0.4% of the arcs of full $S$.

We checked that pruning $S$ does not drastically reduce the $I$ search space: the remaining vocal nuclei of the pruned $S$ fall homogeneously under the different HMM states.

Then, weights need to be balanced between the segmental and intonation automata $S$ and $I$ as we perform a joint optimization. In practice, as the segmental automaton has been deeply pruned to reduce the computational complexity, we consider that all the remaining paths are satisfactory in a segmental point of view. Therefore we gave a higher weight to the intonation automaton $I$ in the composition. The prototype was implemented using the FSM library from [8].

4. Experiments

4.1. Corpora

We tested our model on 4 French speech corpora designed for the Orange Labs unit selection speech synthesizer (demonstrator available at http://tts.Orange-fr). The corpora were mainly composed of utterances in the domain of IVR (Interactive Voice Response) applications, and were recorded by four professional speakers: Agnes, Lise, Julie and Loic. Agnes was constrained to respect a neutral speaking style, phonetics and break positions. The three other speakers were not constrained to speak in any neutral style; their corpora are thus more expressive than the Agnes corpus and most of the usual TTS corpora.

Table 1 describes the four speech corpora. For each voice, all the recorded utterances constitute the acoustic inventory for unit selection. A subset of the utterances is used for training the intonation model; its number of utterances (# training utts) and the corresponding number of syllables (# training syls) are shown in Table 1. The $F_0$ mean and standard deviation are also measured for each corpus, both in Hertz and semi-tones (st).

<table>
<thead>
<tr>
<th>Voice</th>
<th>Agnes</th>
<th>Lise</th>
<th>Julie</th>
<th>Loic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>M</td>
</tr>
<tr>
<td># utterances acoustic inventory</td>
<td>7855</td>
<td>4477</td>
<td>3207</td>
<td>3000</td>
</tr>
<tr>
<td># training utts intonation model</td>
<td>7232</td>
<td>2139</td>
<td>2830</td>
<td>2747</td>
</tr>
<tr>
<td># training syls intonation model</td>
<td>99348</td>
<td>32437</td>
<td>46125</td>
<td>44629</td>
</tr>
<tr>
<td>Mean $F_0$</td>
<td>182Hz</td>
<td>203Hz</td>
<td>203Hz</td>
<td>111Hz</td>
</tr>
<tr>
<td>Std $F_0$</td>
<td>33 Hz (2.9 st)</td>
<td>41 Hz (3.2 st)</td>
<td>62 Hz (4.6 st)</td>
<td>35 Hz (4.8 st)</td>
</tr>
</tbody>
</table>

4.2. Training the intonation model

The corpora utterances were manually segmented into phonemes and automatically tagged with syllables and linguistic tags. Fundamental frequency was automatically extracted every 10 ms and stylized according to Section 2.1.

For each voice, the intonation model was trained on the training corpus. The model parameters were manually chosen to have a relatively refined tree classification while avoiding over-training, resulting in 10 to 17 tree classes depending on the voice and 8 hidden states per class for each voice.

4.3. Unit selection speech synthesizer

Unit selection is performed at the diphone level and no prosodic modification is done on the selected units. The target and concatenation costs, used in the segmental automaton, were derived from the Orange Labs industrial synthesizer. The
concatenation cost uses acoustic parameters; the target cost uses symbolic prosodic tags but makes no use of any numerical intonation - for instance the distance to a predicted intonation curve - so that it does not constrain too much the unit selection. Therefore, $S$ and $I$ automata evaluate different aspects of the utterance; $S$ evaluates acoustics and symbolic prosody whereas $I$ evaluates numerical intonation.

