

A Metrical Model of Prosody for French TTS

Alex Monaghan & Fred Sannier

Aculab plc
Milton Keynes, MK1 1PT, UK
Alex.Monaghan@aculab.com

Abstract

The model of prosody used for French TTS in the Aculab TTS system is unusual in several respects. Firstly, it is based firmly on current metrical theories of French prosody. Secondly, it is entirely knowledge-based: there are no stochastic components in the model. Thirdly, it makes use of a pseudo-random element to avoid the predictability of synthetic prosody. Fourthly, it is designed to facilitate adaptation to other languages, particularly languages with similar prosody such as Spanish and Italian. The design and implementation of this model are presented here, as well as our plans for its extension to other languages.

1. Introduction

Aculab's multilingual TTS system has been developed over the past five years to provide unparalleled accuracy, quality and efficiency for Computer Telephony (CT) applications. The current version supports six languages (French, UK and US English, German, Dutch, and Latin American Spanish), with multiple voices configurable for each language. Further languages and voices are under development. Telephone bandwidth versions (8kHz sampling rate) of all voices are available.

Aculab TTS is a concatenative synthesis system which uses a deliberately small database of natural speech to perform mixed-unit waveform concatenation. Any necessary modification of pitch and duration is performed by time-domain techniques. Aculab's approach to multilingual TTS is based on a multilingual system architecture, detailed linguistic analysis and sophisticated symbolic processing, and the careful design of speech databases: these are discussed in some detail elsewhere [1].

Aculab TTS is specifically designed to run on Aculab's award-winning CT hardware, providing over 100 channels of telephone bandwidth TTS on a single DSP card. The only restriction on the number of channels is the number of cards which can be accessed by the application. At the time of writing, Aculab TTS is available free of charge to users of Aculab's DSP hardware. For more information about these products, see Aculab's website at <http://www.aculab.com>.

The present article concentrates on the design and implementation of prosody modules for French TTS. Our approach to French prosody is based on two pillars of previous work: the theory and implementations of intonation for TTS systems developed by Monaghan and his colleagues [2-5], and the model of French prosody proposed by Di Cristo and colleagues [6-8]. We have also drawn on results from Zellner [9], Santi [10] and Guaitella [11].

We will first set out the theoretical background to our work, and then present in detail the design and implementation of our French prosody modules.

2. A Metrical Model of French

2.1. Intonation in TTS systems

The theory of intonation for TTS described in [2] recognises three levels of intonational domain (TU, IU and IP), three degrees of pitch prominence (primary, secondary and tertiary), and three basic tone types (high, mid and low).

The lowest level of intonational domain, the tone unit or TU, corresponds roughly to a syntactic phrase and is characterised by a final primary accent but no boundary tone or pausing. The next level up, the intonation unit or IU, contains one or more TUs and is characterised by a boundary tone. This domain generally contains the text between two punctuation marks, and in such cases there is a corresponding pause at the end of the IU. The highest level of domain, the intonational phrase or IP, corresponds to a sentence-level chunk of text and is obligatorily associated with a domain-final pause. An IP contains one or more IUs.

In neutral declarative intonation, a primary accent is generally realised by a leading mid tone, an accent-lending (starred) high tone, and a trailing low tone: this corresponds to a nuclear accent in the British tradition [12]. A secondary accent is similar but without the trailing tone: this is the normal non-nuclear accent type. Tertiary accents are identical to secondary accents, except that they have the effect of removing the leading tone from the accent which follows them: this results in tone-linking, and occurs where two accents are placed in close structural or linear proximity.

Accents are generally assigned to the stressed syllables of content words, with the final accent in an IU being primary and the others secondary. The application of a rhythm rule [2,3] leads to the deletion of some secondary accents. In cases where the rhythm rule would cause deletion of the first accent in a particular domain, instead of being deleted this accent is demoted to a tertiary accent. There are thus five levels of prominence at the syllable level which are available to drive the assignment of segmental durations: primary accented syllables, secondary accented syllables, tertiary accented syllables, unaccented stressed syllables (where the accent has been deleted by the rhythm rule), and unstressed syllables. We assume, following [2,3], that lexical stress corresponds to the accentability of a syllable, and that unaccentable function words therefore do not have stressed syllables.

Note that all the computation so far, as far as the association of tones with syllables on the basis of the five levels of emphasis just described, has been done at an abstract, symbolic level. The translation of these symbols into absolute values in the time and frequency domains is postponed for as long as possible, avoiding any dependence on the characteristics of individual speakers or utterances.

