Synthesizing Elaborate Intonation Contours in Text-to-Speech for French

Piet Mertens

Centre for Computational Linguistics,
Leuven University (K.U.Leuven)
Piet.Mertens@arts.kuleuven.ac.be

Abstract

This paper presents a modular TTS system (called MINGUS) which exploits syntactic information contained in the input and allows additional annotation of the input in order to obtain particular intonation contours or to vary most prosodic parameters. This system is based on a tonal representation of French intonation, on a model of the interaction between syntax and prosody, and on a model of the semantic and pragmatic aspects of intonation.

1. Introduction

Whereas segmental quality of synthetic speech has improved over the last decades up to a level where synthetic speech can hardly be distinguished from natural speech, much progress can still be made at the level of prosody to obtain natural and varied intonation contours, to exploit their communicative functions, and to take into account discourse structure.

2. Global aspects of prosody

In modelling speech prosody, one can distinguish global and local properties. Among the global properties are the overall pitch range typical of a given speaker, the actual pitch range used in the utterance, the amount of declination, the rate of speech, rhythm variations, and so on. Although such properties are essential for simulating emotions or speaking styles, one makes abstraction of them when interpreting the linguistic functions of intonation (such as prosodic boundaries, prosodic organisation and focus). The underlying assumption is that a (structural) pitch pattern (configuration, contour) may be modulated by global parameters in order to express information carried by the pitch pattern and by global properties simultaneously.

3. A tonal model of French intonation

At the level of linguistic structure, a tonal model of prosody is adopted [2, 3, 4, 5, 7].

Pitch levels are defined on the basis of (1) the speaker's pitch range (L- = top and H+ = bottom), (2) major pitch intervals (L = low, H = high), and (3) minor pitch intervals, which raise or lower the L and H levels to /L, \L, /H or \H.

Two types of stress are distinguished: final and initial stress, labelled AF (*accent final*) and AI (*accent initial*) respectively. This distinction is based on contour distribution, pause distribution, and phonatory effort.

A tone (or intonation morpheme) is defined as the sequence of pitch levels at a particular syllabic location, where locations are defined relative to stress position.

Three prosodic domains are used: the intonation group, the stress group, and the intonation package.

1. The *intonation group* (IG, *groupe intonatif*) is defined on the basis of the presence of a stressed syllable of type AF. This final stress may be preceeded by a sequence (labelled NA) of one or more unstressed syllables, by a stressed syllable of type AI, and by an unstressed series preceeding the AI. It may also be followed by an *appendix*, i.e. a sequence of unstressed syllables with special distributional properties. The structural scheme of the intonation group may be represented as follows (brackets indicate optionality):

[[NA] AI] [NA] AF [NA]

(Adjacent unstressed syllables in an IG is treated as a single subunit in the structure because of the monotonic contour (either level, rising or falling) characterizing it. As a result, when the IG does not contain an AI, there is only one NA sequence to the left of the AF.)

Figure 1 illustrates how this general scheme may be realized in some intonation groups.

AF	oui HL-
NA AF	sûrement ll HL-
AI AF	assez H L-L-
AI NA AF	précisément H h.l L-L-
NA AI NA AF	dans ma situation l....l H ll HH
NA AI AF	pour ce projet l.....l H L-L-

Figure 1. Samples of intonation groups (right) illustrating location sequences (at left) derived by the general structure of the intonation group.

Figure 2 shows a representative (although not complete) set of contours for an intonation group consisting of the words "elle est sympathique".

For a given syntactic structure the number of intonation groups may vary. This can be accounted for on the basis of an underlying prosodic unit, the stress group.

2. In French a *stress group* (SG, *groupe accentuel*) consists of a word *w* carrying lexical stress (word stress) and of all adjacent clitic words (without lexical stress) that are governed by *w*. As a result the number of SGs is fixed for a given syntactic structure. When a SG is indeed stressed, the stress is located on its last full syllable (containing a vowel other than schwa) of the SG which coincides with the AF of the resulting IG. When it is unstressed, it will be integrated in the IG on its right.

