Symbolic Modeling of Prosody: From Linguistics to Statistics

Nicolas Obin, *Member, IEEE,* Pierre Lanchantin, *Member, IEEE*

***Abstract*—The assignment of prosodic events (accent and phrasing) from the text is crucial in text-to-speech synthesis systems. This paper addresses the combination of linguistic and metric constraints for the assignment of prosodic events in text-to-speech synthesis. First, a linguistic processing chain is used to provide a rich linguistic description of a text. Then, a novel statistical representation based on a hierarchical HMM (HHMM) is used to model the prosodic structure of a text: the root layer represents the text, each intermediate layer a sequence of intermediate phrases, the pre-terminal layer the sequence of accents, and the terminal layer the sequence of linguistic contexts. For each intermediate layer, a segmental HMM and information fusion are used to fuse the linguistic and metric constraints for the segmentation of a text into phrases. A set of experiments conducted on multi-speaker databases with various speaking styles reports that: the rich linguistic representation improves drastically the assignment of prosodic events, and the fusion of linguistic and metric constraints significantly improves over standard methods for the segmentation of a text into phrases. These constitute substantial advances that can be further used to model the speech prosody of a speaker, a speaking style, and emotions for text-to-speech synthesis.**

Index Terms: text-to-speech synthesis, speech prosody, speaking style, prosodic events, surface/deep syntactic parsing, hierarchical HMMs, segmental HMMs, Dempster-Shafer fusion.

I. Introduction

SPEECH PROSODY - "*the music of speech*" - denotes the long-term variations of speech that convey a large variety of information in a speech communication, from linguistic (e.g., meaning) to para/extra-linguistic (e.g., intentions, emotions, origins of a speaker). In particular, speech prosody constitutes the vocal signature of a speaker - his speaking style -, which contributes as a part of his identity. Modeling and adapting the speaking style of a speaker is desired for natural and expressive text-to-speech synthesis [1], [2]. The description of speech prosody can be divided into symbolic and acoustic characteristics: the symbolic representation accounts for the identification of prosodic events (i.e., accent and phrasing); the acoustic representation accounts for the acoustic variations over speech units (i.e., F0 and durations). A large number of methods have been proposed for the statistical modeling of the symbolic [3]–[8] and acoustic [8]–[12] characteristics of the speech prosody of a speaker, and the modeling and adaptation of speaking styles and emotions [13], [14].

Nicolas Obin is Associate Professor at IRCAM, UMR STMS IRCAM-CNRS-UPMC, Paris, France.

Pierre Lanchantin is Research Associate at the Cambridge University Engineering Department, Cambridge, UK.

A text-to-speech synthesis system requires the symbolic and acoustic representation of speech prosody. First, the position of prosodic events (accent and phrasing) is assigned from text analysis. Then, text information and prosodic events are combined to determine the sequence of speech parameters corresponding to the text. In this system, the symbolic modeling of speech prosody is crucial: the intelligibility and the naturalness of the synthesized speech is conditioned by the correct assignment of accents and phrases (Figure 1).

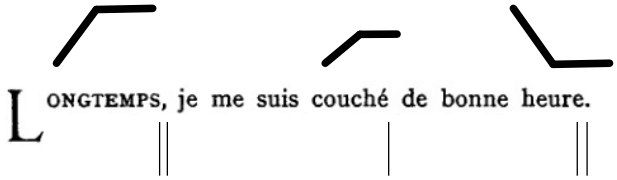

Fig. 1. Symbolic description of speech prosody corresponding to the utterance: "*Longtemps, je me suis couché de bonne heure*" ("*For a long time I used to go to bed early*") as read by a professional actor. On bottom: segmentation into prosodic phrases (/ and // denote minor and major phrases, respectively). On top: description of the corresponding prosodic contours.

Among the number of linguistic representations proposed for the description of prosodic events, ToBI (American-English [15]) is widely considered as a standard. However, the representation may substantially differ from one language to the other (i.e., numerous alternatives exist for French: INTSINT [16], IVTS [17], and Prosogram [18]). The symbolic modeling of speech prosody ranges from expert to statistical models - from formal rules derived from of a small number of linguistic observations, to the modeling of statistical regularities observed over large speech databases.

Research into linguistics generally assumes that a prosodic structure results from the integration of various constraints: in particular, linguistic and metric constraints ([19]–[24] for English; [18], [25]–[29] for French). A prosodic structure is primarily produced to clarify the linguistic structure of an utterance (linguistic constraint). Simultaneously, secondary extra-linguistic constraints tend to produce an optimal prosodic structure (e.g., metric constraint [25], [30]). These constraints conflict in the production of a prosodic structure, and secondary extra-linguistic constraints may override the primary linguistic constraint. The linguistic constraint mostly concerns the specification of prosodic boundaries that correspond to syntactic boundaries - e.g., syntactic constituency ([19], [20], [24], [29]), and syntactic dependency ([22], [26]). The metric constraint is considered as a secondary term used to adjust the length of prosodic phrases in the prosodic structure [20], [26], [29].

The statistical modelling used for the assignment of prosodic event from the text aims at deriving the prosodic structure from the linguistic structure of a sentence. From the text processing, surface syntactic parsing is currently a standard (part-of-speech, punctuation markers). Alternatively, recent advances tend to replace standard linguistic description of a text (e.g., part-of-speech) by unsupervised statistical words representation (e.g., Latent Semantic Analysis) [31], [32]. From the statistical modeling, a large number of methods have been proposed: from static representations (decision-tree [33], [34]), to sequential (HMM, N-grams [4]–[6], [35]), and hierarchical (hierarchical HMM [3]) representations. Also, some recent advances have introduced explicit formulation for the integration of the metric constraint into the statistical modeling for the segmentation of a text into phrases (segmental models [7], [13], [36]). From this point, current statistical methods used for the assignment of prosodic events mainly suffer from the following limitations: the statistical modeling is limited to the surface structure and does not exploit the deep structure of the sentence; the statistical modeling is generally focused on the linguistic constraint only, and there is still a room for a proper fusion of multiple constraints into the statistical modeling.

The main contribution of this paper is the integration of linguistic and metric constraints into the statistical symbolic modeling of speech prosody for text-to-speech synthesis. First, the paper explores the role of the linguistic description for the assignment of prosodic events. A linguistic processing chain that is used to provide a rich syntactic description of a text (surface and deep syntactic structure) is presented in sections II. Then, a novel statistical representation based on a hierarchical HMM (HHMM) is presented in section III. In this representation, the root layer represents the text, each intermediate layer a sequence of intermediate phrases, the pre-terminal layer the sequence of accents, and the terminal layer the sequence of linguistic contexts. For each intermediate layer, a segmental HMM and information fusion are used to fuse the linguistic and metric constraints for the segmentation of a text into phrases. The roles of the linguistic description and the combination of linguistic and metric constraints are presented in section IV.

II. Rich Linguistic Description

A. Surface and Deep Syntactic Parsing

This section presents the details of the automatic linguistic processing chain used to provide a rich syntactic representation of a text for the assignment of prosodic events. The main contributions here are the description of the deep syntactic structure of a text - in terms of constituency and dependency structures -, and the identification of a variety of syntactic constructions during the deep syntactic parsing (e.g., incises, parentheses, subordinate clauses).

The ALPAGE linguistic processing chain[1] is a full linguistic processing chain for French that is organized as a sequence of processing modules:

- a lexer module (LE*fff*: a French morphological and syntactic lexicon [37]; SXPIPE: a full linguistic pre-processing chain for French [38]);
- a parse module (DYALOG: a parser compiler and logic programming environment; FRMG: a FRench Meta Grammar [39]),

The lexer module segments a raw text into sentences and words [38], and processes surface parsing (morpho-syntactic and syntactic) for each sentence [37]. Then, deep parsing is processed to retrieve the syntactic structure of each sentence. Deep parsing is performed by the FRMG parser, a symbolic parser based on a compact TREE ADJOINING GRAMMAR (TAG) for French that is automatically generated from a META-GRAMMAR (MG) [39].

A Tree Adjoining Grammar [40] is a tree automaton composed of a finite set of *elementary trees* and a set of *operations* that are used to derive trees from elementary trees. Tree Adjoining Grammar accounts for all of the linguistic structures that can be derived from elementary trees by successive applications of the operations that are included in the grammar.

Elementary trees are minimal linguistic structures (*initial trees* and *auxiliary trees*). *Initial trees* are non-recursive linguistic structures that contain the essential structure of a sentence (e.g., phrasal structure), and *auxiliary trees* are recursive linguistic structures that contain the non-essential structures (e.g., adjective, adverb, clause) that can be adjoined to the essential structure of a sentence. Each elementary tree is associated with a lexical item which constitutes the anchor of the elementary tree.

Operations of *substitution* and *adjunction* are used to derive trees from elementary trees. The operation of *substitution* is associated with initial trees, and used to insert a non-recursive tree in a tree. The operation of *adjunction* is associated with auxiliary trees, and used to insert a recursive tree in a tree.

The FRMG parser provides a derivation tree that represents the most-likely structure derived from the sentence, and indicates all TAG *operations* (substitution, adjunction) that were used during the derivation.

The deep parsing processed with the TAG formalism allows the representation of a sentence in terms of *constituency* and *dependency* structures [41], and the description of a large variety of syntactic constructions through the *adjunction* operation.

