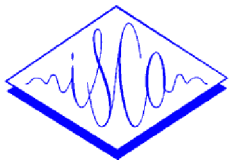
ISCA Archive
http://www.isca-speech.org/archive



7th ISCA Workshop on
Speech Synthesis (SSW-7)
Kyoto, Japan
September 22-24, 2010

Joint Prosodic and Segmental Unit Selection for Expressive Speech Synthesis

Christophe Veaux, Pierre Lanchantin, Xavier Rodet

IRCAM – CNRS STMS,
Analysis-Synthesis Team,
1, place Igor Stravinsky,
75004 Paris, France
{veaux,lanchantin,rod}@ircam.fr

Abstract

One problem in concatenative speech synthesis is how to incorporate prosodic factors in the unit selection. Imposing a predicted prosodic contour as target specification is error-prone and does not benefit from the natural variability contained in the database. This paper introduces a method that searches for the optimal unit sequence by maximizing a joint likelihood at both segmental and prosodic level. At the segmental level, the concatenation cost and target cost are reformulated in terms of conditional and a priori probabilities which are combined with probabilistic models of fundamental frequency and duration at the syllable level and the phrase level. A generalized version of the Viterbi algorithm is used to take into account the long-term dependencies introduced by the prosodic models during the search of the optimal unit sequence. This method has been implemented in a unit selection synthesizer using an expressive speech database and a subjective evaluation shows an improvement in the prosodic quality, although the overall quality is only slightly enhanced.
Index Terms: speech synthesis, unit selection, prosody

1. Introduction

Unit selection based synthesizers are capable of producing synthetic speech with a segmental quality very close to natural speech. However, the prosodic control of the unit selection remains their weakest link in the overall quality of the resulting speech. The usual approach to incorporate a prosodic control in unit selection synthesis relies on a two-stage decision process. A prosodic model is first used to predict a F_0 contour and segmental durations that subsequently serve as fixed targets during the search of the optimal sequence of segmental units. CARTs [1] and linear regression models [2] are typically used to predict these prosodic targets.

However, a major drawback of this method is that the prosodic sequence is chosen independently of the segmental units, which may result in concatenation artifacts if no suitable sequence of segmental units exists in the database. Indeed, a unit selection synthesizer without any prosodic model often sounds better in overall quality. This has led to a conservative strategy that consists of using database of neutral speech and simple constraints to ensure at least a well controlled neutral prosody. While this conservative approach can be suitable for some applications, it precludes the use of synthesized speech in many other situations that would require a more expressive and natural sounding prosody. Specifically, for applications like book reading, video games or dialog systems, the synthesized speech is expected to reproduce natural changes in speech rate, as well as emphasis and contrasts in the intonation. Unit selection synthesizers must be able to use expressive speech database to bring this extent of prosodic diversity. In return, an expressive database has much more sparsity than a neutral database and the specification of a single prosodic target for the search of the segmental units sequence becomes even more inappropriate.

In this paper, 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. To this end, the segmental target and concatenation costs traditionally used in unit selection synthesis are reformulated into a probabilistic framework that is extended to incorporate the probabilistic models of F_0 and durations. Independent statistical models of F_0 and durations are learned at different levels (phone, syllable and phrase) which permits to represent the prosodic variations at the level where they are best described. 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 propose to use a generalization of the Viterbi algorithm [3] which offers the possibility of delayed decisions by relaxing the constraints over the searched paths.

1.1. Related works

There have been previous research efforts [4-6] 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. They differ in the way that these FSA are combined during the search for the optimal unit sequence. A parallel search with token passing between the two FSA is done in [5] whereas the authors in [4,6] 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, our approach can be seen as a dynamic pruning of the less probable states as explained in section 4. Therefore, the joint search can be performed without increasing the search space. Furthermore, the proposed unit selection is formulated in a unified probabilistic framework that reduces the needs of manually balancing the weights between the segmental and the supra-segmental factors.

1.2. Paper content and organization

In section 2, we introduce the probabilistic framework of unit selection incorporating multi-level features. Then, the feature models for each level (phone, syllable and phrase) are detailed in section 3. The principle of the generalized Viterbi algorithm and its application to unit selection are presented in section 4. Finally, we detail the implementation of the new unit selection

synthesizer in section 5, and a subjective evaluation of its performance is presented and discussed in sections 6 and 7.

