HAL
open science

Towards Improved HMM-based Speech Synthesis Using High-Level Syntactical Features.

Nicolas Obin, Pierre Lanchantin, Mathieu Avanzi, Anne Lacheret, Xavier Rodet

To cite this version:

Nicolas Obin, Pierre Lanchantin, Mathieu Avanzi, Anne Lacheret, Xavier Rodet. Towards Improved HMM-based Speech Synthesis Using High-Level Syntactical Features.. Speech Prosody, 2010, Chicago, United States. pp.2000-2004. hal-00589083

HAL Id: hal-00589083
https://hal.science/hal-00589083v1

Submitted on 27 Apr 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

TOWARD IMPROVED HMM-BASED SPEECH SYNTHESIS USING HIGH-LEVEL SYNTACTICAL FEATURES

Nicolas Obin[1], *Pierre Lanchantin*[1], *Mathieu Avanzi*[2,3], *Anne Lacheret-Dujour*[3], Xavier Rodet[1]

[1]Analysis-Synthesis Team, IRCAM, Paris, France
[2] Neuchâtel University, Neuchâtel, Switzerland
[3]Modyco Lab., Paris-Ouest University, Nanterres, France

nobin@ircam.fr, lanchantin@ircam.fr, mathieu.avanzi@unine.ch, anne@lacheret.com, rodet@ircam.fr

ABSTRACT

A major drawback of current Hidden Markov Model (HMM)-based speech synthesis is the monotony of the generated speech which is closely related to the monotony of the generated prosody. Complementary to *model-oriented* approaches that aim to increase the prosodic variability by reducing the "over-smoothing" effect, this paper presents a *linguistic-oriented* approach in which high-level linguistic features are extracted from text in order to improve prosody modeling. A linguistic processing chain based on linguistic preprocessing, morpho-syntactical labeling, and syntactical parsing is used to extract high-level syntactical features from an input text. Such linguistic features are then introduced into a HMM-based speech synthesis system to model prosodic variations (f_0, duration, and spectral variations). Subjective evaluation reveals that the proposed approach significantly improve speech synthesis compared to a baseline model, event if such improvement depends on the observed linguistic phenomenon.

Index Terms— HMM-based speech synthesis, Prosody, High-Level Syntactical Analysis

1. INTRODUCTION

Research on speech synthesis has brought significant improvements over the past decade that makes possible to generate natural speech from text [1, 2]. However, if the synthesized speech sounds *acoustically* natural, it is often considered poor according to the *way of speaking* (prosodic artefacts and monotony). Now, modeling the variability in the way of speaking (variations of prosodic parameters) is required to provide natural expressive speech in many applications of high-quality speech synthesis such as multi-media (avatar, video game, story telling) and artistic (cinema, theater, music) applications. Despite growing attention to prosody modeling over the past few years, one of the major drawbacks of actual prosody models remains the monotony of the generated prosodic parameters. The prosody monotony is both related to *poor dynamics* ("averaging problem") and *poor variability* of the generated prosodic parameters. Poor variability basically results in the generation of stereotypical prosody and is mainly due to the lack of linguistic knowledge extracted from the text .

In order to model the variability of prosodic parameters, it is necessary to determine an appropriate representation of prosodic parameters and then to extract and estimate the effects of high-level linguistic features (syntactic, semantic, discursive) on the observed prosodic parameters. At the signal level many approaches have been proposed to represent the variations of the prosodic parameters ([1, 3] for fundamental frequency variations and [4, 5, 6] for durations) based on statistical parametric modeling. At the symbolic level, prosodic variations are affected in many ways by linguistic constraints (phonologic, syntactic, semantic, discursive,...) occurring on a set of more or less linguistically well-defined units. Strictly speaking of syntactic-prosodic relationships, syntactic structure does not affect prosodic structure [7, 8] only, but acoustic variations as well, as shown in recent studies (for instance: *oral parenthesis* [9, 10]). This paper presents an exploratory study that aims to incorporate high-level syntactical features extracted from text within a HMM-based speech synthesis framework in order to improve prosodic model variability.

The paper is organized as follows: section 2 presents the linguistic processing chain and the syntactical features extracted from text; section 3 introduces the HMM-based synthesis framework; thus perceptual experiment and results are presented in sections 4 and 5.

2. HIGH-LEVEL SYNTACTICAL ANALYSIS

2.1. Linguistic Processing Chain

At the symbolic level, the input text (sentence, set of sentences or raw text) is processed by an automatic linguistic parser (*Alpage Linguistic Processing Chain* [1]) in order to extract high-level linguistic features (morpho-syntactical and syntactical) within the sentence level.