Constituency and dependency structures. Constituency and dependency provide complementary representations of a sentence: the constituent structure represents the phrasal structure of a sentence (each constituent is a group of words which forms a phrase - e.g., verbal phrase, noun phrase), and the dependency structure

[1]The ALPAGE linguistic processing chain is available at: http://alpage.inria.fr/alpc.en.html

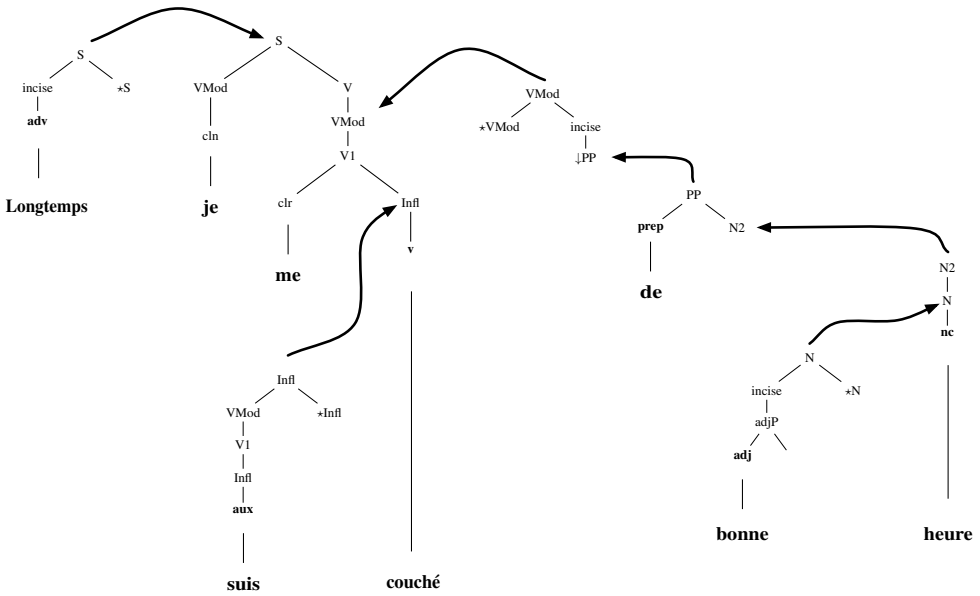

Fig. 2. Derivation structure of the sentence: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early*"). Arrows represent the insertion of elementary trees, and bold font indicates the lexical anchor of each elementary tree.

represents the local relation that connect each word of a sentence (each dependency connects a governor word to a governee word). An example of a dependency structure and a constituency structure are presented in Figure 2 and Figure 3.

Adjunctions. *Substitution* is mostly used to handle the essential phrase structure of a sentence (e.g., subject, verb, object). *Adjunction* is mostly used to handle non-essential modifiers (adjective, adverb, clause). In consequence, adjunction can derive a large amount of syntactic structures: from the adjunction of a single word (e.g., adjective, adverb), to the adjunction of complete structures (e.g., clauses, embedded clauses). In particular, adjunction covers a variety of syntactic constructions (e.g., incises, parentheses, subordinate clauses) that are possibly important for the modeling of speech prosody. An example of adjunctions is presented in Figure 4.

B. Extraction of Rich Syntactic Features

This section presents a description of the syntactic features extracted during the linguistic processing (surface and deep parsing), and further used as context-dependent labels for the assignment of prosodic events. The feature sets are composed of the three main syntactic classes presented in the previous section: morpho-syntactic (M), dependency (D), and constituency (C). An additional feature set that covers adjunctions (A) is additionally introduced.

1) Morpho-Syntactic: The morpho-syntactic features constitute the standard surface syntactic information:

- the *morpho-syntactic category* (part-of-speech) of a word;
- the *morpho-syntactic class* (i.e., function/content) of a word.

2) Dependency: The dependency structure is represented by:

- the *category* and *class* of the governor and governees of a word (as defined above);
- the *edge type* and *label* of the dependencies that connect a word to its governor and governees (e.g., type: substitution, adjunction; and label: anchoring of an adverb, a nominal phrase, a parenthesis, a subordinate clause);
- the *signed dependency distance* between a word and its governors and governees (measured in words and in chunks).

3) Constituency: The constituent structure is first converted into a sequence of chunks - defined as the terminal syntactic constituents of the constituent tree [42] -, and then represented by:

- the *chunk category* of each chunk (e.g., nominal phrase, verbal phrase, adverbial phrase), and each governor / governee chunk;
- the *edge type* and *label* of the dependencies that connect a chunk to its governor and governees;
- the *signed dependency distance* between the chunk and each governor / governee chunk (measured in words and in chunks);
- the *depth* of a chunk in the constituent tree (defined as the depth of the chunk node in the constituent tree, measured from the root node - i.e., the sentence);
- the *depth difference* between the current chunk and the left / right chunks.

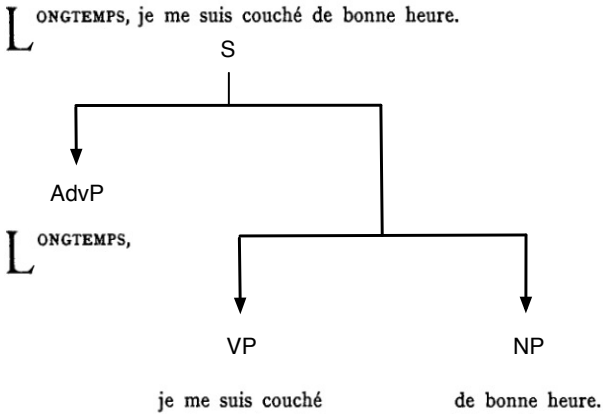

Fig. 3. Constituent tree derived from the sentence: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early*"). S, AdvP, VP and NP denote respectively sentence, adverbial, verbal, and nominal phrases.

4) Adjunction: Finally, adjunctions are retrieved by considering the complete descendency from the introducer of the adjunction. The introducer serves to introduce the syntactic construction that is adjuncted: introducers are commonly relative pronouns (e.g., who, which, that), subordinating conjunction (e.g., as, after/before, though, because), coordinating conjunction (e.g., and, or), and can be empty (e.g., for incises and parentheses). Adjunctions are represented by:

- the *category* and *class* of the introducer of the adjunction (e.g., relative pronoun, subordinating conjunction, coordinating conjunction, empty), and its governor / governee (e.g., noun, verb).
- the *edge type and label* that connect the introducer of the adjunction to its governor and governee (e.g., incise, parenthesis, subordinate clause, coordinate clause);
- the *signed dependency distance* between the introducer of the adjunction and its governor (in words and in chunks).

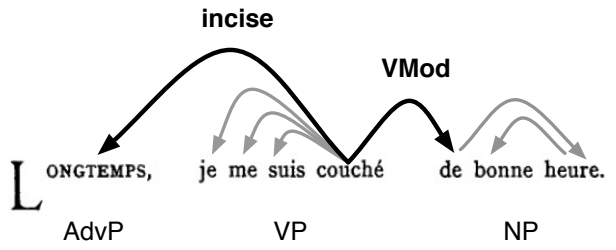

Fig. 4. Dependency structure of the sentence: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early*"). Arrows represent dependencies: grey arrows, substitutions; and black arrows, adjunctions. Black arrows indicate the left adjunction of the adverbial phrase (AdvP) "*longtemps*" ("*For a long time*") as an incise to the verbal phrase (VP) "*je me suis couché*" ("*I used to go to bed*"), and the right adjunction of the nominal phrase (NP) "*de bonne heure*" ("*I used to go to bed early*") as a verb modifier to the verbal phrase (VP).

Here, the nature of the introducer, governor, and governee, and their edge labels suffices to identity a large variety of syntactic constructions (e.g., incises, parentheses, subordinate clauses). For instance, a subordinating conjunction that modifies a noun and introduces an adjunction labelled as a subordinate clause instantiates a subordinate relative clause; in the case the introducer modifies a verb, then the adjunction instantiates a subordinate conjunctive clause.

III. Integration of Metric and Linguistic Constraints

In this section, a statistical method that combines linguistic and metric constraints for the modeling of prosodic events is introduced based on segmental HMMs and Dempster-Shafer fusion. A hierarchical HMM (HHMM) is used to model a sequence of prosodic events conditionally to the sequence of observed linguistic contexts (section III-A), in which phrasing (segmentation into prosodic phrases) is represented by a segmental HMM that accounts explicitly for the metric constraint, and the sequence of prosodic events corresponding to a prosodic phrase is represented by a standard HMM that accounts for the linguistic constraint (section III-B). Then, Dempster-Shafer fusion is used to combine the linguistic and the metric constraints into the segmental HMM (section III-C).

A. Hierarchical HMMs

The assignment of prosodic events (accent and phrasing) from the text can be formulated as a hierarchical processing: first, the segmentation of a text into prosodic phrases (phrasing); and second, the assignment of remaining prosodic events (accent). The phrasing comprises a structure of speech units which cover major phrases and a number of intermediate phrases. For each major phrase and each intermediate phrase, linguistic and metric constraints are combined for the segmentation. Then, the remaining prosodic events are conditioned by the linguistic constraint only.

A hierarchical HMM (HHMM) [43] is presented here to model jointly the segmentation into prosodic phrases and the sequence of prosodic events. Theoretically, the HHMM can assume any number of intermediate layers to represent additional intermediate phrases. For simplicity here, a single layer is used for the segmentation into major phrases (phrasing), and intermediate phrases are processed in the pre-terminal layer with the remaining prosodic events (accent). First, the text is converted into sequence of linguistic contexts. The first layer of the HHMM represents the phrasing of a speaker, the second layer represents the sequence of prosodic events emitted by the major phrase, and the third layer represents the observed sequence of linguistic contexts. For the first layer, the joint process of the major phrases and the prosodic events is modeled by a segmental HMM in order to account explicitly for the contribution of the metric constraint in the phrasing. For the second layer, the joint process of the prosodic events and the linguistic contexts is modeled by a HMM. An illustration of the HHHM is provided in Figure 5.