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 [7]. 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 specifications 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 synthesis, we want to find the sequence of units that maximizes some observation probability $P(O(u) \mid s)$, i.e.

$$u^* = \arg\max_u P(O(u) \mid s) \quad (1)$$

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 current approach, we introduce three levels of observation O_{phr}, O_{syl} and O_{pho} which correspond respectively to the phrase, the syllable and the phone level. Assuming that these observations are independent[1], the best unit sequence can be searched by maximizing the product of the observation probabilities associated to each level, since we have,

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

where the dependency on the context s is omitted for clarity sake. Furthermore, for a given level l, the observation probability $P(O_l(u))$ can be estimated from the conditional observation probabilities over each element in that level[2],

$$P(O_l(u)) = \prod_{i=1}^{N_l} P(O_l(u_{l(i)}) \mid O_l(u_{l(i-1)}), .., O_l(u_{l(1)})) \quad (3)$$

where $u_{l(i)}$ denotes the sequence of units associated with the element of index *i* within the level *l* (e.g. $u_{syl(i)}$ is the group of units that belong to the syllable of index *i*).
If we now assume that the observation $O_l(u_{l(i)})$ associated to the element i is dependent only on the L elements before that, the observation probability for the level *l* reduces to,

$$P(O_l(u)) = \prod_{i=1}^{N_l} P(O_l(u_{l(i)}) \mid O_l(u_{l(i-1)}), .., O_l(u_{l(i-L)})) \quad (4)$$

With this assumption, the maximization of equation (1) can be carried out recursively and thus the search for the optimal unit sequence can still be done under a dynamic programming approach. However, the phrase and syllable levels introduce long-term dependencies since each sequence $u_{l(i)}$ can span several segmental units for these levels. One way of solving this problem is to use a list-type generalization of the Viterbi algorithm that we present in section 4.

[1] An assumption we will discuss in section 3.

[2] The subscript *l* stands either for the phrase, the syllable or the phone level.

3. Multi-level Feature Models

In this section, we detail the features $\{O_{pho}, O_{syl}, O_{phr}\}$ 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 of the prosodic features. In this multi-level model, each prosodic feature estimated at a given level is relativized with respect to its mean over the considered level (except for the highest level). In this way, we obtain a set of orthogonal features between levels and can partially meet the assumption of independence stated in equation (2).

3.1. Phone level

At the phone level, it is usual to assume a temporal dependency of $L = 1$ which means that the observations over a given phone depend only on the preceding phone. Assuming for simplicity sake that segmental units are phones, the equation (4) can be reduced to,

$$\begin{aligned} P(O_{pho}(u) \mid s) &= \prod_{i=1}^{N_{pho}} P(O_{pho}(u_i) \mid O_{pho}(u_{i-1}), s) \\ &= \prod_{i=1}^{N_{pho}} P(O_{pho}(u_i) \mid s) P(h(u_i) \mid t(u_{i-1}), s) \end{aligned} \quad (5)$$

The probability $P(O_{pho}(u_i) \mid s)$ corresponds to the traditional target cost of unit selection whereas the conditional probability $P(h(u_i) \mid t(u_{i-1}), s)$ corresponds to the concatenation cost. Similarly to [7], 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, $\log F_0$, *VUF* and their delta values}

where the *VUF* is the voiced/unvoiced cut-off frequency. All these features are smoothed over a 10 ms window at the head and the tail of the unit, respectively.
The observation features O_{pho} consist of:

- {8 MFCC (c0 excluded), z-$\log F_0$ and loudness} measured at the head, middle and tail of the phone.
- {phone duration}

where z-$\log F_0$ denotes the mean-normalized $\log F_0$ over the phone. This feature is evaluated only for the sonorant phones, i.e. vowels, glides, liquids and nasals.

With this choice of features, the phone model accounts for the goodness of concatenation and for the segmental prosodic pattern. The observation probability is represented by a gaussian model $P(O_{pho} \mid s) = \mathcal{N}(O_{pho}; \mu_s^{pho}, \Sigma_s^{pho})$ with mean vector μ_s^{pho} and covariance matrix Σ_s^{pho} whereas the conditional probability $P(h(u_i) \mid t(u_{i-1}), s)$ is represented by a simple model of zero-order linear prediction at the concatenation point, i.e,

$$P(h(u_i) \mid t(u_{i-1}), s) = \mathcal{N}(h(u_i); t(u_{i-1}) + \delta_s, \Sigma_s^{\delta}) \quad (6)$$

where $\delta_s = E(h(u_i) - t(u_{i-1}) \mid s)$ and Σ_s^{δ} is a diagonal covariance matrix. It can be noticed that with these simplifications, our transition model (6) is similar to the distance measure proposed in [8].

