

# Prosodic control of unit-selection speech synthesis: A probabilistic approach



# Prosodic control of unit-selection speech synthesis: A probabilistic approach

**CONFERENCE PAPER** *in* ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 1988. ICASSP-88., 1988 INTERNATIONAL CONFERENCE ON · MAY 2011

Impact Factor: 4.63 · DOI: 10.1109/ICASSP.2011.5947569 · Source: DBLP



**2 AUTHORS**, INCLUDING:

Christophe Veaux
The University of Edinburgh
**33** PUBLICATIONS **102** CITATIONS

SEE PROFILE

# PROSODIC CONTROL OF UNIT-SELECTION SPEECH SYNTHESIS: A PROBABILISTIC APPROACH

*Christophe Veaux, Xavier Rodet*

IRCAM – CNRS STMS,
Analysis-Synthesis Team,
1, place Igor Stravinsky,
75004 Paris, France

{veaux,rod}@ircam.fr

## ABSTRACT

One problem in concatenative speech synthesis is how to incorporate prosodic factors in the unit selection. Imposing a predicted prosodic target is error-prone and does not benefit from the prosodic variability of the database. In this paper, we assume that several prosodic contours exist in the database for a same symbolic entry. This variability is represented by probabilistic models of the prosodic contours and the optimal sequence of units is searched by maximizing a joint likelihood at both segmental and prosodic levels. A generalized Viterbi algorithm is used to take into account the long-term dependencies introduced by the prosodic models. This method has been implemented in a unit selection synthesizer using an expressive speech database and a subjective experiment shows an improvement of the speech naturalness compared to a conventional unit-selection method.

**Index Terms**: speech synthesis, unit selection, prosody

## 1. INTRODUCTION

The prosodic control of the unit selection speech synthesizers remains one of their weakest links in the overall quality of the resulting speech. One approach relies on a two-stage decision process where a prosodic model is used to predict $F_0$ and duration targets for the unit-selection [1]. A major drawback of this method is that the prosodic sequence is chosen independently of the segmental units, which can create concatenation artifacts if no suitable sequence of units exists in the database. Furthermore, the prediction of a prosodic target often results in a stereotyped prosody that does not reflect the natural variability of the prosodic contours. Indeed, a unit selection synthesizer without any prosodic model often sounds better in overall naturalness.

In this paper, we assume that several possible prosodic realizations can be found in the database for a same symbolic entry. Therefore, instead of predicting a deterministic prosodic target at an early stage, we rely on probabilistic models of $F_0$ contour and durations and propose a method that searches for the optimal unit sequence by maximizing a joint likelihood at both segmental and prosodic levels. Since the prosody is intrinsically a supra-segmental phenomenon, the search of the optimal unit sequence has to consider several segmental units over time before making any decision. Therefore we use a generalized Viterbi algorithm (GVA) which offers the possibility of delayed decisions by relaxing the constraints over the searched paths [2].

There have been other proposals [3-5] to perform a joint search of the prosodic and segmental sequences. They are all based on the use of separate Finite State Automata (FSA) for the segmental and supra-segmental levels but differ in the way that these FSA are combined. A parallel search with token passing between the two FSA is done in [4] whereas the authors in [3,5] propose a composition of both FSA. In all cases, the search turns out to be computationally expensive and some pruning of the states must be performed to reduce this complexity. Interestingly, the GVA can be seen as a dynamic pruning of the less probable states as explained in section 3. Therefore, the joint search can be performed without increasing the search space.

This paper pursues a work initiated in [6]. Here, we propose to use gaussian mixture models of the $F_0$ contour and durations in order to better capture the prosodic variability of the database. These context-dependent models are learned over the syllables and the phrases by decision-tree growing based on Minimum Description Length [7] criterion.

The paper is organized as follows. In section 2, we shortly introduce the probabilistic framework of unit selection synthesis and detail the probabilistic models learned for each level (phone, syllable and phrase). The principle of the GVA and its implementation are presented in section 3. Finally, a subjective evaluation of the speech naturalness is presented and discussed in section 4.