The remaining of the section details the use of the segmental HMM for the integration of the metric constraint, and the statistical fusion of the linguistic and the metric constraints for the segmentation of a text into major phrases.

B. Segmental HMMs

The segmental HMM is used to integrate the metric constraint for the segmentation of a text into major phrases. The main advantage of segmental HMMs [44], [45] (also referred as Hidden Semi Markov Models - HSMM) over hidden Markov models is the representation of a sequence of states as a sequence of segments whose duration (respectively, length) is explicitly modeled. In particular, segmental HMMs can be exploited for the integration of the metric constraint: major phrases are represented as segments whose statistical distribution of length can be modeled straightforwardly.

A segmental hidden Markov model $\boldsymbol{\lambda}$ is defined in a similar manner to the hidden Markov model with a reformulation of the state sequence into a segment sequence with associated segment duration distribution:

$$\boldsymbol{\lambda} = (\boldsymbol{\Pi}, \mathbf{A}, \mathbf{B}, \mathbf{D}) \tag{1}$$

where: $\boldsymbol{\Pi}$, $\mathbf{A}$, $\mathbf{B}$ denote the initial state probability, the transition probability, and the observation probability distributions, respectively; and $\mathbf{D}$ the segment duration probability distribution.

Let $\mathbf{c}_1^N = [\mathbf{c}_1, \ldots, \mathbf{c}_N]$ be the observed sequence of N linguistic contexts, where $\mathbf{c}_n = [c_n(1), \ldots, c_n(C)]^\top$ is the (Cx1) linguistic context vector that describes the linguistic characteristics associated with the n-th syllable; let $\mathbf{l}_1^N = [l_1, \ldots, l_N]$ be the corresponding sequence of N hidden prosodic events, where l_n denotes the prosodic event (accent and break) associated with the n-th syllable; let $\mathbf{s}_1^K = [s_1, \ldots, s_K]$ be the associated sequence of K major phrases s_k, and $\mathbf{d}_1^K = [d_1, \ldots, d_K]$ be the corresponding sequence of K hidden major phrase lengths, where d_k denotes the length of the k-th major phrase s_k (here, counted in syllables). Here, the linguistic contexts $\mathbf{c}_n$ are observed, the labels of the major phrases s_k, the duration d_k of the major phrases, and the labels l_n of the prosodic events are hidden variables. Finally, the labels s_k are fully known since the layer of phrasing is only composed of major phrases.

A major phrase is defined as the segment left/right bounded by major phrase breaks, so the correspondence of a major phrase s_k to the sequence of prosodic events can be written as follows:

$$s_k = [\, l_{n_{k-1}+1}^{n_k-1} = \bar{b},\ l_{n_k} = b \,] \tag{2}$$

where: $[n_{k-1}, n_k]$ denotes the position of the left/right major phrase breaks of the k-th major phrase with length d_k, and b and $\bar{b}$ denote the presence/absence of a major phrase break, respectively. In this representation, each l_n for $n \in [n_{k-1}+1 : n_k - 1]$ can be any prosodic event (accent, intermediate break) with the exception of a major phrase break, and l_{n_k} can only be a major phrase break.

The remaining of this section details the use of segmental HMMs for the segmentation of a text into major phrases and the complete assignment of the sequence of prosodic events.

1) Parameters Training: The model parameters are estimated from annotated speech databases, in which all variables are observed. The linguistic model $\boldsymbol{\lambda}^{(\text{linguistic})}$ is a context-dependent HMM model derived from decision-tree based parameter clustering. First, linguistic contexts are clustered so as to derive a context-dependent tree based on maximum-likelihood minimum-description-length (ML-MDL, [46]). Then, a context-dependent model $\boldsymbol{\lambda}^{(\text{linguistic})} = (\lambda_{S_1}^{(\text{linguistic})}, \ldots, \lambda_{S_M}^{(\text{linguistic})})$ is constructed from the set of terminal contexts $S = (S_1, \ldots, S_M)$ of the context-dependent tree. Here, the observation probabilities $\mathrm{p}(l_n|c_n)$ are determined from the context-dependent tree, and the stationary distribution and transition probabilities $\{\mathrm{p}(l_n), \mathrm{p}(l_n|l_{n-1})\}$ are estimated from the number of occurrences in the speech database. The segment duration model $\boldsymbol{\lambda}^{(\text{metric})} = \{\mathrm{p}(d)\}_{d=1}^{D}$ is a normal distribution estimated from the length distribution of major phrases in the speech database, where D corresponds to the maximal length allowed (in syllables).

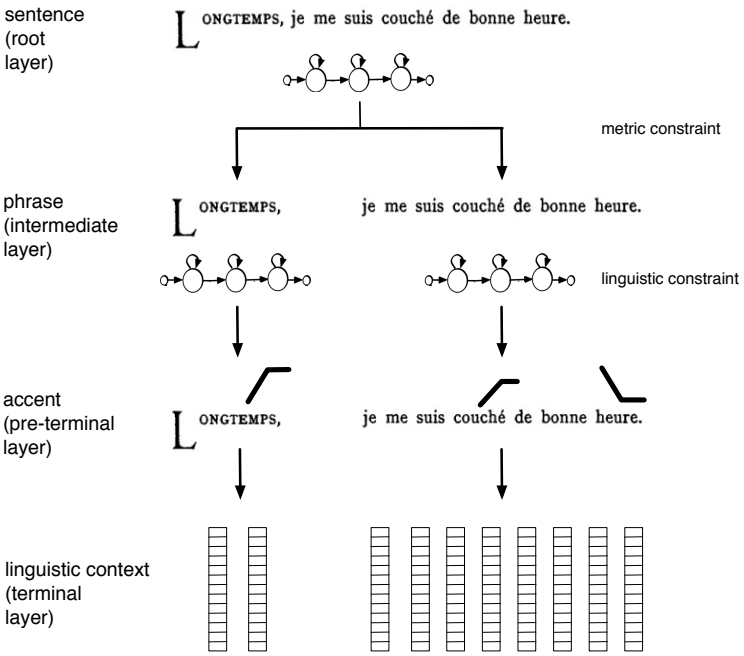

Fig. 5. Schematic illustration of the HHMM for the assignment of prosodic events.

2) Parameters Inference:

■ Standard HMM

In the standard HMM, the random process of prosodic events $\mathbf{l}_1^N$ is a stationary Markov chain and the observed linguistic contexts $\mathbf{c}_1^N$ are conditionally independent on the prosodic events $\mathbf{l}_1^N$, so that:

$$\mathrm{p}(\mathbf{l}_1^N) = \prod_{n=1}^{N} \mathrm{p}(l_n|l_{n-1}) \tag{3}$$

$$\mathrm{p}(\mathbf{c}_1^N|\mathbf{l}_1^N) = \prod_{n=1}^{N} \mathrm{p}(\mathbf{c}_n|l_n) = \mathrm{p}(\mathbf{c}_1^N) \prod_{n=1}^{N} \frac{p(l_n|\mathbf{c}_n)}{p(l_n)} \tag{4}$$

The optimal sequence of prosodic events $\widehat{\mathbf{l}_1^N}$ is then determined by maximising $\mathrm{p}(\mathbf{l}_1^N|\mathbf{c}_1^N)$ as:

$$\widehat{\mathbf{l}_1^N} = \underset{\mathbf{l}_1^N}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{l}_1^N|\mathbf{c}_1^N) \tag{5}$$

with:

$$\mathrm{p}(\mathbf{l}_1^N|\mathbf{c}_1^N) = \frac{\mathrm{p}(\mathbf{l}_1^N)\mathrm{p}(\mathbf{c}_1^N|\mathbf{l}_1^N)}{\mathrm{p}(\mathbf{c}_1^N)} = \prod_{n=1}^{N} \mathrm{p}(l_n|l_{n-1})\frac{\mathrm{p}(l_n|\mathbf{c}_n)}{\mathrm{p}(l_n)} \tag{6}$$

■ Segmental HMM

In the proposed HHMM, the first layer of the HHMM represents the phrasing of a speaker, the second layer represents the sequence of prosodic events emitted by the major phrase, and the third layer represents the observed sequence of linguistic contexts. On the one hand, the joint random process of the major phrases $(\mathbf{s}_1^K, \mathbf{d}_1^K)$ and the prosodic events $\mathbf{l}_1^N$ is modeled by a segmental HMM. On the other hand, the joint process of the prosodic events $\mathbf{l}_1^N$ and the linguistic contexts $\mathbf{c}_1^N$ is modeled by a HMM.