Since an IG consists of one or more adjacent SG which are syntactically related and SGs are predicted by syntax, stress groups are used to determine possible intonation groups.

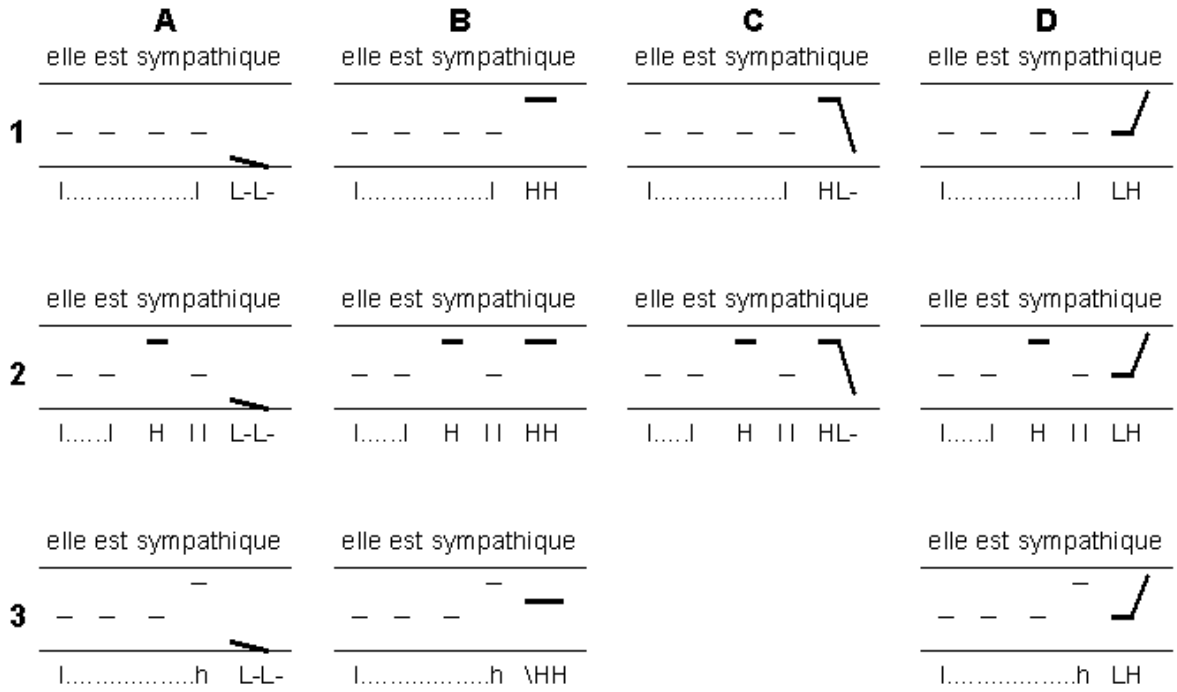

Figure 2. Schematic representation of some intonation contours generated by the tonal model for a string consisting of one intonation group. The tonal notation is shown below the contour. Row 2 adds an initial stress to the corresponding contour of row 1. Row 3 adds a high tone at the penultimate syllable.

The presence of lexical stress can be determined on the basis of the part-of-speech, at least in most cases [4, 7].

3. A boundary level is associated with each tone of type AF. A sequence of intonation groups therefore implies a sequence of prosodic boundaries resulting in a recursive and hierarchical organisation of the intonation groups into *intonation packages* (*paquet intonatif*). The largest package is that which carries a maximal boundary.

An initial stress (AI, sometimes called "emphatic stress") may be added to a word (or morpheme) of the intonation group [2]; see row 2 of figure 2. It will be located at the first syllable of the word if its onset contains a consonant, and at the first or second syllable otherwise (e.g. "exactement", "absolument"). Initial stress typically is accompanied by increased vocal effort and lengthening of the syllabic onset, and often by a pause preceeding the stressed word.