## 2. PROBABILISTIC FRAMEWORK

In the traditional approach for unit selection synthesis, the best sequence of units is searched by minimizing a weighted sum of target costs and concatenation costs. This cost based view has been reformulated in a probabilistic framework in [8]. However, in both cases, the observations are at the segmental level only. Here we start from a more general view of unit selection in order to include the supra-segmental observations. Let $s = s_1, ..., s_K$ be a sequence of symbolic entries derived from the textual input and $u = u_1, ..., u_K$ a sequence of segmental units. The segmental units $u_k$ are generally phone-sized units. In a probabilistic framework of unit selection, we want to find the sequence that maximizes some observation probability $P(O(u) \mid s)$, i.e.

$$\hat{u} = \arg\max_{u} P(O(u) \mid s) \qquad (1)$$

where $O(u)$ denote the observation features associated to the sequence of units $u$, such as spectral features, energy and $F_0$ contours, segmental and syllabic durations. In our approach, we decompose these observation features into three levels of observation $O_{phr}, O_{syl}$ and $O_{pho}$ which correspond respectively to the phrase, the syllable and the phone level. Assuming statistical independence between these observations, we have,

$$P(O(u)) = P(O_{phr}(u))P(O_{syl}(u))P(O_{pho}(u)) \qquad (2)$$

Therefore, the best unit sequence can be searched by maximizing the product of the observation probabilities

associated to each level. In the following subsections, we detail the features and the statistical models learned separately for the phone, syllable and phrase levels. The spectral features are represented only at the phone level whereas we adopt a multi-level representation inspired by [9] for the prosodic features. In this way, we obtain a set of features that can fulfill at least partially the assumption of independence between levels stated in equation (2).

### 2.1. Phone level

Following the probabilistic model proposed in [8], we have,

$$P(O_{pho}(u)\,|\,s) = \prod_{k=1}^{N_{pho}} P(O_{pho}(u_k)\,|\,s) P(h(u_k)\,|\,t(u_{k-1}), s) \qquad (3)$$

where $u_k$ denotes the phone unit of index *k*. The probability $P(O_{pho}(u_k)\,|\,s)$ corresponds to the traditional target cost of unit selection whereas the conditional probability $P(h(u_k)\,|\,t(u_{k-1}), s)$ corresponds to the concatenation cost. We denote *h* and *t* the feature vectors associated with the beginning (head) and end (tail) of the units. In our current implementation, these feature vectors comprise:

- 13 MFCC coefficients and $\log F_0$ with their delta values

All these features are smoothed over a 10 ms window at the head and the tail of the unit. The observation features $O_{pho}$ are:

- 13 MFCC coefficients averaged at the middle of the phone.
- phone duration

In the current stage of our implementation, the observation probability is represented by a single gaussian model $P(O_{pho}\,|\,s) = \mathcal{N}(O_{pho}; \mu_s^{pho}, \Sigma_s^{pho})$ with mean vector $\mu_s^{pho}$ and diagonal covariance matrix $\Sigma_s^{pho}$. The conditional probability $P(h(u_k)\,|\,t(u_{k-1}), s)$ is represented by a simple model of zero-order linear prediction at the concatenation point, i.e,

$$P(h(u_k)\,|\,t(u_{k-1}), s) = \mathcal{N}(h(u_k); t(u_{k-1}) + \delta_s, \Sigma_s^{\delta}) \qquad (4)$$

where $\delta_s = E(h(u_k) - t(u_{k-1})\,|\,s)$ and $\Sigma_s^{\delta}$ is a diagonal covariance matrix. It can be noticed that with these simplifications, our transition model (4) is similar to the distance measure proposed in [10].