The optimal segmentation into major phrases $\widehat{\mathbf{d}_1^K}$ and the sequence of prosodic events $\widehat{\mathbf{l}_1^N}$ are jointly determined by maximising $\mathrm{p}(\mathbf{l}_1^N, \mathbf{d}_1^K|\mathbf{c}_1^N)$ as:

$$(\widehat{\mathbf{l}_1^N, \mathbf{d}_1^K}) = \underset{\mathbf{l}_1^N, \mathbf{d}_1^K}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{l}_1^N, \mathbf{d}_1^K|\mathbf{c}_1^N) \tag{7}$$

Durations $\mathbf{d}_1^K$ being independent and identically distributed, and assuming that the joint sequences $\{\mathbf{l}_{n_{k-1}}^N, \mathbf{c}_{n_{k-1}}^N\}_{k=1}^K$ are independent conditionally on $\mathbf{d}_1^K$, its distribution is given by:

$$\mathrm{p}(\mathbf{l}_1^N, \mathbf{d}_1^K|\mathbf{c}_1^N) = \prod_{k=1}^{K} \underbrace{\mathrm{p}_l(k)}_{\substack{\text{linguistic}\\ \text{constraint}}} \times \underbrace{\mathrm{p}_m(k)}_{\substack{\text{metric}\\ \text{constraint}}} \tag{8}$$

where:

$$\mathrm{p}_l(k) = \mathrm{p}(\mathbf{l}_{n_{k-1}+1}^{n_k}|\mathbf{c}_{n_{k-1}+1}^{n_k}) \tag{9}$$

$$= \prod_{n=n_{k-1}+1}^{n_k} \mathrm{p}(l_n|l_{n-1})\frac{\mathrm{p}(l_n|\mathbf{c}_n)}{\mathrm{p}(l_n)} \tag{10}$$

$$\mathrm{p}_m(k) = \mathrm{p}(d_k) \tag{11}$$

The notations $\mathrm{p}_l(k)$ and $\mathrm{p}_m(k)$ are introduced here as short-cuts for the contribution of the linguistic and the metric constraints whose expressions are presented in equations (10) and (11). These notations will be used for clarity in the remaining of the paper (especially in section III-C).

The solution to equation (8) is obtained by using a dynamic programming algorithm which is a reformulation of the standard VITERBI decoding algorithm (VA) (as detailed in [45]). The main modification to the standard VITERBI search stands in the add of all possible major phrase segmentations as a supplementary dimension in the search space. The remaining of this section describes the details of the dynamic programming algorithm.

Define δ_n the log-probability of the most-likely sequence of major phrase lengths $\mathbf{d}_1^k$ and the corresponding sequence of prosodic events $\mathbf{l}_1^n$ that end a major phrase at time n, conditionally to the sequence of linguistic contexts $\mathbf{c}_1^n$:

$$\delta_n = \max_{\mathbf{l}_1^n, \mathbf{d}_1^k} \log \mathrm{p}(\mathbf{l}_1^n, \mathbf{d}_1^k|\mathbf{c}_1^n) \tag{12}$$

The traceback information is stored in ψ_n, which contains the length of the major phrase d_k and the sequence of prosodic events $\mathbf{l}_{n-d_k+1}^n$ corresponding to the k-th major phrase conducting to δ_n.

Then, the decoding algorithm can be described as follows:

■ initialization: $n = 1, d_1 = 1$

$$\delta_1 = \log \mathrm{p}(l_1)\mathrm{p}(d_1)\mathrm{p}(l_1|\mathbf{c}_1) \tag{13}$$

■ recursion: $n \in [2, N], d_k \in [1, D]$

$$\delta_n = \max_{\mathbf{l}_{n-d_k+1}^n, d_k} \delta_{n-d_k} + \log \mathrm{p}_l(k) + \log \mathrm{p}_m(k) \tag{14}$$

$$= \max_{\mathbf{l}_{n-d_k+1}^n, d_k} \delta_{n-d_k} + \log \mathrm{p}(\mathbf{l}_{n-d_k+1}^n|\mathbf{c}_{n-d_k+1}^n) + \log \mathrm{p}(d_k) \tag{15}$$

$$\psi_n = \underset{\mathbf{l}_{n-d_k+1}^n, d_k}{\operatorname{argmax}} \delta_{n-d_k} + \log \mathrm{p}(\mathbf{l}_{n-d_k+1}^n|\mathbf{c}_{n-d_k+1}^n) + \log \mathrm{p}(d_k) \tag{16}$$

Here, the calculation of δ_n requires to compute for each possible major phrase length d_k the posterior probability of the most-likely sequence of prosodic events corresponding to the major phrase.

Finally, the most-likely sequences of major phrases and prosodic events $(\widehat{\mathbf{l}_1^N, \mathbf{d}_1^K})$ are retrieved through backtracking:

■ initialization: $n = N$, $k = K$

$$(\widehat{\mathbf{l}_{N-d_K+1}^N, d_K}) = \psi_N \tag{17}$$

■ recursion: $n \in [1, N[$

$$n' = n - d_k \tag{18}$$

$$k = k - 1 \tag{19}$$

$$(\widehat{\mathbf{l}_{n'-d_K+1}^{n'}, d_k}) = \psi_{n'} \tag{20}$$

Here, the number K of major phrases remains actually unknown until the backtracking is completed.

C. Segmental HMMs & Dempster-Shafer Fusion

In equation (8), the linguistic and the metric probabilities are equally considered. However, the linguistic and the metric constraints are not necessarily equally important for the segmentation into major phrases. Consequently, a proper fusion of the linguistic and the metric probabilities into the segmental HMMs must be formulated. The Dempster-Shafer fusion is here presented to perform the statistical fusion of the linguistic and metric constraints in the segmental HMM.

Dempster-Shafer theory of evidence [47] is a mathematical theory commonly used as a method for sensor fusion in statistical signal processing. In particular, Dempster-Shafer

theory provides a powerful framework for information fusion in which the *reliability* that can be conferred to different sources of information can be explicitly formulated. In the Dempster-Shafer fusion, probability density functions (PDFs) can be reformulated into mass functions (MFs) to account for the *reliability* that can be conferred to each source of information, and then combined with the Dempster-Shafer fusion rule. The principle of the Dempster-Shafer combination is shortly described, and its integration into the segmental HMM for the segmentation into major phrases is detailed.

1) Mass Function: An elementary mass function m is a function of $\mathcal{P}(\mathcal{C})$ in $\mathbb{R}_+$ that verify:

$$\begin{cases} \mathrm{m}(\emptyset) = 0 \\ \sum_{\mathcal{A}\in\mathcal{P}(\mathcal{C})} = 1 \end{cases} \tag{21}$$

where $\mathcal{C}$ is the state alphabet, and $\mathcal{P}(\mathcal{C})$ is the power set of $\mathcal{C}$.

Mass functions present the advantage over standard probabilities that a mass can be assigned to composite classes rather than singletons only. It allows one to specify a degree of *ignorance* instead of being forced to supply probabilities that add to unity. It can be used to model the reliability of a source of information during the fusion by assigning more or less weight to the composite class composed of all the classes.

2) Dempster-Shafer Fusion: The Dempster-Shafer fusion of two masses is given by:

$$m(A) = (m_1 \oplus m_2)(A) \tag{22}$$
$$\propto \sum_{B_1\cap B_2=A} m_1(B_1) \times m_2(B_2) \tag{23}$$

Hence, the Dempster-Shafer fusion of a mass m and a probability p is a probability given by:

$$(m \oplus p)(x) = \frac{\sum_{x\in u} m(u)p(x)}{\sum_{x'\in\mathcal{C}} \sum_{x'\in u'} m(u')p(x')} \tag{24}$$

where $m(u)$ ($u \in \mathcal{P}(\mathcal{C})$) denotes the mass associated with a source of information for which the reliability may vary and p the probability associated with another source of information.

In order to control the relative importance of the linguistic constraint $\mathrm{p}_l(k)$ and the metric constraint $\mathrm{p}_m(k)$ during their combination in the segmental HMM (equation (8)), PDFs are replaced by the following mass functions (MFs):

$$m_l(k) = \alpha\ \mathrm{p}_l(k) \qquad m_l(\mathcal{C}) = 1-\alpha \tag{25}$$
$$m_m(k) = \beta\ \mathrm{p}_m(k) \qquad m_m(\mathcal{C}) = 1-\beta \tag{26}$$

where $\alpha \in [0,1]$ and $\beta \in [0,1]$ are coefficients which denote the reliability that is associated with the linguistic constraint $p_l(k)$ and metric constraint $p_m(k)$ respectively, and $m_l(\mathcal{C})$ and $m_m(\mathcal{C})$ model the ignorance of each source.

The Dempster-Shafer fusion of m_l and m_m is then given by:

$$(m_l \oplus m_m)(k) \propto \alpha(1-\beta)\mathrm{p}_l(k) + \alpha\beta\ \mathrm{p}_l(k)\mathrm{p}_m(k) + \beta(1-\alpha)\mathrm{p}_m(k) \tag{27}$$

Hence,

$$(m_l \oplus m_m)(k) \propto \begin{cases} \mathrm{p}_l(k), & \alpha=1,\ \beta=0 \quad ① \\ \mathrm{p}_m(k), & \alpha=0,\ \beta=1 \quad ② \\ \mathrm{p}_l(k)\ \mathrm{p}_m(k), & \alpha=1,\ \beta=1 \quad ③ \end{cases} \tag{28}$$

① denotes that only the metric probability is considered (metric constraint), ② denotes that only linguistic probability is considered (linguistic constraint), and ③ denotes that the linguistic and metric probabilities are equally considered (linguistic/metric constraints).

The reliability coefficients α and β are rewritten into a single coefficient $\gamma = (\alpha, \beta)$ so that the fusion is always a probability, and so that the relative importance of linguistic and the metric probabilities is linearly interpolated from the metric constraint solely to the linguistic constraint solely. Thus: $\gamma = -1$ will refer to $\alpha = 0$ and $\beta = 1$, $\gamma = 0$ to $\alpha = 1$ and $\beta = 1$, and $\gamma = +1$ to $\alpha = 1$ and $\beta = 0$.