3.2. Syllable level

Assuming that the observations over a given syllable depend on the two preceding syllables ($L = 2$), the equation (4) can be reduced to,

$$P(O_{syl}(u)) = \prod_{i=1}^{N_{syl}} P(O_{syl}(u_{syl(i)}) \mid O_{syl}(u_{syl(i-1)}), O_{syl}(u_{syl(i-2)})) \quad (7)$$

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.

$$Z_{syl}(i) = O_{syl}(u_{syl(i)}) \mid O_{syl}(u_{syl(i-1)}), O_{syl}(u_{syl(i-2)}) \quad (8)$$

The observation features Z_{syl} consist of:

- {delta and delta-delta syllable duration}
- {delta and delta-delta $\log F_0^{(sylll)}$}

where the delta and delta-delta are estimated with respect to the preceding syllables and $\log F_0^{(sylll)}$ is the average of $\log F_0$ over the vocalic part of syllable.

We do not incorporate the absolute duration and absolute $\log F_0^{(sylll)}$ in the syllable model since these features will be part of the phrase model. Consequently, the syllable level describes mainly the local prosodic prominence. The conditional probability in equation (7) is represented by a gaussian model $P(Z_{syl} \mid s) = \mathcal{N}(Z_{syl}; \mu_s^{syl}, \Sigma_s^{syl})$ with mean vector μ_s^{syl} and covariance matrix Σ_s^{syl}.

3.3. Phrase level

We assume temporal independency between phrases, i.e.

$$P(O_{phr}(u) \mid s) = \prod_{i=1}^{N_{phr}} P(O_{phr}(u_{phr(i)}) \mid s) \quad (9)$$

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 also 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). In our current implementation, the feature vector O_{phr} consists of:

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

A similar parameterization has already been proposed in [10]. It represents only the smooth variations of the prosodic curves, which is adequate in our framework since the local prominences are represented at the syllable level.

The observation probability is represented by a gaussian model $P(O_{phr} \mid s) = \mathcal{N}(O_{phr}; \mu_s^{phr}, \Sigma_s^{phr})$ with mean vector μ_s^{phr} and covariance matrix Σ_s^{phr}.

3.4. Learning of the models

In order to handle unseen context, the context-dependent parameters of the model densities are estimated by decision-tree clustering. In the current stage of our implementation, we assume that all the model densities have diagonal covariance matrices. With this simplification, we can use a variance criterion for the decision-tree growing and the learning of the contextual models can be split in two separate steps. In the first step, the feature vectors for each level are mean and variance normalized and the Euclidian distances between each pair of vectors are calculated over the training set. Theses pair-wise distances are then used as an impurity measure for the decision-tree growing in a similar fashion as in [9]. In a second step, the gaussian models are estimated within each cluster (i.e. either a leaf or a node of the tree) by calculating the mean and the variance of the feature vectors in that cluster.

After the learning process, the statistical model for each level consists in a tree of context-dependent gaussian models. At synthesis time, given an input specification s, the gaussian models that best match the local context at each instant[1] i 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.

4. 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 optimal path within a trellis whose states at time *k* are the candidate units $\{u_k\}$.

Now, if we consider a minimization criterion deriving from equation (2) with only the phrase level for simplicity:

$$L(u) = -\sum_{i=1}^{N_{phr}} \log P(O_{phr}(u_{phr(i)}) \mid s) \quad (10)$$

Using a Viterbi algorithm to minimize (10) supposes building a trellis of $N^{p(i)}$ states at each phrase of index i with $p(i) = \mathrm{card}(u_{phr(i)})$ which is impractical. This complexity comes from the fact that all the transitions between successive states are considered. However, within all the states 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 states are stored into M lists and the best S candidate paths are selected from each list. An illustration of this approach is given in Figure 1 with the settings $N = 6, M = 2$ and $S = 3$. It shows that the GVA can retain survivor paths that would otherwise be merged by the classical Viterbi algorithm. In this way, the long-term dependencies between units can be considered without increasing the dimensionality of the search space.