The *Alpage Linguistic Processing Chain* is a full linguistic processing chain for French based on a sequence of processing modules: a *lexer* module, a *parse* module, and a post-processing module. The input text is first preprocessed by the lexer module that output Direct Acyclic Graphs (DAGs) combined with lexical information retrieved from a morphological and syntactic lexicon. Then deep parsing is performed by the parser (FRMG), a symbolic parser based on a compact *Tree Adjoining Grammar* (TAG) for French that is automatically generated from a meta-grammar. The parsing result is then enriched by a series of post-processing modules whose role is to organize all the informations retrieved along those steps.

The output of the parser is a shared derivation forest that represents all derivation structures that the grammar can build for the input sentence, and indicates which TAG operation (substitution, adjunction,

[1]http://alpage.inria.fr/alpc.en.html

anchoring) took place on a given node of a given tree for a given chunk. This forest is then transformed into a shared dependency forest: anchors of trees related to a given node label are put into a dependency relationship with this label. A dependency forest is represented into a *DEP XML* format that incorporates the following items:

- **clusters** that are associated to the forms of the sentence;
- **nodes** that point on a given cluster and associated to a lemma, a syntactical category and a set of derivations;
- **edges** that connect a source node with a target node with a given label. More precisely, a given edge is associated with a set of *derivations* related to this edge and the related source and target chunk *operations*.

At last the forest is disambiguated by a heuristic-based module that outputs a unique dependency tree. An example of a disambiguated dependency graph as provided by the parser is shown in figure 1.

2.2. Syntactical Features Extraction

From the output of the linguistic process described in 2.1, a set of more or less high-level linguistic features were extracted.

morpho-syntactical form features are extracted from the surface processing.

- form segment;
- form *lexical category* and *class* (function vs. content form);

form dependencies are extracted from the deep parse processing. This set basically encodes the relationship between forms.

- {governor, current, governee} form *lexical category* and *class*;
- *edge type and label* between current form and {governor, governee} form;
- *signed dependency distance* between current form and {governor, governee} forms (in forms and in chunks);

recursive chunk are retrieved in a top-down process according to the operations and associated derivations. As a chunk of a given level can have several derivations (i.e governee chunks), we chose to stack governee chunk of a given level from left to right in order to provide a binary tree chunk representation. For our example sentence (cf. fig. 1), the transformed recursive chunks are:

(*S* (*AdvP* Longtemps) ((*VP* je me suis couché) (*NP* de bonne heure)))

Recursive chunks were finally transformed into non-recursive chunks by extracting only leaves of the transformed chunk tree (and encoding only partially recursivity through chunk depth estimation).

- {governor, current, governee} *chunk category*;
- *edge type and label* between current chunk and {governor, governee} chunks;
- *signed dependency distance* between current chunk and {governor, governee} chunks (in forms and in chunks);
- *chunk depth*;

adjunction represents a specific type of syntactical phenomena. In particular, adjunctions could concern different text spans (from a single form to a full sentence). Adjunction covers a large amount of syntactical phenomena such as incises, parentheses, subordinate and coordinate clauses, enumerations, ...

In the FRMG parser formalism, adjunctions can be easily extracted according to some specific pattern matching (including *modified*, *introducer*, and *modifier* nodes). Full adjunction is thus extracted by retrieving the full dependency descendance from the introducer. Modifier, introducer and modified category as well as edges type and label are then used to identify adjunction's type (such as relative clauses, adjective clauses, adverb clauses, ...).

- {modified, introducer, modifier} form *lexical category* and *class*;
- *adjunction type* retrieved from an adjunction dictionary;
- *edge type and label* between modified and introducer nodes and between introducer and modifier nodes;
- *signed dependency distance* between the adjunction's introducer and the modified node (in forms and in chunks);

As adjunctions have recursive property (a given adjunction can be embedded within another adjunction, or within an adjunction of the same type), features were extracted separately for each known adjunction type (with related pattern). Then, in case of recursivity of a given adjunction's type, adjunction with the larger span was extracted only.

3. INTEGRATION IN THE HMM-BASED SPEECH SYNTHESIS

Over the last decade, HMM-based speech synthesis system has grown in popularity [1, 2, 11]. This system models spectrum, excitation, and durations in a unified HMM framework. Compared to unit-selection systems it offers the ability to model different voices, speaking styles or emotional speech without requiring the recording of large databases. The implementation of our HMM-based speech synthesis system is based on the *HTS Toolkit*.