Finally, the Dempster-Shafer fusion of the linguistic and the metric constraints as expressed in equation (27) is used to replace the standard fusion in equation (8):

$$(\widehat{\mathbf{l}_1^N, \mathbf{d}_1^K}) = \underset{\mathbf{l}_1^N, \mathbf{d}_1^K}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{l}_1^N, \mathbf{d}_1^K | \mathbf{c}_1^N) \tag{29}$$
$$= \underset{\mathbf{l}_1^N, \mathbf{d}_1^K, \gamma}{\operatorname{argmax}} \prod_{k=1}^{K} (m_l \oplus m_m)(k) \tag{30}$$

The decoding algorithm is simply rewritten in order to account for the combination of the metric and linguistic constraints $(m_l \oplus m_m)(k)$. For each possible coefficient γ, the most-likely sequences of prosodic events and major phrases $(\widehat{\mathbf{l}_1^N, \mathbf{d}_1^K})$ are determined. Then, the most-likely sequences of major phrases and prosodic events $(\widehat{\mathbf{l}_1^N, \mathbf{d}_1^K})$ are determined so as to maximise the conditional probability to the combination γ.

IV. Experiments

In order to investigate the role of the linguistic constraint, and the combination of linguistic and metric constraints for the assignment of prosodic events, two objective experiments were conducted: first, the role of the linguistic constraint for the assignment of prosodic events (accent and phrasing) is addressed in section IV-C. Then, the combination of the linguistic and the metric constraints for the segmentation into major phrases is addressed in section IV-D. The experiments were conducted on a multi-speaker speech database with various speaking styles (read and spontaneous speech). Finally, a subjective experiment was conducted to address the combination of the linguistic and the metric constraints in the context of text-to-speech synthesis.

A. Speech Material and Annotations

This study investigates the assignment of prosodic events in the context of French text-to-speech synthesis. Contrary to research for text-to-speech synthesis in English for which a

number of large speech databases is available with manual annotations of prosodic events (accent and/or major phrases) (e.g., [48], [49]), there is currently no comparable resources available for French. In consequence, a French speech database containing multi-speaker with various speaking styles (read and spontaneous speech) was specifically designed and used for the experiments (see [8] for details). Speaking styles include read speech and a number of more/less spontaneous speech.

- A READ corpus is composed of short sentences, selected in order to design a phonetically well-balanced speech database for text-to-speech synthesis. Each prompt is read by a non-professional French speaker and recorded in an anechoic room (9 hrs.).
- A SPONTANEOUS SPEECH corpus is composed of multi-speaker French spontaneous speech recordings in 4 speaking styles: church offices, political speech, journalistic chronicles, and sports commentary (4 hours in total, and around 1 hour for each speaking style). Speech recordings were collected from broadcast (radio, television).

A short description of the French speech database is presented in table I.

TABLE I
OVERVIEW OF THE SPEECH DATABASES: NUMBER OF SPEAKERS, NUMBER OF UTTERANCES, TOTAL DURATION.

SPEAKING STYLE	# SPEAKER	# UTTERANCE	TOTAL DURATION
READ	1	1030	1h00
CHURCH	7	598	1h20
POLITICS	5	454	1h10
JOURNAL	5	840	1h10
SPORT	4	743	0h35

Firstly, speech recordings come with manually processed text transcriptions. Then, speech-to-text alignment was processed by IRCAMALIGN [50] - a HMM-based speech segmentation system for French based on the HTK toolkit [51] and trained on the BREF [52] multi-speaker French speech database. Front-end processing for speech-to-text alignment includes the LIAPHON system [53] for text-to-phoneme, -syllable, and -word conversion. The resulting alignment was then manually corrected.

Secondly, the alphabet of prosodic events used is a compact representation of speech prosody in terms of accent and phrasing. The alphabet is composed of: a single element for accent (prominence: P), and two degrees of break for phrasing (intermediate break: F_m, and major phrase break F_M, respectively) [54]. Prosodic events were automatically transcribed based on the IRCAMPROM system - a GMM-based system for the automatic transcription of prosodic events [55] -, and then manually corrected. Statistics of prosodic events and phrases for the various speaking styles are presented in table II.

Finally, text parsing (surface and deep) was processed by the ALPAGE linguistic processing chain [39], and then converted into context-dependent labels aligned to the speech signal

TABLE II
STATISTICS OF PROSODIC EVENTS AND PHRASES FOR THE SPEAKING STYLES: MEAN PROPORTION OF PROSODIC EVENTS (%), AND MEAN LENGTH OF PROSODIC PHRASES IN SYLLABLES.

SPEAKING STYLE	PROSODIC EVENTS F_M (%)	F_m (%)	P (%)	TOTAL (%)	PHRASE LENGTH (syl.)
READ	11	10	5	26	7.5
CHURCH	14	10	11	35	5.5
POLITICS	13	7	10	30	5.3
JOURNAL	6	12	7	25	10.1
SPORT	17	7	10	34	6.8

(as presented in section II). In this study, prosodic events and context-dependent labels were aligned on syllables in order to provide a common unit for the assignment of all prosodic events: accents (mostly, aligned on the syllable), and breaks (mostly, aligned on the word). Also, linguistic information extracted from text and computed on word unit were aligned on syllable unit, and then converted into context-dependent labels. Additional position/number information were added to the context-dependent labels: position/number of units (syllable/word/chunk/adjunction) within larger units (word/chunk/adjunction/sentence) - also aligned on syllable unit.

B. Experimental Procedure

Objective experiments were conducted in order to investigate: 1) the role of the linguistic constraint, and 2) the combination of the linguistic and the metric constraints for the assignment of prosodic events. For the role of the linguistic constraint: the assignment of prosodic events (accent P and phrasing F_m and F_M) is assessed depending on the nature of the linguistic information: morpho-syntactic (M), dependency (D), constituency (C), and adjunction (A). For the combination of linguistic and metric constraints: the segmentation of a text into major phrases (F_M) is assessed depending on the combination of linguistic and metric constraints, and the nature of the linguistic information.

Experiments were conducted based on a 10-fold cross-validation. First, the whole database was split into 10 folds. For each fold, linguistic (linguistic constraint) and metric (metric constraint) PDFs were estimated on the 10 folds minus one. Then, the sequence of prosodic events was determined for each sentence of the remaining fold. Finally, the assigned sequence of prosodic events are compared with the reference sequence of prosodic events. The F-measure (harmonic mean of recall and precision) was used to measure the performance of the assignment. Experiments were conducted with comparison to a STANDARD model based on surface information only (M), and a BASELINE punctuation rule-based model (PUNC) in which a major phrase break (F_M) is inserted after each punctuation marker.

C. Role of Linguistic Contexts

Tables III and IV summarize the mean performance and 95% confidence interval obtained for the read and spontaneous speech databases with a standard HMM (as defined in section III-B2). For clarity, tables III and IV report only the performance obtained for the best combinations of linguistic

information, and the individual performance of all linguistic information.

TABLE III
ROLE OF THE LINGUISTIC CONTEXT ON READ SPEECH: MEAN F-MEASURE (AND 95% CONFIDENCE INTERVAL) FOR THE ASSIGNMENT OF PROSODIC EVENTS (F_M, F_m, P) DEPENDING ON THE LINGUISTIC CONTEXT (M/D/C/A).

READ			
	F_M	F_m	P
MDCA	**94.8 (±0.9)**	**55.0** (±1.5)	**34.5 (±1.1)**
MCA	94.2 (±0.7)	53.3 (±1.6)	33.8 (±1.2)
MDA	91.8 (±1.1)	52.3 (±1.1)	33.0 (±1.3)
MDC	83.7 (±1.1)	48.9 (±1.3)	34.4 (±1.2)
CA	94.5 (±1.2)	53.7 (±1.4)	34.3 (±1.1)
⋮	⋮	⋮	⋮
A	90.9 (±0.9)	21.4 (±1.3)	0.7 (±1.9)
C	83.2 (±1.0)	39.9 (±1.4)	4.2 (±1.3)
D	60.5 (±1.2)	30.2 (±1.1)	2.2 (±1.6)
M	74.3 (±1.1)	43.5 (±1.2)	33.1 (±1.4)
PUNC.	65.9 (±1.4)	-	-

The role of the linguistic information depends on the prosodic event. Firstly, the assignment of breaks presents substantial (major phrase break: F_M) and moderate (intermediate break: F_m) performance, and accent (P) presents poor performance only. Secondly, the deep linguistic information does not uniformly affects the assignment of prosodic events. On the one hand, the improvement for the assignment of phrase breaks is substantial: drastically significant for major phrase breaks (F_M), and significant for intermediate breaks (F_m) (from surface syntactic information (M) to deep syntactic information (MDCA)). In particular, the constituency and adjunction information (C, A) are particularly significant for the assignment of breaks. On the other hand, the improvement for the assignment of accents (P) is negligible: the surface syntactic information (M) is robust, and the deep syntactic information is insignificant (D, C, A).

This proves the role of deep syntactic information for the assignment of prosodic events. In particular, constituency and adjunction information - as global syntactic constructions - are proved to be extremely significant for the assignment of phrase breaks. This indicates that the prosodic structure more closely relates to global syntactic information (associated with large syntactic units) rather than on local syntactic information (associated with small syntactic units). Finally, the syntactic information is not sufficient for the assignment of accents (which mainly reflects semantic focus).

Additionally, there is a large difference for the assignment of prosodic events between read and spontaneous speech, especially for major and intermediate phrase breaks. This is principally due to the syntactic and prosodic variability of spontaneous speech compared to that of read speech. In particular, the syntactic parsing is less reliable in spontaneous speech, and the prosodic structure depends less on the syntactic information in spontaneous speech (e.g., pragmatics, discourse).