2.2. The Di Cristo model

The Di Cristo model of French prosody contains many of the principles and ideas just described, and also adds many details specific to French. We will describe its main points here.

The two main principles for prominence assignment in this model are the *Accentual Bipolarisation Principle* (BPP) and the *Accentual Hierarchisation Principle* (AHP). The BPP essentially states that at each level of prosodic structure - word, phrase, utterance, etc. - there is a tendency for the first and last items to be relatively prominent. The AHP states that at each level of structure the rightmost prominent item will receive extra prominence. These two principles together allow a metrical tree or grid to be constructed, wherein the first and last items of any unit are metrically strong and the rightmost strong item at any level is the strongest (designated terminal element or DTE in metrical terminology, nucleus in British intonation terminology).

The application of the BPP in French assigns lexical stress to the first and last full syllables (non-schwa) in French content words, and assigns accents to the first and last stressed syllables in a TU. At the IU level, the BPP can be used to assign boundary tones to the last accent and extra pitch prominence to the first accent.

The AHP promotes the rightmost strong element at each level to the next level of prominence, allowing a metrical grid to be generated directly. It also assigns primary (nuclear) status to the rightmost accent in the appropriate domain: the choice of level of domain depends on various factors, including speech rate, speaking style, and affect.

The details of tone types, tonal alignment and scaling in the Di Cristo model are rather different from our own model. As our model applies to six languages, and there is considerable evidence that many of these details are language-independent [13], we have not adopted a specific approach for French.

We do, however, share many underlying assumptions with the Di Cristo model. Perhaps the most important is that, in the context of TTS systems, the details of syntactic structure are largely irrelevant to prosody. Di Cristo asserts that eurhythmy (a regular prosodic structure) is more important than isomorphy with syntax: he sees rhythm in French as the result of a sequence of accented and unaccented syllables, the correspondingly different lengths of those syllables, and the pauses and other boundaries in the sequence. It is therefore unnecessary to perform a full parse of the text or to derive any deeper a hierarchy than the prosodic hierarchy which will be realised by the synthesiser. The conclusion is that shallow parsing is sufficient for TTS systems.

On the other hand, Di Cristo shares the view expressed in [4] that the determinants of prosody are actually above the level of syntax, and may indeed be the same as the determinants of syntactic structure: semantic, pragmatic, and cognitive factors (expressed through devices such as focus, pronominalisation, and sequential order of constituents) are the ideal predictors of prosody, but unfortunately we are still a long way from being able to apply such levels of analysis in TTS systems. We are therefore obliged to assume a neutral information structure as the input to TTS conversion, and in this case the application of eurhythmic principles seems to be very effective.

2.3. Design criteria

Whereas linguistic theories of prosody can assume all the information which is available to human readers, speakers and hearers, a TTS system can only assume availability of a very limited amount of information as the input to prosodic processing. Aculab TTS is no exception. Its prosody modules are furnished with the following information from previous text processing:

- A sequence of orthographic words
- A syntactic part of speech for each word
- A sequence of syllables for each word
- A sequence of phonemes, including lexical stress information, for each word
- Boundary markers (e.g. punctuation) which can be deduced from the text

While most of this information is correct for any particular input, there are inevitably occasional errors in each type of information. This is another reason why it is preferable to rely on prosodic principles than the 80% or 90% accuracy of a syntactic analysis.

Bearing these unfortunate realities in mind, the design of our French prosody model was undertaken according to three main criteria:

- Robustness (the model should apply to all possible text input without producing gross unnaturalness)
- Intelligibility (the model should err on the side of slow, careful, over-emphasised speech rather than fast, casual, or reduced speech)
- Simplicity (the model should keep processing to a minimum, since its input is inherently unreliable and - like all contemporary TTS systems - lacks much of the information required for accurate prediction of prosody)

The basic algorithm which we developed respects all these criteria, and allows further development when more reliable input is available or when better heuristics are identified. The improvement of this algorithm is one of our ongoing research objectives. The algorithm is currently as follows:

1. Assign IU boundaries corresponding to boundary markers identified in the text, such as punctuation or white space
2. Assign TU boundaries at the edges of syntactic phrases, by default at the boundary between a content word and a following function word
3. Build a metrical grid, with secondary accent-level prominence for the first and last stressable syllable of every content word
4. Assign primary accent-level prominence to the last accent (if any) in each TU
5. Remove any TUs or IUs which do not contain a primary accent
6. Apply eurhythmic principles to equalise the lengths of all IUs, aiming for an optimal number of TUs in an IU: this involves demoting IU boundaries to TU boundaries, and promoting TU boundaries to IU boundaries
7. Apply eurhythmic principles to equalise the lengths of all TUs, aiming for an optimal number of syllables in a TU: this involves inserting and deleting TU boundaries
8. Apply the appropriate rhythm rule to each IU
9. Apply tonal assignment rules, duration rules, tonal alignment rules, and tone scaling rules (these rules are all described separately below)

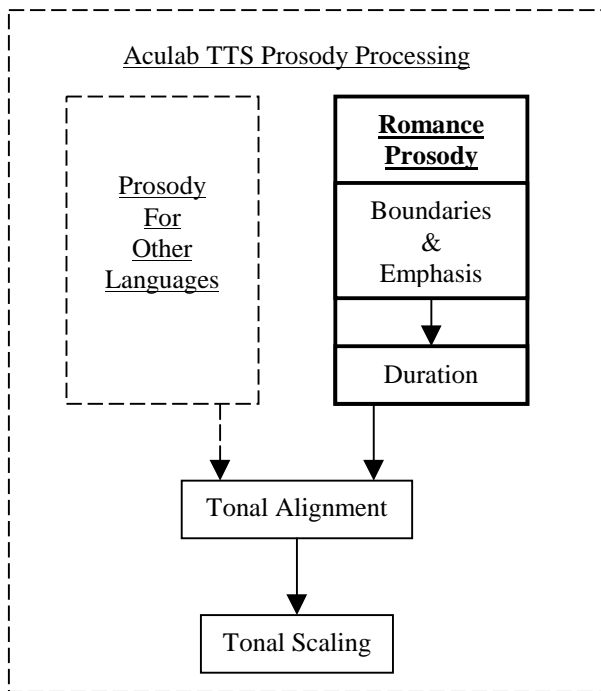


Figure 1. Prosody processing for Romance languages, including French, in Aculab TTS

3. Implementation

As mentioned above, our French prosody model is only part of Aculab's multilingual TTS system. It thus had to be tailored to the existing architecture, and to the principles mentioned in [1] such as the minimal use of language-specific components. There are five main stages in Aculab's prosody generation:

- Assignment of major and minor prosodic domains;
- Assignment of emphasis within each domain;
- Calculation of segmental durations;
- Alignment of pitch targets with phonetic segments;
- Calculation of frequency values for pitch targets.

The first two stages have been outlined above, and their implementation will be discussed here. The third stage has hardly been mentioned so far, but will be described in detail in this section. The last two stages are fully language-independent, and will be outlined below.

Since the release of the first version of Aculab's French TTS early in 2001, the implementation has been generalised to include other Romance languages supported by Aculab TTS. The set of Romance languages available to users of Aculab TTS (currently French and Latin American Spanish) will be enlarged to include Italian and Brazilian Portuguese in the near future. The general structure of prosody processing for Romance languages in Aculab TTS is shown in Figure 1.

The remainder of this section will first describe the implementation of the algorithm given above. Next it will describe our approach to segmental durations for French, and give some details of the implementation of our duration rules. Finally, it will outline the process of aligning and scaling pitch targets.

3.1. Boundary and emphasis assignment

When support for French was added to Aculab TTS, we decided to implement a grid-only metrical approach, i.e. to omit the construction of the metrical tree which we use for Germanic languages. This was partly due to the computational expense of tree construction and manipulation, and partly due to our belief that the computation of emphasis and boundaries for TTS in Romance languages is rather simpler than for the Germanic languages which had so far been supported by Aculab TTS. This is not meant to suggest that we believe French prosody to be less complex than Germanic prosody, but merely that French neutral prosody is more easily predicted to the level of accuracy which one can reasonably expect from a TTS system for unrestricted text. Romance languages tend not to indulge in pragmatic deaccenting to the same extent as Germanic languages, and the location of the nuclear accent or DTE is much less variable, resulting in a more regular assignment of prosodic emphases and boundaries.