During the training, both spectrum and excitation parameters vectors (including f_0) and their dynamic vectors are extracted from the speech corpus and used to train the *context-dependent HMMs* according to the alignment. In this way, a mapping is performed between speech acoustics and linguistic features. Due to the large amount of context-dependent models, the models are hierarchically clustered into acoustically similar models using *decision trees* estimated according to *M*aximum-Likelihood Minimum Description Length criterion (ML-MDL). *Multi-space probability distributions* (MSD) are used to model variable dimensional parameter sequence such as log f_0 with unvoiced regions properly. Each context-dependent HMM has *state duration probability density functions* (PDFs) to capture the temporal structure of speech [4, 5].

During the synthesis, the text to be synthesized is first converted to a context-dependent label sequence. An utterance HMM is then built by concatenating the most appropriate context-dependent HMMs according to the label sequence and the trained decision trees. State durations of the utterance HMM are then determined based on the state duration PDFs. Finally, the speech parameter are generated using a maximum likelihood algorithm including *dynamic features*[13] from which a speech waveform is synthesized using a speech synthesis filter.

It is important to note that by adopting this HMM framework, all prosodic dimensions could be estimated according to the linguistic contextual features (i.e. not only f_0 variations and duration but voice quality as well). The following linguistic feature sets were introduced as contextual features:

- **baseline** linguistic units: phonem, syllable and prosodic group, with phonem and syllable phonological features (phonem phonological features, syllabic structure, and prosodic structure: prosodic frontiers and syllable prominence) [6];

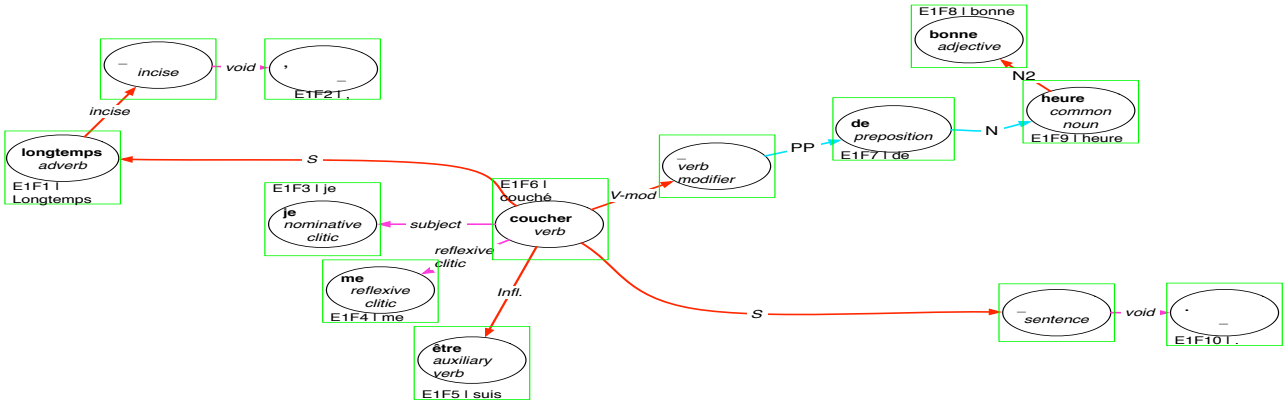

Fig. 1. Disambiguated dependency graph for the sentence: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early*"). Squares represent *clusters* (with associated form), circles represent *nodes* (with associated lemma and lexical category), arrows represents *edges* (with associated dependency label) going from source node (*governor*) to target node (*governee*).

- **morpho-syntactical** linguistic unit: form;
- **dependency** linguistic unit: form;
- **chunk** linguistic unit: chunk;
- **adjunction** linguistic unit: adjunction;

For each feature set, low-level linguistic features were extracted on each linguistic unit with left-to-right contexts (second order for the phoneme features, and first order for the other linguistic units with the exception of adjunction that does not necessarily have direct adjacent context): *locational* (position of a given unit within higher level units) and *weight* (number of observations of a given linguistic unit within higher level units) features.