D. Fusion of Linguistic and Metric Constraints

Tables V and VI summarize the mean performance and 95% confidence interval obtained for the assignment of major phrase breaks (F_M) for various combinations of linguistic and metric constraints: individual contribution of the METRIC constraint ($\gamma = (\alpha = 0, \beta = 1) = -1$), individual contribution of the LINGUISTIC constraint ($\gamma = (\alpha = 1, \beta = 0) = +1$), standard fusion of the METRIC/LINGUISTIC constraint ($\gamma = (\alpha = 1, \beta = 1) = 0$, as in the segmentation HMM), and the optimal FUSION of metric/linguistic constraints.

TABLE V
FUSION OF METRIC/LINGUISTIC CONSTRAINTS ON READ SPEECH: MEAN F-MEASURE (AND 95% CONFIDENCE INTERVAL) FOR THE ASSIGNMENT OF MAJOR BREAKS (F_M) OBTAINED WITH THE METRIC CONSTRAINT, THE METRIC/LINGUISTIC CONSTRAINTS, THE LINGUISTIC CONSTRAINT, AND THE FUSION OF METRIC/LINGUISTIC CONSTRAINTS.

READ				
	METRIC	METRIC/ LINGUISTIC	LINGUISTIC	FUSION
MDCA	**62.1 (±2.5)**	**92.1 (±1.1)**	**94.8 (±0.9)**	**96.6 (±0.8)**
MCA	62.1 (±2.5)	91.9 (±0.9)	94.2 (±0.7)	96.0 (±0.8)
MDA	62.1 (±2.5)	89.4 (±1.1)	91.8 (±1.1)	93.3 (±1.2)
MDC	62.1 (±2.5)	80.5 (±1.2)	83.7 (±1.1)	86.8 (±1.1)
CA	62.1 (±2.5)	92.0 (±1.5)	94.5 (±1.2)	95.9 (±0.9)
⋮	⋮	⋮	⋮	⋮
A	62.1 (±2.5)	87.2 (±1.4)	90.9 (±0.9)	92.7 (±1.2)
C	62.1 (±2.5)	80.7 (±1.2)	83.2 (±1.0)	87.3 (±1.4)
D	62.1 (±2.5)	58.1 (±1.5)	60.5 (±1.2)	63.2 (±1.2)
M	62.1 (±2.5)	72.2 (±1.4)	74.3 (±1.1)	78.5 (±1.1)

On the one hand, the standard fusion of the linguistic/metric constraints does not outperform the linguistic constraint in most of the cases. This indicates that a standard fusion of the linguistic/metric constraints does not suffice to exploit the information provided by the metric constraint. On the other hand, the proposed fusion of linguistic/metric constraints significantly outperforms the standard linguistic/metric constraints and the linguistic constraint in most of the cases. This confirms that a proper fusion of the the linguistic/metric constraints successfully improves the segmentation into major phrases (i.e., under and over segmentations caused by the linguistic constraint only).

E. Subjective Experiment

Finally, a subjective experiment was conducted to compare the quality of the LINGUISTIC, METRIC/LINGUISTIC, and FUSION OF METRIC/LINGUISTIC in speech synthesis. For this purpose, 20 sentences were randomly selected from the French fairy-tale "Le Petit Poucet" ("Little Tom Thumb") by Charles Perrault, and used to synthesize speech utterances for each system. The IRCAMTTS unit-selection speech synthesis system was used for the comparison [56], with the voice of the French speaker of the READ database. The MDCA was used as linguistic contexts, and the model parameters were estimated on the read speech database.

For each sentence, the text was segmented into major phrases according to the considered constraints, and then synthesized with the IRCAMTTS speech synthesis system.

TABLE IV

ROLE OF THE LINGUISTIC CONTEXT ON SPONTANEOUS SPEECH: MEAN F-MEASURE (AND 95% CONFIDENCE INTERVAL) FOR THE ASSIGNMENT OF PROSODIC EVENTS (F_M, F_m, P) DEPENDING ON THE LINGUISTIC CONTEXT (M/D/C/A).

CHURCH

	F_M	F_m	P
MDCA	79.8 (±1.3)	42.0 (±1.4)	30.2 (±1.4)
MCA	**81.1 (±1.2)**	**43.3 (±1.2)**	**31.0 (±1.3)**
MDA	75.6 (±1.4)	40.7 (±1.4)	30.0 (±1.4)
MDC	75.3 (±1.1)	39.6 (±1.5)	29.5 (±1.2)
CA	80.1 (±1.2)	42.5 (±1.3)	5.2 (±2.6)
⋮	⋮	⋮	⋮
A	71.5 (±1.1)	18.3 (±1.8)	1.2 (±2.3)
C	72.2 (±1.3)	36.8 (±1.5)	3.6 (±1.8)
D	42.1 (±1.9)	22.2 (±1.6)	4.1 (±2.0)
M	64.1 (±2.1)	27.3 (±1.5)	24.4 (±1.8)
PUNC.	N/A	-	-

POLITICS

	F_M	F_m	P
MDCA	78.5 (±1.2)	40.5 (±1.6)	29.8 (±1.3)
MCA	**79.9 (±1.2)**	**41.6 (±1.4)**	**30.4 (±1.5)**
MDA	72.3 (±1.5)	39.6 (±1.5)	29.5 (±1.4)
MDC	74.2 (±1.4)	37.2 (±1.6)	28.9 (±1.4)
CA	80.1 (±1.1)	40.4 (±1.5)	4.7 (±1.8)
⋮	⋮	⋮	⋮
A	63.5 (±1.4)	21.4 (±1.3)	2.1 (±1.7)
C	73.1 (±1.5)	39.9 (±1.4)	4.7 (±1.8)
D	40.5 (±1.5)	30.2 (±1.1)	3.4 (±2.1)
M	62.5 (±1.3)	23.5 (±1.2)	25.3 (±1.6)
PUNC.	N/A	-	-

JOURNAL

	F_M	F_m	P
MDCA	**71.1 (±1.0)**	**46.2** (±1.5)	**28.2 (±1.3)**
MCA	68.5 (±1.4)	45.4 (±1.4)	27.8 (±1.4)
MDA	64.0 (±1.2)	42.1 (±1.4)	27.9 (±1.3)
MDC	64.4 (±1.4)	40.5 (±1.5)	27.5 (±1.2)
CA	64.9 (±1.3)	44.4 (±1.5)	4.3 (±1.4)
⋮	⋮	⋮	⋮
A	61.8 (±1.3)	23.2 (±1.5)	0.8 (±1.7)
C	63.1 (±1.1)	40.4 (±1.5)	5.5 (±1.5)
D	52.1 (±1.2)	17.8 (±1.4)	4.1 (±1.9)
M	56.4 (±1.6)	25.8 (±1.6)	20.2 (±1.5)
PUNC.	N/A	-	-

SPORT

	F_M	F_m	P
MDCA	**79.2 (±1.1)**	**37.3** (±1.4)	**23.5 (±1.5)**
MCA	79.1 (±1.4)	36.8 (±1.4)	23.3 (±1.5)
MDA	75.0 (±1.4)	32.4 (±1.5)	23.1 (±1.4)
MDC	71.2 (±1.3)	28.9 (±1.7)	23.2 (±1.6)
CA	80.2 (±1.1)	35.2 (±1.5)	3.1 (±1.8)
⋮	⋮	⋮	⋮
A	71.1 (±1.2)	15.5 (±1.6)	0.5 (±1.7)
C	70.1 (±1.1)	31.8 (±1.5)	2.5 (±2.1)
D	62.0 (±1.4)	18.5 (±1.8)	1.3 (±1.8)
M	65.2 (±1.8)	25.6 (±1.5)	19.8 (±2.3)
PUNC.	N/A	-	-

TABLE VI

FUSION OF METRIC/LINGUISTIC CONSTRAINTS ON SPONTANEOUS SPEECH: MEAN F-MEASURE (AND 95% CONFIDENCE INTERVAL) FOR THE ASSIGNMENT OF MAJOR BREAKS (F_M) OBTAINED WITH THE METRIC CONSTRAINT, THE METRIC/LINGUISTIC CONSTRAINTS, THE LINGUISTIC CONSTRAINT, AND THE FUSION OF METRIC/LINGUISTIC CONSTRAINTS.