Our implementation of the assignment of boundaries and emphases for French is a multi-stage process, with each stage corresponding to a linear pass through the input. The first pass assigns one level on the grid to the initial and final stressable syllables of content words: this applies the Accentual Bipolarisation Principle (BPP) to the word level. The very first such syllable in the input is marked specially, since in French this is guaranteed to be the first pitch accent. A second pass assigns tone unit (TU) boundaries between certain syntactic categories - basically between content words (CWs) and function words (FWs) - to produce an approximate phrase-level structure. Next, a higher grid level (corresponding to a pitch accent) is assigned to the final grid position before each of these CW/FW boundaries, and to the first grid position following each boundary, applying the BPP at the TU level. This produces an over-accented grid for each TU, as illustrated in Figure 2.

```

      x           x           x   x
      x x x   x   x           x   x
Le petit papillon jaune       Un chien blanc

```

Figure 2. Over-accented TU-level grid

The intonation unit (IU) level is handled in the next pass, which assigns IU boundaries at major breaks indicated in the text, such as punctuation marks. Note that some TU and IU boundaries will coincide, and some will not: however, if an IU boundary is assigned where there was no existing TU boundary, then that boundary is automatically also a TU boundary. A higher grid level is then assigned to the final accent before each of these IU boundaries, corresponding to a primary or nuclear accent: this is the DTE of the IU, and reflects the application of the Accentual Hierarchisation Principle (AHP) at the IU level. This produces an over-accented grid for each IU, as shown in Figure 3.

At this stage, the various eurhythmic principles are applied. First, the structure at the IU level is balanced so that all IUs are maximally similar in terms of length and prosodic prominence. If any IUs do not contain a nuclear accent, their boundaries are removed and their contents are merged with a neighbouring IU. We then ensure that all IUs in an utterance are of a similar length, by demoting IU boundaries to TU boundaries and promoting TU boundaries to IU boundaries

where necessary: this process explicitly puts rhythmic considerations before syntactic ones.

Once the IU structure has been optimised, the TU-level structure within each IU is similarly balanced. As above, units with no prosodic prominence are merged with neighbouring units and the remaining units are adjusted for length by adding or deleting TU boundaries.

The final stage of assigning boundaries and emphasis is to apply the rhythm rule [3] at the IU level. This involves another pass through the grid, demoting alternate accents leftwards from the rightmost accent (nucleus or DTE) in each IU. One important refinement of this simple rule is that the leftmost accent in an IU should not be demoted, since according to the BPP it is equally important to mark both ends of these domains with an accent. The output of this final stage is a rhythmic alternation of accents within each IU, as shown in Figure 4.

```

                                     x
      x  x    x  x    x          x  x  x
      x  x    x  x    x x x    x  x    x  x
Un chien blanc va manger le petit papillon jaune de maman

```

Figure 3. Over-accented IU-level grid

```

                                     x
      x  x          x          x          x
      x  x    x  x    x x x    x  x    x  x
Un chien blanc va manger le petit papillon jaune de maman

```

Figure 4. Rhythmic IU-level grid

At the end of the process of boundary and emphasis assignment, we have produced a grid value (ranging from 1 to 5) for each syllable, as well as an implicit three-level prosodic structure of utterance, IUs and TUs. The resultant degrees of emphasis and boundary strength form the input to the duration rules and the language-independent pitch target alignment and scaling functions.

3.2. Duration rules

In Aculab TTS, segmental duration in Romance languages is calculated entirely by rule, with no stochastic component whatsoever. This is a novel approach, combining previous work [14] with new ideas. The syllable-timed nature of French makes syllable durations a very important factor in this approach, but we believe that a similar approach can also be used for languages such as English and German which are not generally considered to be syllable-timed.

The grid provides a level of emphasis for each syllable, which translates into a target duration at the syllable level. A language-specific table of segmental durations provides an intrinsic duration for any syllable, and the mismatch between the intrinsic duration and the target duration results in modifications to the actual duration of individual segments in the particular syllable.

The rules which govern these modifications are applied in discrete steps, until an acceptable compromise is achieved between the intrinsic duration (the sum of the typical segment durations) and the target duration specified by the grid. The number of steps applied therefore depends on the degree of mismatch. The steps are as follows. First, for all segments

duration is modified by 10%. If this is not sufficient, all segment durations are modified once more by a further 5%. Finally, if even more modification is required, those segments which show the greatest elasticity in their durations are modified by another 10%.

These rules are currently quite simple, and will be refined as more languages are handled by the system. This approach produces highly acceptable segmental durations for French in our current system. We expect that with minor modifications the set of rules used for French can be generalised to other Romance languages: they have already been successfully applied to Latin American Spanish. Additionally, our novel combination of segmental and syllable durations with an iterative approximation technique produces a quasi-random variation in the segmental duration values which are assigned, and this enhances the perceived naturalness of the synthetic speech output.