A (compositional) semantic interpretation of the tones has been proposed [5], taking into account the domain which is affected by the tone. As a result, the semantic characterisation is closely related to the tonal model.

4. Synthesis of prosody

4.1 Rationale and system architecture.

Apart from the actual synthesizer, a TTS system generally consists of two major blocks [1]. A first stage performs linguistic processing to provide morphological and syntactic information required for later steps. This also includes the computation of the phonetic transcription (grapheme-to-phoneme conversion). For a detailed description of these modules, see [1] and [7]. In the second stage, called prosody generation, pitch contours are computed.

The approach to prosody generation taken in the MINGUS system is unconventional because of the extensive use of syntactic information and because two levels of prosodic representation are used: a tonal representation in which tones (symbols) are associated with particular syllables, and an acoustic-phonetic representation in which pitch targets are assigned to particular positions in those syllables. The first representation is computed by the intonation generation module and the second by the pitch and duration models. Finally, an annotation system (using XML tags) enables one to select a particular tonal sequence, to adjust global parameters, to assign prosodic functions, and so on. As a result prosody generation itself consists of several steps.

4.2 Intonation generation.

Prior to intonation generation, particular syntactic constructions (cleft, left and right dislocation, parenthesis) affecting prosodic boundaries and tones are identified [6, 7]. Punctuation is used to derive pragmatic tags (e.g. assertion, question, focus), which are linked to substrings of the input.

The syntactic tree is reorganized in such a way that, at some level in the tree, nodes correspond to intonation groups. First, stress groups are obtained on the basis of morphological (part-of-speech) information and syntactic dependency. Syllable count as well as syntactic dependency are used to combine stress groups into intonation groups. Speech rate and utterance length may be used as additional criteria. Each intonation group receives a prosodic boundary level depending upon the syntactic relations encoded in the (modified) syntactic tree. Prosodic tags may be explicit (added to the input text by the user) or inferred by the system on the basis of other information. The combination of all available

information leads to the selection of tones to be associated with syllabic positions in the IG, either final stress (AF), the penultimate syllable, the syllable carrying initial stress (AI), or the appendix. Only the AF tones are necessary; other positions have default (unmarked) values. The resulting tonal representation is speaker-independent: it merely consists of tones (pitch levels and stress type) associated with syllables, without any reference to absolute fundamental frequency.

4.3 Duration model.

The duration model determines the duration of each sound, taking into account sound identity, syllable boundaries, position in the intonation group, tone identity, and speech rate [7]. Syllable duration is calculated first. It results from several factors: (1) the location of the syllable in the IG (200 ms for final stress, 152 ms for the penultimate syllable, 131 ms otherwise); (2) in the case of final stress, the tone carried by the syllable also plays a role (the initial duration is multiplied by a factor between 1.0 and 2.4, which is specified for each tone); (3) rate of speech. In a next step sound duration is calculated using a z-score model such that total syllable duration approaches the estimated syllable duration.

4.4 Contour generation.

The contour model provides the pitch targets (i.e. F_0 values at time instants) associated with a sound. In the current model, pitch targets are generated for vowels only. For each tone there is a list of tone targets, where a tone target specifies the pitch level that is reached at the specified temporal position in the sound. Tone targets are specified relative to total vowel duration; their number varies according to the shape of the contour. For tone 'HL-' there are three (table 1): the high level is reached after 33% of the vowel and maintained at that level up till 50% of the duration of the vowel, then the pitch reaches the bottom level ('L-') at 100% of the duration.

Table 1. Examples of pitch levels targets for a tone.