CHURCH

	METRIC	METRIC/ LINGUISTIC	LINGUISTIC	FUSION
MDCA	35.8 (±3.6)	77.9 (±2.0)	79.8 (±1.3)	83.7 (±1.4)
MCA	**35.8 (±3.6)**	**79.3 (±1.6)**	**81.1 (±1.2)**	**84.0 (±1.4)**
MDA	35.8 (±3.6)	72.4 (±1.1)	75.6 (±1.4)	77.3 (±1.2)
MDC	35.8 (±3.6)	74.9 (±1.2)	75.3 (±1.1)	78.1 (±1.4)
CA	35.8 (±3.6)	70.0 (±1.5)	80.1 (±1.2)	81.8 (±1.1)
⋮	⋮	⋮	⋮	⋮
A	35.8 (±3.6)	70.4 (±1.5)	71.5 (±1.1)	73.3 (±1.3)
C	35.8 (±3.6)	70.7 (±1.0)	72.2 (±1.3)	74.0 (±1.2)
D	35.8 (±3.6)	38.2 (±2.1)	42.1 (±1.9)	42.4 (±1.7)
M	35.8 (±3.6)	61.8 (±2.4)	64.1 (±2.1)	64.7 (±1.5)

POLITICS

	METRIC	METRIC/ LINGUISTIC	LINGUISTIC	FUSION
MDCA	33.8 (±2.8)	74.3 (±1.6)	78.5 (±1.2)	81.0 (±1.3)
MCA	**33.8 (±2.8)**	**75.9 (±1.4)**	**79.9 (±1.2)**	**82.1 (±1.1)**
MDA	33.8 (±2.8)	69.1 (±1.2)	72.3 (±1.5)	74.5 (±1.4)
MDC	33.8 (±2.8)	72.3 (±1.4)	74.2 (±1.4)	77.8 (±1.3)
CA	33.8 (±2.8)	73.9 (±1.4)	80.1 (±1.1)	81.5 (±1.2)
⋮	⋮	⋮	⋮	⋮
A	33.8 (±2.8)	58.6 (±1.5)	63.5 (±1.4)	65.2 (±1.2)
C	33.8 (±2.8)	70.1 (±1.3)	73.1 (±1.5)	76.3 (±1.3)
D	33.8 (±2.8)	43.4 (±1.3)	40.5 (±1.5)	43.1 (±1.3)
M	33.8 (±2.8)	59.1 (±1.5)	62.5 (±1.3)	62.8 (±1.2)

JOURNAL

	METRIC	METRIC/ LINGUISTIC	LINGUISTIC	FUSION
MDCA	**50.1 (±5.1)**	**69.0 (±1.3)**	**71.1 (±1.0)**	**74.3 (±1.1)**
MCA	50.1 (±5.1)	65.7 (±1.5)	68.5 (±1.4)	71.4 (±1.2)
MDA	50.1 (±5.1)	62.8 (±1.2)	64.0 (±1.2)	66.5 (±1.1)
MDC	50.1 (±5.1)	63.9 (±1.3)	64.4 (±1.4)	67.8 (±1.3)
CA	50.1 (±5.1)	61.1 (±1.2)	64.9 (±1.3)	67.0 (±1.1)
⋮	⋮	⋮	⋮	⋮
A	50.1 (±5.1)	60.8 (±1.5)	61.8 (±1.3)	64.2 (±1.3)
C	50.1 (±5.1)	62.7 (±1.3)	63.1 (±1.1)	65.2 (±1.2)
D	50.1 (±5.1)	54.2 (±1.4)	52.1 (±1.2)	57.3 (±1.8)
M	50.1 (±5.1)	56.5 (±2.1)	56.4 (±1.6)	58.9 (±1.8)

SPORT

	METRIC	METRIC/ LINGUISTIC	LINGUISTIC	FUSION
MDCA	**59.2 (±2.3)**	**75.1 (±2.1)**	**79.2 (±1.1)**	**83.2 (±1.2)**
MCA	59.2 (±2.3)	75.2 (±1.2)	79.1 (±1.4)	82.1 (±1.1)
MDA	59.2 (±2.3)	72.3 (±1.5)	75.0 (±1.4)	77.8 (±1.3)
MDC	59.2 (±2.3)	72.5 (±1.6)	71.2 (±1.3)	74.9 (±1.4)
CA	59.2 (±2.3)	76.4 (±1.4)	80.2 (±1.1)	81.8 (±1.2)
⋮	⋮	⋮	⋮	⋮
A	59.2 (±2.3)	71.7 (±1.2)	71.1 (±1.2)	73.4 (±1.1)
C	59.2 (±2.3)	69.8 (±1.4)	70.1 (±1.1)	71.5 (±1.3)
D	59.2 (±2.3)	65.3 (±1.2)	62.0 (±1.4)	66.2 (±1.1)
M	59.2 (±2.3)	67.2 (±2.1)	65.2 (±1.8)	67.8 (±2.3)

10 native French speakers participated in the experiment. The evaluation was conducted according to a *crowd-sourcing* technique using social networks. Pairs of synthesized speech utterances were randomly presented to the participants who were asked to attribute a preference score according to the *naturalness* of the speech utterances on the comparison mean opinion score (CMOS) scale [57]. Participants were encouraged to use headphones.

The preference score obtained for the comparison of the LINGUISTIC, METRIC/LINGUISTIC, and FUSION OF METRIC/LINGUISTIC in text-to-speech synthesis are presented in Figure 6. The preference scores obtained are 41.0% for the FUSION OF METRIC/LINGUISTIC constraints, 28.5% for the LINGUISTIC constraint, 20.5% for the METRIC/LINGUISTIC constraints, and 10% for no preference. This confirms that the fusion of the metric/linguistic constraints constitutes a qualitative advance for text-to-speech synthesis.

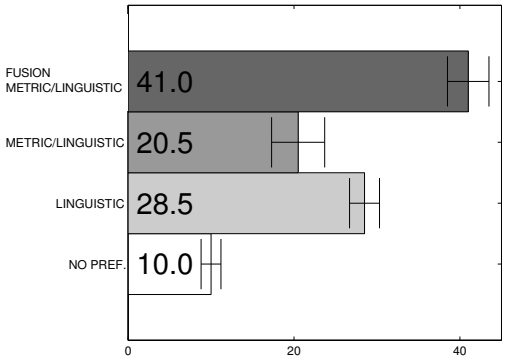

Fig. 6. Mean preference score and 95% confidence interval obtained for: the fusion of the linguistic/metric constraints, the linguistic/metric constraints, the linguistic constraint, and no preference.

V. CONCLUSION

This paper explored the fusion of linguistic and metric constraints for the assignment of prosodic events for text-to-speech synthesis. Firstly, a linguistic processing chain was presented in order to provide the surface/deep syntactic structure of a text for the assignment of prosodic events. Secondly, a hierarchical HMM (HHMM) was introduced to model the prosodic structure of a text: the root layer represents the text, each intermediate layer a sequence of intermediate phrases, the pre-terminal layer the sequence of accents, and the terminal layer the sequence of linguistic contexts. For each intermediate layer, a segmental HMM and Dempster-Shafer fusion are used to combine linguistic and metric constraints for the segmentation of a text into major phrases. A set of experiments conducted on multi-speaker databases with various speaking styles confirms: the role of a deep linguistic representation of a text for the assignment of prosodic events, and the role of the fusion of linguistic and metric constraints for the segmentation of a text into major phrases. This constitutes a substantial advance for the modeling of the speech prosody for text-to-speech synthesis.

ACKNOWLEDGEMENTS

The authors would like to thank Xavier Rodet (professor, IRCAM) for his supervision and his lifetime investment into speech research, Anne Lacheret (professor, MoDyCo Lab. - University of Paris X) for her expertise into linguistics and prosody, and Eric de la Clergerie (researcher, ALPAGE-INRIA) for his expertise into syntax and his help for the use of the ALPAGE linguistic processing chain.

REFERENCES

[1] A. Hunt and A. Black, "Unit Selection in a Concatenative Speech Synthesis System using a Large Speech Database," in *International Conference on Audio, Speech, and Signal Processing*, 1996, pp. 373–376.
[2] H. Zen, K. Tokuda, and A. Black, "Statistical Parametric Speech Synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.
[3] M. Ostendorf and N. Veilleux, "A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary location," *Journal of Computational Linguistics*, vol. 20, no. 1, pp. 27–54, 1994.
[4] K. Ross and M. Ostendorf, "Prediction of Abstract Prosodic Labels for Speech Synthesis," *Computer Speech and Langage*, vol. 10, pp. 155–185, 1996.
[5] A. W. Black and P. Taylor, "Assigning Phrase Breaks from Part-of-Speech Sequences," in *European Conference on Speech Communication and Technology*, Rhodes, Greece, 1997, pp. 995–998.
[6] I. Bulyko and M. Ostendorf, "Joint Prosody Prediction and Unit Selection for Concatenative Speech Synthesis," in *International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, USA, 2005, pp. 781–784.
[7] H. Schmid and M. Atterer, "New Statistical Methods for Phrase Break Prediction," in *International Conference On Computational Linguistics*, Geneva, Switzerland, 2004, pp. 659–665.
[8] N. Obin, "MeLos: Analysis and Modelling of Speech Prosody and Speaking Style," PhD. Thesis, IRCAM - UPMC, 2011.
[9] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-based Speech Synthesis," in *European Conference on Speech Communication and Technology*, Budapest, Hungary, 1999, pp. 2347–2350.
[10] T. Toda and K. Tokuda, "A Speech Parameter Generation Algorithm Considering Global Variance for HMM-based Speech Synthesis," *IEICE Transactions on Information and Systems*, vol. 90, no. 5, pp. 816–824, 2007.
[11] Y. Qian, Z. Wu, and F. K. Soong, "Improved Prosody Generation by Maximizing Joint Likelihood of State and Longer Units," in *International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, 2009, pp. 3781–3784.
[12] Z.-J. Yan, Y. Qian, and F. K. Soong, "Rich Context Modeling for High Quality HMM-based TTS," in *Interspeech*, Brighton, UK, 2009, pp. 4025–4028.
[13] P. Bell, T. Burrows, and P. Taylor, "Adaptation of Prosodic Phrasing Models," in *Speech Prosody*, Dresden, Germany, 2006.
[14] J. Yamagishi, "Average-Voice-based Speech Synthesis using HSMM-based Speaker Adaptation and Adaptive Training," *IEICE Transactions on Information and Systems*, vol. 90, no. 2, pp. 533–543, 2007.
[15] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: a Standard for Labeling English Prosody," in *International Conference of Spoken Language Processing*, Banff, Canada, 1992, pp. 867–870.
[16] D. Hirst, A. Di Cristo, and R. Espresser, *Prosody: Theory and Experiments*. M. Horne, 2000, ch. Levels of Representation and Levels of Analysis for the Description of Intonation Systems.
[17] B. Post, *Tonal and Phrasal Structures in French Intonation*. The Hague: Academic Graphics, 2000.
[18] P. Mertens, "The Prosogram : Semi-Automatic Transcription of Prosody based on a Tonal Perception Model," in *Speech Prosody*, Nara, Japan, 2004, pp. 549–552.
[19] W. E. Cooper and J. Paccia-Cooper, *Syntax and Speech*. Cambridge: Harvard University Press, 1980.
[20] J. Gee and F. Grosjean, "Performance Structures: a Psycholinguistic and Linguistic Appraisal," *Cognitive Psychology*, vol. 15, pp. 411–458, 1983.
[21] E. Selrik, *Phonology and Syntax: The Relation between Sound and Structure*. Cambridge: MIT Press, 1984.
[22] F. Ferreira, "Planning and Timing in Sentence Production: The Syntax-to-Phonology Conversion," PhD. Thesis, University of Massachusetts, 1988.