3.3. Tonal assignment rules

The tonal assignment rules are currently part of the emphasis assignment routine, since they are language specific and entirely predictable from the prosodic structure. In agreement with the work by Di Cristo and Monaghan discussed in Section 2 above, we use the same basic tone types for both declarative and interrogative utterances. This means essentially that boundary tones are low and accent-leading (starred) tones are high.

A primary (nuclear, or DTE) accent is assigned three tones (mid leading tone, high accent-leading tone, low trailing tone). A secondary accent is assigned two tones (leading mid, accent-leading high).

Utterance-final boundary tones are low, but utterance-internal boundary tones are generally not. This produces the desired perception of continuation at utterance-internal boundaries, which contrast with the terminal utterance-final falling boundary.

There are details of downstep and upstep, range overshoot and undershoot, and various types of utterance-internal boundaries which are omitted here to avoid lengthy explanations and the need for a very specialised theoretical background. Most of these details are implemented in a similar way to the implementation discussed in [2], where such matters are explained more fully.

All the tones specified by our tonal assignment rules are associated with the appropriate syllables: no attention is paid to the structure or contents of syllables at this stage. The output of the tonal assignment rules is a sequence of syllables, some of which have one or more tones associated with them. Note once again that we still preserve a purely symbolic tonal representation: there is no mention of absolute frequency or timing values at this stage. The function of converting these abstract symbolic representations into absolute values in the time and frequency domains is left to the final stage of our prosody processing, the tonal alignment and scaling rules.

3.4. Aligning and scaling pitch targets

Pitch target alignment follows the work cited in [13], which appears to generalise across all the languages in question. The calculation of frequency values is based on the original model in [15] and refined in [2]. In both cases, we believe that Aculab TTS is unusual in applying sophisticated phonological models to a commercial speech synthesis system.

Pitch targets are aligned with the segmental content of the relevant syllable, using a set of rules which takes account of the structure of the syllable (open, closed, etc.) and the phonetic nature of the vowels and consonants involved. As a rough guide, accent-lending tones are aligned towards the end of the syllable nucleus, leading tones are aligned with the onset of voicing in the syllable, trailing tones are aligned with the start of the following syllable, and boundary tones are aligned with the ends of their domains. If this alignment will result in tones occurring too close together, or in important pitch movements occurring on unvoiced portions of the utterance, the tonal alignment is adjusted accordingly.

Fundamental frequency values are assigned to targets using a model very similar to that described in [2]. Some aspects of the model (e.g. the lowest and highest possible pitch values, and the normal width of the pitch range) are speaker-specific and others (e.g. the size of downstep, or the position of a tonal target within the pitch range) are speaker-independent. This allows us to synthesise a wide range of different voices simply by modifying a handful of parameters. As with the tonal alignment rules, this model is speaker-independent (given the right parameters) and language-independent. Figure 5 shows the major parameters of the model.

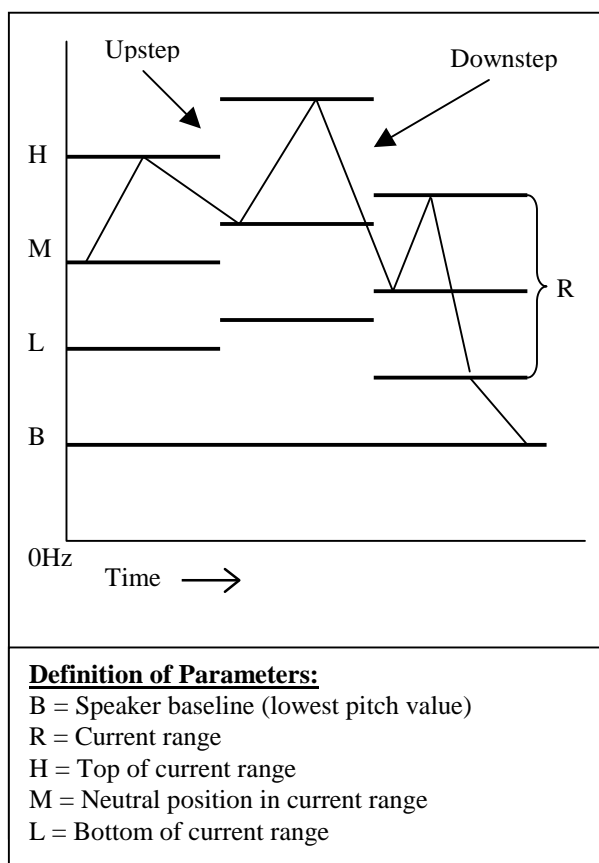


Figure 5. A parametric model of fundamental frequency for speech synthesis.