### 2.2. Syllable level

Assuming that the observations over a given syllable depend on the two preceding syllables, we have,

$$P(O_{syl}(u)) = \prod_{i=1}^{N_{syl}} P(O_{syl}(u_{syl(i)})\,|\,O_{syl}(u_{syl(i-1)}), O_{syl}(u_{syl(i-2)})) \qquad (5)$$

where the sequence of units $u_{syl(i)}$ corresponds to the syllable of index *i* and the dependency on the context $s$ is omitted for clarity. In the following we refer to $Z_{syl}$ as the conditional observation over the syllable, i.e.

$$Z_{syl}(i) = O_{syl}(u_{syl(i)})\,|\,O_{syl}(u_{syl(i-1)}), O_{syl}(u_{syl(i-2)}) \qquad (6)$$

The observation features $Z_{syl}$ consist of:

- 3 DCT coefficients of the $\log F_0$ contour ( $c_0$ excluded)
- delta and delta-delta values of the first DCT coefficient $c_0$
- syllable duration with its delta and delta-delta values

where the delta and delta-delta are estimated with respect to the preceding syllables. The first DCT coefficient $c_0$ which represents the average $\log F_0$ over the syllable is excluded from the syllable model since it will be part of the phrase model.

To account for the natural prosodic variability associated to a given symbolic context, the conditional probability in equation (5) is represented by a gaussian mixture model with mixture weight $\omega_j^{syll}$, mean vector $\mu_j^{syll}$ and diagonal covariance matrix $\Sigma_j^{syll}$

$$P(Z_{syl}\,|\,s) = \sum_j \omega_j^{syll}\ \mathcal{N}(Z_{syl}; \mu_j^{syll}, \Sigma_j^{syll}) \qquad (7)$$

where the dependency on the context $s$ is omitted for clarity.

### 2.3. Phrase level

We assume temporal independency between phrases, i.e.

$$P(O_{phr}(u)\,|\,s) = \prod_{i=1}^{N_{phr}} P(O_{phr}(u_{phr(i)})\,|\,s) \qquad (8)$$

where the sequence of units $u_{phr(i)}$ corresponds to the phrase of index *i*. One motivation behind this choice is to limit the long-term dependencies between units in equation (2) and consequently in the search for the optimal unit sequence. Nevertheless, it seems a reasonable assumption since the dynamic features used at the syllable level can bring to a certain extent some phrase-level information (e.g. $F_0$ resetting between phrases). The feature vector $O_{phr}$ consists of:

- 3 DCT coefficients of $\log F_0^{(syll)}$
- 3 DCT coefficients of syllable duration

where $\log F_0^{(syll)}$ denotes the average of $\log F_0$ over the voiced part of syllable. Similarly to the syllable level, the observation probability (8) is represented by a gaussian mixture model with mixture weight $\omega_j^{phr}$, mean vector $\mu_j^{phr}$ and diagonal covariance matrix $\Sigma_j^{phr}$

$$P(O_{phr}\,|\,s) = \sum_j \omega_j^{phr}\ \mathcal{N}(O_{phr}; \mu_j^{phr}, \Sigma_j^{phr}) \qquad (9)$$

where the dependency on the context $s$ is omitted for clarity.

### 2.4. Learning of the models

In order to handle unseen contexts and achieve robust training, the context-dependent parameters of the model densities must be tied by decision-tree clustering. We use the Maximum Likelihood (ML) as splitting criterion and the Minimum Description Length (MDL) as stopping criterion. The models associated to the phone, the syllable and the phrase are learned independently and separate decision-trees are grown for each of these levels. For the syllable level and the phrase level, the number of components of the gaussian mixtures is set to an initial maximum value of *K* gaussians. Nevertheless, in order to avoid over-fitting when growing the decision-trees, this number can be adaptively reduced by applying the MDL criterion within each node of the trees.

The symbolic features used to learn the decision-trees are of the 4 major classes: type, context (type of the previous/next units, type of the parent linguistic unit), position within the parent unit (with 4 categorical values: head, middle, tail or mono), weight (number of components). More specifically, the type features used for each level is as follows:

- *Phone level*: phonological type, sonority degree, articulation strength.
- *Syllable level*: initial/final phonological class, lexical type of the parent word (lexical or grammatical), onset/coda weight, nucleus position.
- *Phrase level*: mode (interrogative, exclamatory or neutral), initial/final word lexical type.