4. EXPERIMENT

4.1. Experimental setup

For the purpose of that experiment, two models were compared: a *baseline model* including baseline and morpho-syntactical linguistic features; and a *rich linguistic model* including all linguistic features. Such linguistic sets were used to train all speech synthesis parameters (spectral and f_0 variations and durations).
Models were trained on a 1 hour (956 sentences) *laboratory* corpus spoken by a French non professional male speaker and recorded at 16kHz in an anechoic room.
The following processing chain was applied to the training corpus: phonemic segmentation using *ircamAlign* [12]; syllabation on interpausal groups; automatic syllable prominence detection using *ircamProm* [14]. Linguistic feature were extracted from text with the *Alpage Linguistic Processing Chain*. All analysis were conducted within the *ircamCorpusTools* [15] framework.
Due to the high complexity of the linguistic structure as well as their dependencies with speech prosody, estimating precisely the influence of each of the extracted linguistic features on the synthesized speech and prosody is unreachable. It is more likely to reduce the evaluation to a subset of well-defined specific syntactical phenomena. For that reason the evaluation corpus was designed according to syntactical adjunctions only. This choice is motivated by several reasons: adjunctions concerns long-term speech scope (oral parenthesis) which makes it easier to evaluate variations across different models; adjunctions could be described with limited vocabulary which makes easier to generate a limited sentence corpus which is needed in a subjective evaluation experiment; adjunctions has been shown in a previous study to be the most relevant syntactical feature for prosody modeling [16].
The text corpus used for the evaluation has been designed in the following maneer: 10 baseline sentences were chosen without any specific syntactic property (ex: *Le chat a mangé la souris*, *The cat ate the mouse*). These sentences were then enriched according to a set of adjunction type (subordinate participial and relative clauses, coordinate clauses, incises and enumerations). For each adjunction type, original sentences were enriched according to two control parameters: *position* (initial, medial, final) and *complexity* (presence or not of adjunctions within the current adjunction) of the introduced adjunction. This finally lead to a 54 sentences evaluation corpus. For the feasibility of the study, only 20 were selected for the subjective evaluation. Finally speech parameters were synthesized for each sentence according to the considered model and contextual features sequence extracted from text (prosodic structure being inferred from a common model then equal for both compared models).

4.2. Subjective evaluation

The evaluation consists of a subjective comparison between the 2 models. A comparison category rating (*CCR*) test was used to compare the quality of the synthetic speech generated by *baseline* and *rich linguistic* models. The evaluation were conducted according to *source-crowding* using social networks [2]. 50 French native speakers (including 17 experts and 33 naive listeners) compared a total of 20 speech sample pairs. They were asked to attribute a preference score according to the quality of each of the sample pairs on the comparison mean opinion score (CMOS) scale. Listening test was performed with headphones.

5. RESULTS & DISCUSSION

Analysis of variance (*ANOVA*) has been estimated on the resulting preference distributions in order to asses perceptive differences be-

[2]this perceptual evaluation group is available on `http://www.facebook.com/group.php?gid=150354679034&ref=ts` and the test on `http://recherche.ircam.fr/equipes/analyse-synthese/lanchant/index.php/Main/TestSP`.

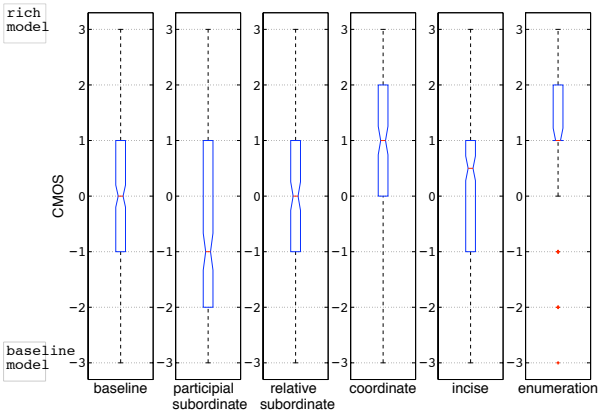

Fig. 2. CMOS preference score for the different adjunction types

tween the two models. An overall comparison shows a slight but significant preference for the rich linguistic model ($F(1, 1999) = 31$, $p - value < 10^{-5}$). A closer look to the results (fig.2) reveals noticeable differences among the different sentence sets. There is a slight but significant difference ($F(1, 499) = 8.4$, $p - value = 4.10^{-3}$) for the baseline sentences. This is actually in agreement to what could be expected since such sentences were defined without any particular syntactical property. Scores obtained for the enriched sentences reveals the strongest preference to the rich linguistic model, this result is not observed for all types of adjunction. If there is a clear preference to the rich linguistic model for the last three adjunction types (coordinate: $F(1, 299) = 50$, $p - value < 10^{-5}$; incise: $F(1, 399) = 38$, $p - value < 10^{-5}$; enumeration: $F(1, 99) = 78.4$, $p - value < 10^{-5}$), there is however a preference to the baseline model in case of subordinate sentences (clear preference in case of participial subordinate: $F(1, 399) = 17.58$, $p - value < 10^{-5}$ and no preference in the case of relative subordinate ($F(1, 299) = 1.5$ with $p - value = 0.2$).
The prosodic improvements of the synthesized speech can be listed as follows:

- accentuation: more prosodic variations observed on the accented syllables; accent form modification (conclusive, continuative accent);
- prosodic phrasing: local improvement of the prosodic phrasing (pitch variations and local speech rate).