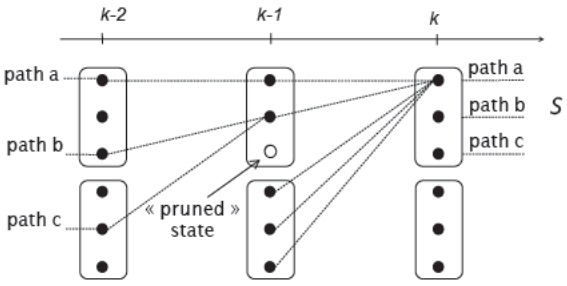


Figure 1: Principle of the GVA. The boxes represent the lists of states among which the best S paths are selected. Some survivor paths can share the same previous states whereas some states may have no survivors.

[1] The index *i* is relative to the level (phone, syllable or phrase).

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. The tradeoff between the optimality of the selection algorithm and its long-term memory is set by the couple of parameters (M, S).
Finally, our unit selection procedure can be described as follows.

• Initialization:

For each linguistic level (phone, syllable and phrase), the gaussian models that best match the symbolic context at each time are selected from the contextual trees learned at this level. The contextual trees trained at the phone level[1] are also used to preselect a subset of N units $\{u_k\}$ for each symbolic input s_k. These candidate units are then stored in M lists with N / M units per list[2].

• 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:* An observation memory is associated to each survivor path. This memory stores the features estimated along that path and is 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.

5. System Implementation

5.1. Speech corpus

The speech corpus used in our system comes from the recordings of a French actor that were originally intended for television dubbing. This corpus presents a high prosodic variability which makes it a good candidate for an expressive speech synthesizer although its phonetic coverage was not optimized for speech synthesis. It contains more than 3000 sentences which represents a total of 4 hours of active speech. All the symbolic analyses are derived from LiaPhon [12]. They consist in phonetization and syllabification, part-of-speech tagging and detection of breath groups (phrases). These symbolic tags were automatically aligned with the speech samples using ircamAlign [13] without any manual correction.

5.2. Training

For each level (phone, syllable and phrase), the symbolic features used to learn the decision-trees are of the 4 major classes:

- Type features
- Contextual features (type of the previous/next units or even a larger context, type of the parent /child units)
- Categorical positional features relative to parent linguistic units (with 4 values: head, middle, tail or mono)
- Weight features (number of components)

More specifically, the type features used for each level were 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.

We use wagon [11] to learn the decision trees after these symbolic contexts. The corpus has been split into a training set of 1000 sentences, a validation set of 400 sentences and a testing set of 400 sentences. The validation set was used to prune the tree during the learning process and a stop value of 20 is set as stopping criterion.

5.3. Optimization of the unit selection

Different values of the parameters (M, S) have been tested starting from the classical Viterbi $(M = N$ and $S = 1)$ to the List-type selection $(M = 1$ and $S = N)$. Informal listening tests yield to the setting $(N = 50, M = 10, S = 5)$. Obviously, this setting depends on the corpus content.

6. Experiment

In this experiment, we aimed to evaluate the extent to which the proposed selection method accounts for the supra-segmental components of the prosody. Therefore, we have compared two settings of our unit selection synthesizer:

- **Baseline** synthesis, which uses only the phone model with the classical Viterbi search $(N = 50, M = 50, S = 1)$.
- **Multi-level** synthesis, which uses all the three levels with the GVA search $(N = 50, M = 10, S = 5)$.

The baseline synthesis incorporates some prosodic control since the phone model represents the segmental duration and the mean normalized F_0 curve over the phone. However, it relies solely on the smoothness constraint at the concatenation point to produce a consistent prosody at supra-segmental levels (syllable duration, syllable prominence and prosodic phrase curves).

A Comparison Category Rating test[3] (CCR) [14] was set up to compare both synthesizers. A set of 15 speech utterances was randomly selected from the test corpus, with 19 syllables in average (from 7 to 44 syllables). 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. Additional information was asked to the subjects: speech expertise (expert, naïve), language (native French speaker, French speaker, or non French speaker) and listening conditions (headphones or loudspeakers).

A total of 24 subjects performed the test (all French native speakers; 13 experts and 11 naïve listeners). The results for the whole set of subjects is shown in Figure 2. In this case, the

[1] Actually since we use diphones units, we derive a diphone tree from the question-lists learned at the phone-level.
[2] For simplicity we assume a constant N for each subset $\{u_k\}$.