After the learning process, the statistical model for each level consists of a tree of context-dependent multi-gaussian models[1]. At synthesis time, given an input specification $s$, the models that best match the local context at each time instant are searched through each tree. These "predicted" models are finally combined according to equation (2) in the dynamic search for the optimal unit sequence that we present in the following section.

## 3. GENERALIZED VITERBI SEARCH

In a traditional unit selection synthesizer, a subset of $N$ units $\{u_k\}$ is preselected for each symbolic input $s_k$. A Viterbi algorithm is then used to find the best path within a trellis whose states at time *k* are the candidate units $\{u_k\}$. This algorithm considers all the transitions between successive units $u_k$ but keeps only the best path (survivor) leading to a given candidate unit $u_k$. However, within all the candidate units at a given time only a few of them will belong to a probable path and it seems reasonable to omit the others.

The generalized Viterbi algorithm (GVA) is one such modification of the Viterbi algorithm: at each time *k*, the *N* candidates units are stored into *M* lists and the best *S* candidate paths are selected from each list. By relaxing the constraint over the survivor paths, the GVA can retain survivor paths that would otherwise be merged by the classical Viterbi algorithm. Furthermore, by pruning the less probable candidate units, the long-term dependencies between units can be considered without increasing the dimensionality of the search space. Since we lose the systematic structure of the Viterbi algorithm, the search for the best sequence of candidate units becomes sub-optimal. Nevertheless, this loss of optimality is negligible as long as we can assume that only a limited number of unit sequences are likely to be a good solution.

An illustration of this approach is given in Figure 1 with the settings *N*=3, *M*=1, and *S*=3. With this particular setting, the GVA reduces to the List-type algorithm, in fact the GVA allows a wide range of settings, ranging from the classical Viterbi (*M*=*N*, *S*=1) to the List-type algorithm (*M*=1, *S*=*N*).

### 3.1. Unit selection procedure

The unit selection implemented in our speech synthesizer is illustrated in Figure 2 and can be described as follows.

• Initialization:

For each linguistic level (phone, syllable and phrase), the probabilistic models that best match the symbolic context at each time are selected from the contextual tree learned at this level. The contextual tree trained at the phone level is also used to preselect a subset of $N$ units $\{u_k\}$ for each symbolic input $s_k$. These candidate units are sorted according to their observation probability $P(O_{pho}(u_k)\,|\,s_k)$ and then alternately distributed along *M* lists with *N*/*M* units per list.

• Recursion:

1) *Path extension:* At time $k$, the *S* survivor paths are extended by one unit to yield *NS* candidate paths, and these candidates are classified into *M* lists.
2) *Update of observation memories:* Observation memories are associated to each survivor path. These memories store the features estimated for each level along that path and are updated each time a new observation is available.
3) *Path selection:* Using the statistical models of each linguistic level, the a posteriori probability (2) is evaluated from the available observations along each candidate path. Finally, the best *S* paths from each list are selected for the next step.

### 3.2. Optimization of the unit selection

In our implementation of the GVA algorithm, we impose *S*=*N*/*M* in order to keep a constant number of competing paths. The ratio *N*/*M* controls the tradeoff between the optimality of the unit selection at the segmental level (*M*=*N*) and its long-term memory ($M \ll N$). Obviously this parameter depends on the prosodic models learned at the syllable and phrase levels, and must be optimized over the training corpus. Imposing the number of candidate units *N*, we derive the optimal number of lists *M* as follows.