Tone	Targets
HL-	(33, H), (50, H), (100, L-)
HH	(50, H), (80, H)

To convert pitch levels into acoustic F_0 values, a *pitch range model* is used involving 6 parameters that define a configuration of lines corresponding to the pitch levels in the symbolic representation, and which we call the *grid.*

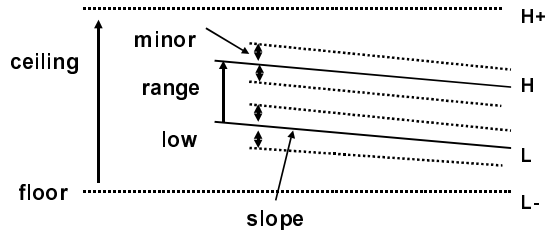

Figure 3. Grid system used in the contour model. See text for a definition of the parameters.

Figure 3 illustrates these parameters. *Floor* and *ceiling* correspond to the lower and upper limits of the speaker's pitch range. The *low* parameter corresponds to the frequency of the low level at the beginning of the utterance. The *range* parameter sets the melodic interval between the low and high pitch levels. The *slope* parameter can be used to simulate declination or inclination. Finally the parameter *minor* sets the melodic interval between the primary (low or high) level and its lowered or raised counterpart. The parameters *floor* and *low* are specified as absolute values in Hz; all others are specified relatively in semitones. In addition there is a command to shift the central part (all parameters except floor, ceiling and slope) of the grid up or down, which is used to simulate register changes.

The pitch range model is in agreement with [9] who distinguish two aspects of pitch range: pitch *span* and *level*, corresponding to the *range* and *floor* parameters in our model.

The above model is used to compute F_0 values starting from the pitch levels specified in the tone targets. For each target, one computes its position on the time axis (by adding total elapsed time and the portion of sound duration specified in the pitch target) and obtains the F_0 value at the intersection of the line corresponding to the pitch level specified in the pitch target. This results in pitch targets expressed as frequency values. The MBROLA synthesizer, which is used in the MINGUS system, linearly interpolates F_0 values between the pitch targets. As a result the actual pitch contour will consist of a sequence of straight lines. This is illustrated in figure 4, where thin lines indicate pitch levels as defined above, and thick lines indicate the resulting contour.

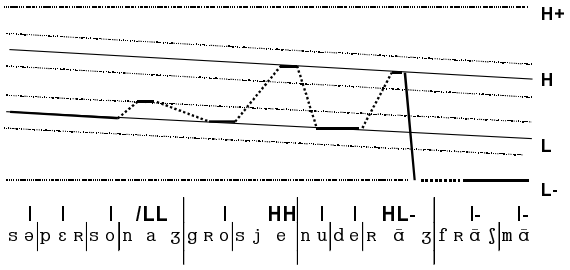

Figure 4. Pitch contour resulting from the transformation of tone targets in the contour model.

It will be clear that the duration model plays an important role for the actual shape of syllabic contours, since it affects the timing of the pitch targets. However al this is hidden in the specification of the targets.

5. Tags for synthesis of prosody

The MINGUS system uses available syntactic information and punctuation to select tones and to obtain a grammatical intonation contour. This kind of intonation, sometimes referred to as "neutral intonation", is quite predictable and gets boring very soon. Indeed, syntax alone rarely tells us what parts of the utterance the speaker wants to focus or put in the background, what kind of emotion he wants to convey, and so on. These elements require additional information. In the future some of this information could be obtained from discourse (or dialogue) representation.

To test hypotheses about pragmatic aspects of prosody, an annotation scheme [8] (in XML format) was defined enabling the system to generate a variety of pitch contours, while keeping in line with syntactic structure.

The prosodic tags are divided into 3 classes (see table 2). (1) Acoustic tags are used to modify global prosodic aspects (such as the grid parameters, speech rate, voice) and to insert pauses or sounds. (2) Tonal tags enable us to force a

Table 2. Tags used for synthesis of prosody.