3.5. Summary

We have presented a metrical model of French prosody, based on current prosodic theory and making no use of stochastic techniques. This model has been implemented in the current version of Aculab TTS. Evaluations carried out by ourselves and our customers indicate that the output of the model is of a very high quality, sufficient for commercial applications: the details of these evaluations are unfortunately not available for publication.

The implementation of this model has followed the principles of placing more importance on prosodic structure than on syntactic structure, and of preserving an abstract symbolic representation for as long as possible. It has also incorporated several novel techniques into Aculab TTS: grid-only calculation of prosodic structure, grid-based duration rules, a hybrid model of syllabic and segmental duration, and a pseudo-random duration component.

Our implementation of French prosody is robust and simple, and produces highly intelligible output. It maintains Aculab's reputation for fast and reliable TTS of a high quality for CT applications.

4. Conclusions

Aculab TTS is a multilingual system which provides a framework for the development of any language. We have developed modules for French prosody within this framework, combining the approaches of Monaghan and Di Cristo with novel techniques. We believe that Aculab's application of linguistic approaches to the generation of prosody, particularly our implementation of grid-based metrical models, is unique among commercial TTS systems. In our model of French prosody, we have combined the advantages of a symbolic metrical representation with a well-developed approach to the generation of prosody in TTS systems.

Our model does not rely on accurate or detailed syntactic analysis, and concentrates on those aspects of prosody which are most important in TTS systems: naturalness, intelligibility, and the right balance between smoothness and spontaneity. Our use of abstract speaker-independent representations allows us to synthesise a wide range of prosodic variants in a very efficient manner. We expect to apply our grid-only approach to the generation of prosody for the Germanic languages in the near future.

5. References

- [1] Monaghan, A., Kassaei, M., Luckin, M., Amador-Hernandez, M., Lowry, A., Faulkner, D., and Sannier, F., "Multilingual TTS for Computer Telephony: The Aculab Approach", submitted to Eurospeech 2001, Aalborg, Denmark.
- [2] Monaghan, A., *Intonation in a Text-to-Speech Conversion System*, PhD thesis, University of Edinburgh, 1991.
- [3] Monaghan, A., "Rhythm and Stress Shift in Speech Synthesis", *Computer Speech and Language* 4:71-78, 1990.
- [4] Monaghan, A., "What Determines Accentuation?", *Journal of Pragmatics* 19:559-584, 1993.

- [5] Monaghan, A., "Heuristic Strategies for Higher-Level Analysis of Unrestricted Text", in G. Bailly & C. Benoit (eds), *Talking Machines* 143-161, Elsevier, Amsterdam, 1992.
- [6] Di Cristo, A., "Le Cadre Accentuel du Français: Essai de Modélisation", *Langues* 2:184-205 and 258-269, 1999.
- [7] Di Cristo, A., "Intonation in French", in D. Hirst and A. Di Cristo (eds), *Intonation Systems* 195-218, Cambridge University Press, Cambridge, 1998.
- [8] Di Cristo, A., Di Cristo, P., Campione, E., and Véronis, J., "A Prosodic Model for Text-to-Speech Synthesis in French", in A. Botinis (ed), *Intonation: Analysis, Modelling and Technology* 321-355, Kluwer, Amsterdam, 2000..
- [9] Zellner, B., *Caractérisation et Prédiction du Débit de Parole en Français*, doctoral dissertation, Université de Lausanne, 1998.
- [10] Santi, S., *Synthèse Vocale de Sons du Français*, doctoral dissertation, Université de Provence.
- [11] Guaitella, I., *Rythme et Parole*, doctoral dissertation, Université de Provence.
- [12] Crystal, D., *Prosodic Systems and Intonation in English*, Cambridge University Press, Cambridge.
- [13] Arvaniti, A. and Ladd, D. R., "Tonal Alignment and the Representation of Accentual Targets", *Procs ICPhS* vol. 4:220-223, Stockholm, 1995.
- [14] Campbell, W. N., Isard, S. D., Monaghan, A., and Verhoeven, J., "Duration, Pitch and Diphones in the CSTR TTS System", *Procs ICSLP* vol. 2:825-828, Kobe, Japan, 1990.
- [15] Ladd, D. R., "A Phonological Model of Intonation for use in Speech Synthesis by Rule", *Procs European Conference on Speech Technology* vol. 2:21-24, 1987.