For each sentence *s* of the training corpus and a set of discrete values of *M* with $1 \le M \le N$, we compute the normalized selection cost:

$$Q(M,s) = \frac{\log(P(O_{phr}(\hat{u})))}{N_{phr}} + \frac{\log(P(O_{syl}(\hat{u})))}{N_{syll}} + \frac{\log(P(O_{pho}(\hat{u})))}{N_{pho}}$$

where $\hat{u}$ is the unit sequence selected by the GVA with the parameter setting *M*. The optimal value of *M* is the one that maximize the average of $Q(M,s)$ over the training corpus:

$$\hat{M} = \arg\max_{M} \frac{1}{N_s} \sum_{s} Q(M,s) \qquad (10)$$

where $N_s$ is the number of sentences of the training corpus.

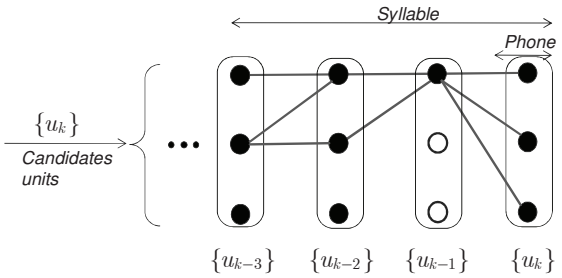


*Figure 1: Principle of the GVA unit selection. The boxes represent the lists of candidate units among which the best S paths are selected. The blank nodes are pruned units.*

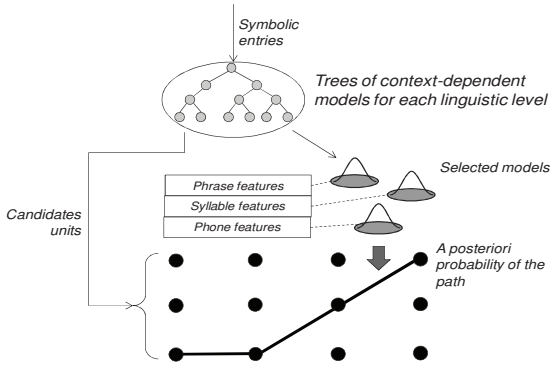


*Figure 2: Outline of the proposed unit-selection synthesis. The probability of a given unit sequence is estimated from the phone, syllable and phrase features evaluated along this path.*

[1] A tree of single gaussian models is used for the phone level.

## 4. EXPERIMENT

The speech corpus used in our system comes from the recordings of a French actor and presents a high prosodic variability which makes it a good candidate for evaluating the prosodic control of the unit-selection. It contains more than 3000 sentences which represents a total of 4 hours of active speech. This corpus has been split into a training set of 1800 sentences, a validation set of 800 sentences and a testing set of 400 sentences. The syllable and phrase models were learned with a maximum number of components of 8 gaussians. The optimal number of lists of the GVA selection was estimated to be $M$=10 for a given number of $N$=50 candidate units.

### 4.1. Subjective evaluation

We designed a subjective experiment in order to evaluate the extent to which the long-term dependencies introduced by the syllable and the phrase models could enhance the overall speech naturalness. Therefore, we have compared two settings of our unit selection synthesizer:

- **Baseline** synthesis which uses only the phone model with the classical Viterbi search ($N$=*50*, $M$=$N$, $S$=1).
- **Multi-level** synthesis which uses the phone, syllable and phrase models with optimal settings ($N$=*50*, $M$=10, $S$=5).

A Comparison Category Rating test (CCR) [11] was set up to compare both synthesizers. A set of 15 speech utterances was randomly selected from the test corpus, ranging from 5 to 26 syllables and from 1 to 3 phrases per utterance. These utterances have been synthesized by both baseline and multi-level systems, and the synthesized samples were presented by pairs in random order. The subjects were asked to judge the overall naturalness of the speech, i.e. its prosodic quality as well as its acoustical quality. The ranking of the two methods was evaluated by averaging the scores of the CCR test for each method. A total of 25 subjects performed the test (all French native speakers; 10 experts and 15 naïve listeners). The result is illustrated in Figure 3. The mean rating value is 0.96 in favor of the multi-level synthesis with a p-value << 1.e-4. No significant difference was found between experts and naïve listeners.