[3] The subjective test is available on-line at: "http://recherche.ircam.fr/equipes/analyse-synthese/veaux/index.php/Main/TestSubjectif"

mean rating value is 0.44 in favor of the multi-level synthesis (with a p-value << 1.e-6). Interestingly, the analysis of variance reveals a significant dependency on the listening conditions (p-value = 0.002). Subjects using loudspeakers tend to prefer the multi-level synthesis whereas subjects using headphones are more indeterminate as illustrated on Figure 3.

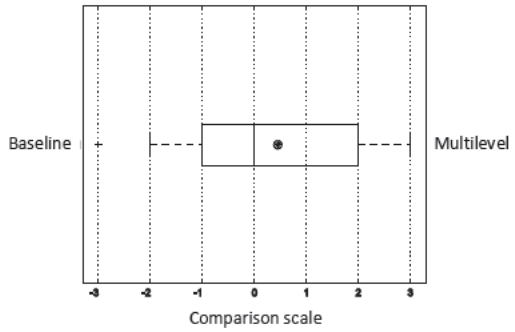


Figure 2: Average CCR between the proposed system (multilevel) and baseline from the 24 subjects (interquartiles, median, mean and standard deviation).

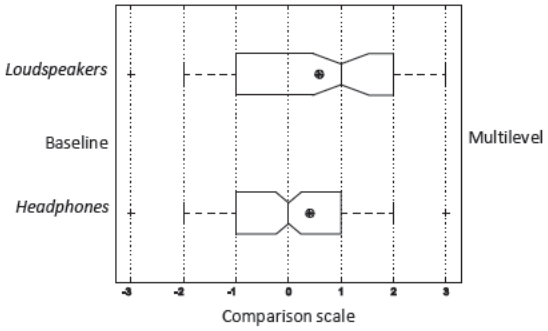


Figure 3: Average CCR between the proposed system (multilevel) and baseline from 11 subjects using loudspeakers and 13 subjects using headphones (interquartiles, median, mean and standard deviation).

7. Discussion

The proposed synthesis brings only a slight improvement in terms of subjective ratings (although statistically significant) compared to a baseline system that use segmental features only. Many factors can explain this result. First, the subjects reported that it was rather difficult to assess both acoustical and prosodic qualities. The significant dependency on the listening conditions (loudspeakers or headphones) tends to prove that in optimal listening conditions, subjects focused more their attention on the segmental quality. Second, the corpus itself may have too much prosodic variability to be correctly controlled without introducing concatenation artifacts. However, apart from the segmental quality, the prosody of the synthesized speech turns out to be consistently better with the proposed approach.

Nevertheless, there is a lot of room for improvement in our current implementation. The symbolic features used to train the prosodic models provide only a limited representation of the linguistic structure. A more elaborate symbolic analysis, with enriched linguistic information is certainly necessary to achieve a significant enhancement of the prosodic quality. Moreover, the accuracy of the models could be increased by considering full covariance matrices and by adopting a Maximum Likelihood (ML) approach for the decision-tree growing instead of the two steps learning process described in section 3.4.

8. Conclusions

We propose a method that searches for the optimal unit sequence by maximizing a joint likelihood at both segmental and prosodic level. It is based on a generalized version of the Viterbi algorithm which can take into account long-term dependencies between units during the search of the best unit sequence without increasing the dimensionality of the search space. 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 prosodic quality, although the overall quality is only slightly enhanced. Further work will focus on the improvement of the prosodic modeling accuracy.

9. Acknowledgment

This research was conducted with the support from FEDER project "Respoken".

10. 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] A. Black and A. Hunt, "Generating f0 contours from ToBI labels using linear regression," *ICSLP 96*, Philadelphia, 1996.

[3] 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.

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

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

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

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

[8] 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.

[9] A. Black and P. Taylor, "Automatically Clustering Similar Units for Unit selection in Speech Synthesis," *Eurospeech 1997*, Rhodes, Greece, 1997.

[10] 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.

[11] P. Taylor, R. Caley, and A. Black. *The Edinburgh Speech Tools Library*. CSTR, 1998.

[12] F. Béchet, "Lia_Phon: un système complet de phonétisation de textes," TAL, vol. 42, no. 1, pp. 47-68, 2001.

[13] P. Lanchantin, A. C. Morris, X. Rodet, and C. Veaux, "Automatic Phoneme Segmentation with Relaxed Textual Constraints," *LREC*, Marrakech, Morocco, 2007.

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