These are encouraging preliminary results even if such improvements are not systematic: this is probably due to sparse observations (number of occurrences for each extracted syntactical phenomenon) for which parameters of the model could not be robustly estimated. Further experiments with more training observations will be carry out.

6. CONCLUSION & FURTHER WORK

We have presented a speech synthesis system using high-level linguistic features in order to improve prosody modeling. Perceptual evaluation shows that the proposed approach improve speech synthesis compared to the state-of-the-art approach. However this improvement still depends on the observed linguistic phenomenon which is probably due to sparse observations. In further work such approach will be evaluated with more linguistic observations and according to the linguistic complexity of an observed corpus. A joint model approach will be used in order to improve the estimation of the dependencies between high-level linguistic features and prosodic variations by jointly modeling prosodic variations over different time scales. Furthermore, one needs to define evaluation measurement that could more precisely account for prosodic variability.

7. ACKNOWLEDGMENTS

This study was partially supported by:

- ANR Rhapsodie 07 Corp-030-01; reference prosody corpus of spoken French; French National Agency of research (ANR); 2008-2012.;
- Programmes Exploratoires Pluridisciplinaires (PEPS), CNRS/ST2I, 2008-2010;
- Swiss National Science Foundation, grants n° PBNEP1-127788 and n°100012-113726/1, Neuchâtel University.

8. REFERENCES

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," in *Proc. of Eurospeech*, Paris, France, 1999, pp. 2347–2350.

[2] A.W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *Proc. ICASSP*, 2007, pp. 1229–1232.

[3] J. Latorre and M. Akamine, "Multilevel parametric-base f0 model for speech synthesis," in *Interspeech*, Brisbane, Australia, 2008.

[4] Heiga Zen, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura, "Hidden semi-markov model based speech synthesis," in *Proc. ICSLP*, 2004, pp. 1397–1400.

[5] B. Gao, Y. Qian, Z. Wu, and F. Soong, "Duration refinement by jointly optimizing state and longer unit likelihood," in *Proc. of Interspeech*, Brisbane, Australia, 2008.

[6] N. Obin, X. Rodet, and A. Lacheret-Dujour, "A multi-level context-dependent prosodic model applied to durational modeling," in *Proc. of Interspeech*, Brighton, U.K., 2009.

[7] A. Black and P. Taylor, "Assigning intonation elements and prosodic phrasing for english speech synthesis from high level linguistic input," in *Proc. of ICSLP*, 1994, pp. 715–718.

[8] V.K. Rangarajan Sridhar, S. Bangalore, and S.S. Narayanan, "Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 4, pp. 797–811, 2008.

[9] N. Dehé, "Clausal parentheticals, intonational phrasing, and prosodic theory," *Journal of Linguistics*, vol. 45, no. 3, pp. 569–615, 2009.

[10] F. Gachet and M. Avanzi, "Les parenthèses en français : Etude prosodique," *Verbum*, 2009.

[11] P. Lanchantin, G. Degottex, and X.Rodet, "A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method," in *submitted to ICASSP 2010*.

[12] P. Lanchantin, A.C. Morris, X. Rodet, and C. Veaux, "Automatic phoneme segmentation with relaxed textual constraints," in *Proc. of ELREC*, Marrakech, Morroco, 2008.

[13] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and S. Imai, "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," in *Proc Eurospeech'95*, 1995.

[14] N. Obin, X. Rodet, and A. Lacheret-Dujour, "A syllable-based prominence model based on discriminant analysis and context-dependency," in *Proc. of SPECOM*, St-Petersburg, Russia, 2009.

[15] C. Veaux, B. Beller, D. Schwarz, and X. Rodet, "Ircamcorpustools: an extensible plateform for speech corpora exploitation," in *Proc. of ELREC*, Marrakech, Morroco, 2008.

[16] N. Obin, X. Rodet, and A. Lacheret-Dujour, "Using high-level syntactical features to improve prosodic modeling," in *Submitted to ICASSP*, Dallas, U.S.A, 2010.