ACOUSTIC TAGS

<**voice** db=*Name*/>	Select voice, i.e. diphone database.
<**grid** [range=*R*] [slope=*S*] [floor=*F*] [ceiling=*C*] [low=*L*]/> <grid reset/> <grid keep/>	Set grid parameters. (see fig. 2) Reset grid parameters to default values for active voice. Keep (don't reset) current grid.
<**register** normal/low/high> ...	Select register
<**rate** normal/low/high/>	Set speech rate
<**pause** [len=*N*]/>	Insert pause of length *N*
<**breathe**/>	Breathing

TONAL TAGS

<**tone** af=*Tone*> ... </tone> <tone ai=*Tone*> ... </tone> <tone penult=*Tone*> ... </tone>	Apply the indicated tone to the syllable with final stress. Apply the indicated tone to the syllable with initial stress Apply the indicated tone to the penultimate syllable of group.
<**boundary** terminal/internal />	Insert boundary of maximal/medium strength

FUNCTIONAL TAGS

<**focus**> ... </focus>	Focus, highlight, emphasize	use HL
<**topic**> ... </topic>	Topic (sentence-initial position)	use HH or H/H
<**tail** low/high> ... </tail>	Background info (sentence-final position)	use low/high appendix
<**e**> ... </e>	Emphatic stress	use AI
<**question**> ... </question>	Question intonation (rising)	use H/H
<**probe**> ... </probe>	Probe/Poll listener.	use ..h for penultimate
<**assert**> ... </assert>	Assertiveness effect	use h L-L-
<**invite**> ... </invite>	Invite listener	use LH
<**cite**> ... </cite>	"call contour", Citation effect	use h \HH
<**parenthesis** low/high> ... </parenthesis>	Parenthesis	use low/high register

particular tone in a particular position. They override the tones that would be generated by the system, but preserve tones at positions for which no tags are specified. (3) Functional tags are more abstract; they specify prosodic functions (such as focus). It is left to the system to decide what prosodic events are required to perform those functions. In the current implementation functional tags behave as meta-tags: they are translated into tonal or acoustic tags before prosody generation starts.

6. Conclusion

In this paper we described prosody generation in the MINGUS text-to-speech system. Various aspects were covered including stress group formation, intonation group formation, tone assignment, duration modelling, pitch contour generation, syntax-prosody interface, global prosodic features. In addition an annotation scheme is proposed which enables one to adjust or add prosodic features. This prosodic annotation scheme prepares the system for future developments that would integrate a discourse representation theory.

Text-to-speech provides an ideal test platform for validating theories about prosody. The system gives control over almost any aspect of prosody. Samples of synthetic speech generated by the MINGUS system are available at http://bach.arts.kuleuven.ac.be/pmertens/ .

References

[1] Dutoit, Th., 1997. *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic: Dordrecht.

[2] Mertens, P., 1987. *L'intonation du français. De la description linguistique à la reconnaissance automatique*. Ph.D., Université de Leuven.

[3] Mertens, P., 1990. L'intonation. In *Le français parlé. Etudes grammaticales*, Blanche-Benveniste, C.; Bilger, M.; Rouget, C.; Eynde, K. van den (eds) Paris: Editions du CNRS, 159-176.

[4] Mertens, P., 1993. Accentuation, intonation et morphosyntaxe. *Travaux de Linguistique* 26, 21-69.

[5] Mertens, P., 1997. De la chaîne linéaire à la séquence de tons. *Traitement Automatique des Langues* 38, 27-51.

[6] Mertens, P., 1999. Un algorithme pour la génération de l'intonation dans la parole de synthèse. *Traitement Automatique du Langage Naturel 1999.* Cargèse, 233-242.

[7] Mertens, P.; Goldman, J-Ph.; Wehrli, E.; Gaudinat, A., 2001. La synthèse de l'intonation à partir de structures syntaxiques riches. *Traitement Automatique des Langues* 42 (1), 145-192.

[8] Mertens, P.; Auchlin, A.; Goldman, J-Ph.; Grobet, A., 2001. L'intonation du discours: une implémentation par balises ; motifs et premiers résultats. *Journées Prosodie*. Grenoble.

[9] Patterson, D.; Ladd, R., 1999. Pitch Range Modelling: Linguistic dimensions of variation. *International Congress of Phonetic Sc.*, San Francisco, 1169-1172.