[23] S. Abney, "Prosodic Structure, Performance Structure and Phrase Structure," in *Human Langage Technology: Proceedings of the workshop on Speech and Natural Language*. Morristown, NJ, USA: Association for Computational Linguistics, 1992, pp. 425–428.
[24] D. Watson and E. Gibson, "The Relationship between Intonational Phrasing and Syntactic Structure in Language Production," *Language and Cognitive Processes*, vol. 6, pp. 713–755, 2004.
[25] F. Dell, "L'Accentuation dans les Phrases en Français," *Forme Sonore du Langage: Structure des Représentation en Phonologie*, pp. 65–122, 1984.
[26] G. Bailly, "Integration of Rhythmic and Syntactic Constraints in a Model of Generation of French Prosody," *Speech Communication*, vol. 8, no. 2, pp. 137–146, 1989.
[27] P. Monnin and F. Grosjean, "Les Structures de Performance en Français : Caractérisation et Prédiction," *L'année Psychologique*, vol. 93, pp. 9–30, 1993.
[28] R. D. Ladd, *Intonational Phonology*. Cambridge: Cambridge University Press, 1996.
[29] E. Delais-Roussarie, "Vers une Nouvelle Approche de la Structure Prosodique," *Langue Française*, vol. 126, 2000.
[30] P. Fraisse, *Psychologie du Rythme*. Presses Universitaires de France, 1974.
[31] O. Watts, J. Yamagishi, and S. King, "Unsupervised Continuous-Valued Word Features for Phrase-Break Prediction without a Part-Of-Speech Tagger," in *Interspeech*, Florence, Italy, 2011, pp. 2157–2160.
[32] O. Watts, S. Gangireddy, Y. Yamagishi, S. King, S. Renals, A. Stan, and M. Giurgiu, "Neural Net Word Representations for Phrase-Break Prediction without a Part Of Speech Tagger," in *International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, 2014, pp. 2618–2622.
[33] J. Hirschberg, "Using Text Analysis to Predict Intonational Boundaries," in *European Conference on Speech Communication and Technology*, Genova, Italy, 1991, pp. 1275–1278.
[34] A. Black and P. Taylor, "Assigning Intonation Elements And Prosodic Phrasing For English Speech Synthesis From High Level Linguistic Input," in *International Conference on Spoken Language Processing*, Yokohama, Japan, 1994, pp. 715–718.
[35] M. Atterer and E. Klein, "Integrating Linguistic and Performance-Based Constraints for Assigning Phrase Breaks," in *International Conference on Computational Linguistics*, Taipei, Taiwan, 2002, pp. 995–998.
[36] N. Obin, P. Lanchantin, A. Lacheret, and X. Rodet, "Reformulating Prosodic Break Model into Segmental HMMs and Information Fusion," in *Interspeech*, Florence, Italy, 2011, pp. 1829–1832.
[37] B. Sagot, "The Lefff, a Freely Available and Large-Coverage Morphological and Syntactic Lexicon for French," in *International Conference on Language Ressources and Evaluation*, Valletta, Malte, 2010, pp. 2744–2751.
[38] B. Sagot and P. Boullier, "From Raw Corpus to Word Lattices: Robust Pre-parsing Processing with SxPipe," *Archives of Control Sciences. Special Issue on Language and Technology. Human Language Technologies as a Challenge for Computer Science and Linguistics*, vol. 15, no. 4, pp. 653–662, 2005.
[39] E. Villemonte de La Clergerie, "From Metagrammars to Factorized TAG/TIG Parsers," in *International Workshop On Parsing Technology*, Vancouver, Canada, Oct. 2005, pp. 190–191.
[40] A. Joshi, L. Levy, and M. Takahashi, "Tree Adjunct Grammars," *Journal of the Computer and System Sciences*, vol. 10, no. 1, pp. 136–163, 1975.
[41] E. Villemonte de La Clergerie, "Convertir des dérivations TAG en dépendances," in *Traitement Automatique des Langues Naturelles*, Montréal, Canada, 2010.
[42] E. Selrik, "On Prosodic Structure and its Relation to Syntactic Structure," in *Nordic Prosody II*, Trondheim, Norway, 1981, pp. 111–140.
[43] S. Fine, Y. Singer, and N. Tishby, "The Hierarchical Hidden Markov Model: Analysis and Applications," *Machine Learning*, vol. 32, p. 41–62, 1998.
[44] M. Gales and S. Young, "Segmental HMMs for Speech Recognition," in *European Conference on Speech Communication and Technology*, Berlin, Germany, 1993, pp. 1579–1582.
[45] M. Ostendorf, V. Digalakis, and O. Kimball, "From HMM's to Segment Models: a Unified View of Stochastic Modeling for Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 360–378, 1996.
[46] J. O'Dell, "The Use of Context in Large Vocabulary Speech Recognition," PhD. Thesis, Cambridge University, 1995.
[47] G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press, 1976.
[48] P. Roach, G. Knowles, T. Varadi, and S. Arnfield, "MARSEC: a Machine-Readable Spoken English Corpus," *Journal of the International Phonetic Association*, vol. 23, no. 1, p. 47–53, 1993.
[49] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus," Boston University Technical Report, Tech. Rep. ECS-95-001, 1995.
[50] P. Lanchantin, A.-C. Morris, X. Rodet, and C. Veaux, "Automatic Phoneme Segmentation with Relaxed Textual Constraints," in *International Conference on Language Resources and Evaluation*, Marrakech, Morroco, 2008, pp. 2403–2407.
[51] S. Young, "The HTK Hidden Markov Model Toolkit: Design and Philosophy," *Entropic Cambridge Research Laboratory, Ltd*, vol. 2, pp. 2–44, 1994.
[52] L. F. Lamel, J.-L. Gauvain, and M. Eskénazi, "BREF, a Large Vocabulary Spoken Corpus for French," in *Eurospeech*, Genova, Italy, 1991, pp. 505–508.
[53] F. Béchet, "Lia_Phon : un système Complet de Phonétisation de Textes," *Traitement Automatique des Langues*, pp. 47–67, 2001.
[54] A. Lacheret, A.-C. Simon, M. Avanzi, and N. Obin, *Rhapsodie: a Prosodic and Syntactic Treebank for Spoken French*. Benjamins, 2013, ch. The Prosodic Annotation of the Rhapsodie Corpus.
[55] N. Obin, X. Rodet, and A. Lacheret-Dujour, "French Prominence: a Probabilistic Framework," in *International Conference on Audio, Speech, and Signal Processing*, Las Vegas, U.S.A, 2008, pp. 3993–3996.
[56] N. Obin, C. Veaux, and P. Lanchantin, "Making Sense of Variations: Introducing Alternatives in Speech Synthesis," in *Speech Prosody*, Shanghai, China, 2012.
[57] " ITU-T P.800. Methods for Subjective Determination of Transmission Quality - Series P: Telephone Transmission Quality; Methods for Objective and Subjective Assessment of Quality," Tech. Rep., 1996.

Nicolas Obin (M'13) is Associate Professor at the Institute for Research and Coordination in Acoustics & Music (IRCAM), and the University of Pierre and Marie Curie (UPMC). He received a MSc degree in Acoustics, Signal Processing, and Computer science applied to Music and a Ph.D. in Computer Sciences from the University of Paris VI in 2006 and 2011, respectively. Also, he received a MSc degree in Musicology, Arts, and Aesthetic from the University of Paris VIII in 2007. During his Ph.D., he studied speech processing, statistical modeling, and computational linguistics for the modeling of speech prosody and speaking style for text-to-speech synthesis. He received the award for the best French Ph.D. thesis in computational sciences from "La Fondation Des Treilles" in 2011. His primary research interests cover signal processing and statistical modeling, for human-computer interaction, speech and music technologies.

Pierre Lanchantin (M'11) is Research Associate in the Speech Research Group at the Cambridge University Engineering Department (CUED). He received a MSc degree in Acoustics, Signal Processing and Computer science applied to Music from Paris VI University and a Ph.D. in Statistical Signal Processing from Telecom SudParis, France. His research interests include statistical modeling of signals, speech processing and their applications to music. During his Ph.D., he studied generalizations of hidden Markov models (HMM) called Pairwise and Triplet Markov chains with applications to image segmentation. He then directed his research toward speech processing and joined the Institute for Research and Coordination in Acoustics & Music (IRCAM), working on speech recognition, speech synthesis and voice conversion. He is currently working on advanced learning and adaptation techniques for speech recognition and synthesis.