4.4. Experiment

Small scenarios, containing 3 sentences on average, were designed in the domain of IVR services for evaluation purposes. For each voice, 11 scenarios were synthesized in two versions: SEG stimuli correspond to the result of the segmental unit selection (target and concatenation costs only), i.e. the concatenation of the units of the automaton $S$ best path; JOINT stimuli correspond to the result of the joint segmental and prosodic unit selection (target, concatenation costs and intonation model), i.e. to the concatenation of the units of the automaton $J$ best path. Additionally, NAT scenarios represent "natural" speech and were realized by concatenating up to 3 recorded sentences. Fifty native listeners were each asked to rate 18 scenarios on a 1-10 scale, 1 for bad, 10 for excellent.

Figure 6: Mean ratings of the scenarios with their 95% confidence intervals

Figure 6 gives the results of the subjective evaluation for the 4 voices and for the overall mean rating. The NAT version is rated between 7.4 and 8.1, always significantly better than the versions JOINT and SEG. The rated version is rated slightly better than the SEG version in overall (difference of 0.23) and 2 of the 4 voices, Lise and Julie. For Agnes and Loic, the difference between the two versions is lower than 0.1.

Table 2 gives the p-values of the t-test evaluating if the JOINT mean rating is greater than the SEG mean rating. With a significance level of 95% and a Bonferroni correction for 5 simultaneous comparisons, the maximum p-value for a significant comparison is 0.01. Therefore, none of the JOINT versions is judged significantly better than the SEG version.

Table 2: $p$-values of the t-test "JOINT mean rating greater than SEG mean rating"

<table>
<thead>
<tr>
<th>Voice</th>
<th>Agnes</th>
<th>Lise</th>
<th>Julie</th>
<th>Loic</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value</td>
<td>0.38</td>
<td>0.16</td>
<td>0.06</td>
<td>0.58</td>
<td>0.10</td>
</tr>
</tbody>
</table>

4.5. Discussion

Although the versions SEG and JOINT do not differ significantly in terms of subjective ratings, they do sound different. Table 3 shows the F0 features (mean and standard deviation) measured on the synthesized scenarios. It appears that the F0 features of the JOINT scenarios are globally closer to the corpus intrinsic values than the F0 features of the SEG scenarios. This confirms the beneficial influence of the intonation model in the unit selection.

Other factors might have prevented the system from a larger improvement in the ratings. First, the joint selection realizes a compromise between segmental and intonation constraints; the improvement in the intonation is therefore sometimes counterbalanced by a loss in segmental quality. Also, the other prosodic factors have been voluntarily kept away for this study, like phone duration and break positions. They need to be taken in account for a larger improvement.

Table 3: F0 mean and standard deviation of the scenarios synthesized with SEG and JOINT versions

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Agnes</th>
<th>Lise</th>
<th>Julie</th>
<th>Loic</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEG scenarios</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean F0</td>
<td>190 Hz</td>
<td>213 Hz</td>
<td>220 Hz</td>
<td>125 Hz</td>
</tr>
<tr>
<td>Std F0</td>
<td>(3.1 st)</td>
<td>(3.6 st)</td>
<td>(4.8 st)</td>
<td>(6.0 st)</td>
</tr>
<tr>
<td>JOINT scenarios</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean F0</td>
<td>178 Hz</td>
<td>191 Hz</td>
<td>203 Hz</td>
<td>110 Hz</td>
</tr>
<tr>
<td>Std F0</td>
<td>(2.8 st)</td>
<td>(3.1 st)</td>
<td>(4.7 st)</td>
<td>(5.1 st)</td>
</tr>
</tbody>
</table>

5. Conclusions

We propose to control the intonation of unit selection speech synthesis with a mixed CART-HMM intonation model. Three finite state automata are built and composed to form a joint automaton incorporating both segmental and prosodic constraints. The joint segmental-and-prosodic unit selection then finds the best path in the joint automaton.

Subjective experiments have not shown significant improvement in the quality of the synthesized utterances although the generated intonation is clearly different and closer to the natural intonation.

Future work will involve of including other prosodic factors in the intonation model, like phone duration.

6. Acknowledgments

The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 216594 (CLASSIC project: www.classic-project.org).

7. References