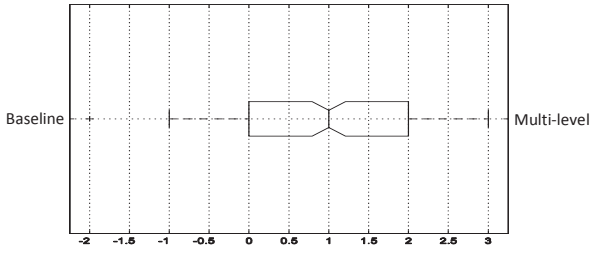


*Figure 3: Average CCR between the proposed system (multilevel) and baseline from the 25 subjects (inter-quartiles, median, and standard deviation).*

### 4.2. Discussion

The multi-level control of the unit-selection synthesis brings an improvement in terms of overall naturalness compared to a system that uses the segmental level only. The baseline system is the optimal setting for the segmental quality, but it relies solely on the smoothness constraint at the concatenation point to produce a consistent prosody. Therefore this result tends to demonstrate that the GVA combined with probabilistic models of the prosodic contours improves the prosodic quality without introducing too much segmental artifacts. Nevertheless, it seems obvious that the symbolic features used to train the prosodic models provide a rather limited representation of the linguistic structure. Therefore we are working on using high level linguistic features which may bring further enhancement of the prosodic quality.

## 5. CONCLUSION

We propose a probabilistic approach to the prosodic control of unit selection. The prosodic variability of the database is represented by multi-gaussian models of $F_0$ contour and durations. A generalized version of the Viterbi algorithm is used to search for the optimal unit sequence by maximizing a joint likelihood at both segmental and prosodic levels. This method has been implemented in a unit selection synthesizer using an expressive speech database and a subjective evaluation shows a consistent improvement in the speech naturalness. Further work will focus on the use of high level syntactical features in order to enhance the prosodic modeling accuracy.

## 6. REFERENCES


[1] K. Dusterhoff, A. Black and P. Taylor, "Using decision trees within the Tilt intonation model to predict F0 contours", *Eurospeech 1999*, Budapest, 1999.

[2] T. Hashimoto, "A List-Type Reduced-Constraint Generalization of the Viterbi Algorithm," in *IEEE Transactions on Information Theory*, vol. 33, no. 6, 1987, pp. 866-876.

[3] I. Bulyko and M. Ostendorf, "Joint Prosody Prediction and Unit Selection for Concatenative Speech Synthesis," *ICASSP 2001*, Salt Lake City, USA, 2001.

[4] R. Clark and S. King, "Joint Prosodic and Segmental Unit Selection Speech Synthesis," *Interspeech 2006*, Pittsburgh, PA, 2006.

[5] C. Boidin, O. Boeffard, T. Moudenc, and G. Damnati, "Towards Intonation Control in Unit Selection Speech Synthesis," *Interspeech*, 2009, Brighton, UK, 2009.

[6] C. Veaux, P. Lanchantin and X. Rodet, "Joint Prosodic and Segmental Unit selection for Expressive Speech Synthesis," *Proc. 7th Speech Synthesis Workshop*, Kyoto, Japan, 2010.

[7] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," J. Acoust. Soc. Jpn. (E), vol. 21, no. 2, pp. 79–86, 2000.

[8] S. Sakai and H. Shu, "A Probabilistic Approach to Unit Selection for Corpus-Based Speech Synthesis," *Interspeech 2006*, Pittsburgh, PA, 2006.

[9] Z. Wu, Y. Qian, F.K. Soong and B. Zhang, "Modeling and Generating Tone Contours with Phrase Intonation for Mandarin Chinese Speech," *ISCSLP 2008*, Kunling, China, 2008.

[10] R. Donovan, "A New Distance Measure for Costing Spectral Discontinuities in Concatenative Speech Synthesisers," in *Proc. of the 4th Workshop on Speech Synthesis*, Scoltand, 2001.

[11] ITU-T. P800, "Methods for Subjective Determination of Transmission Quality," *ITU Recommendations*, 1996.