SYNTHESIZING FRENCH PROSODY FROM A PHONOLOGICAL REPRESENTATION

Pierre Larrey, Nadine Vigouroux, Guy Pérennou
IRIT, Université Paul Sabatier, Toulouse, France

ABSTRACT

The work presented here aims to implement a concept-to-speech synthesizer that would allow a dialog manager to make use of prosody to achieve communicative goals. The approach chosen makes use of abstract phonological commands that can drive a context-sensitive prosody. The paper focuses on the phonetic implementation of the prosodic commands to drive a speech synthesizer. Using auditory tests to ensure perceptual equivalence, an intermediate phonetic representation is proposed to reduce the perplexity of the phonetic implementation. This representation permits testing and refining the phonetic and the phonological of the model. Perceptual and statistical results are presented showing the validity of the chosen approach.

1. INTRODUCTION

1.1. Objectives

We are developing Interactive Vocal Systems that are able to generate speech under semantic and pragmatic control. In this framework, we are looking for a set of commands that would permit the Dialogue Manager to control the prosody of the produced messages. Recent progress in Intonational Phonology allow descriptions of intonation with generative capacities. This paper presents our work on the phonetic implementation of one particular phonological model for French intonation.

1.2. Applications

This work offers testing and refining the phonetic and the phonological of the model, and proposes an interface between speech synthesis and language generation but also opens opportunities in several other contexts :

- formal study of French tones patterns and their syntactic, semantic and pragmatic correlates
- modular assessment of speech synthesis components
- perceptual categorization of utterances, and then objective similarity measurements
- use of prosody as a complementary modality for speech synthesis in Interactive Systems (e.g. for blind people)
- prosodic labeling of speech databases
- speaking styles characterization
- phonetic science teaching or second language acquisition

1.3. Underlying Principles

The mapping between the prosodic commands and the acoustical prosodic information is by no mean straightforward. We believe this process must be an incremental and reversible stylization and symbolic coding of acoustic parameters. Thus we proceed by defining several intermediate representations that allow reversibility, modularity and separate validation of the different algorithms. The underlying principles of this approach are *perceptual equivalence* [12] and *perplexity reduction*. Perceptual equivalence is defined by listeners auditory judgements, while we define perplexity as the maximal number of different surface forms. The goal is to reduce perplexity gap between adjacent representations and to validate the mapping by means of perceptual equivalence tests.

Those perceptual and computational constraints can be solved by assuming a phonological description of intonation. Indeed, a phonological description is at least composed of : a level of description where continuous speech parameters are described by discrete and distinctive categories, and implementation rules that describe the mapping the categories and the continuous parameters [8]. This is the reason why we decided to implement a phonological model of French intonation.

2. PHONOLOGICAL MODEL

2.1. Choice of the model

One particular approach of intonation phonology is generative *per se*, the autosegmental metrical theory, see [8] for a review. The phonological models for French in this context are Hirst [6], Post [13], Jun [7], and Mertens [10]. Hirst and Mertens models are closer to phonetic representation, and their assumptions proved to be useful for the definition of an intermediate phonetic representation. Post model is strictly phonological but emphasizes syntactic constraints and fail to capture melodic properties essential in French, like for instance the particular treatment of the initial accent. Because, it has not been implemented yet and because it does not show the same outcomes, we chose the Jun and Fougeron [7] model for French intonation.

2.2. Description of the model

The model admits two level tones (L, H) and distinguishes Hi, the initial accent and H* the final accent. The lowest prosodic domain is the Accentual Phrase (AP) which has the underlying tonal pattern of LHiLH* that can be realized or not. Two levels of boundary tones have been defined (H-,L-) and (H%,L%) respectively delimiting two other domains in the hierarchy : the intermediate phrase (ip) and the intonation phrase (IP).

[_____] IP
[_____] [_____] ip
[_____][_____][_____] AP
[____] [____][____] [_____] [_____] [_____] W
[__][_][____][__][___] [____] [____] [___] [__][_][___] F
[__][_][____][__][___] [____] [____] [___] [__][_][___] S.

Fig 1. Prosodic hierarchy of the model (S=syllable, F=foot, W=phonologic word).

To represent the phonological description that can drive the prosody generation in a single string, we defined a Prosody Synthesis Markup Language (PSML), thus the prosodic tree is represented by a set of tags and attributes. The focus features have not been implemented in the present study.

Tag	Attributes
UTT	pitch, ratio, voice
IP	register, amplitude, **final** (boundary tone)
inter (ip)	**final** (boundary tone)
AP	focus
tone	**pa** (pitch accent), height, weight, range

Table 1. PSML components (bold attributes are obligatory ones, others have default values)

Here is an example of an utterance described by a PSML string : "Votre ordinateur ne se trouve pas configuré pour un affichage en deux cent cinquante six couleurs." *(Your computed is not configured for a 256 colors display).*

```
<UTT pitch=0.8><IPfinal=L%><inter final=H->
   <AP> vo <tone pa=L> tRoR di na <tone
   pa=H*> t9R
<inter final=H->
   <AP> <tone pa=L> n2 s2 <tone pa=Hi> tRuv
   <tone pa=H*> pa
   <AP> <tone pa=Hi> ko~ fi <tone pa=L> gy
   <tone pa=H*> Re
<inter final=H->
   <AP> <tone pa=L> pu R9~ <tone pa=Hi> na
   fi <tone pa=H*> SaZ
<inter final=L->
   <AP> <tone pa=L> a~ <tone pa=Hi> d2 sa~
   se~ ka~t si <tone pa=L> ku <tone pa=H*>
   l9R
</UTT>
```

2.3. Phonetic Implementation Guidelines

Some phonetic implementation rules given in the literature (see [13] for a review) are compatible with the chosen model. Those are guidelines that we had in mind when searching for an intermediate phonetic level.

1. Metrical strength : the metrical tree defined by the hierarchy determines relative prominence of syllables, which in turn influences duration (among other constraints)
2. Upstep : a high boundary tone enhances the level of the last pitch accent which becomes the most prominent
3. Downstep : the level of successive H* is progressively reduced except for the last one (upstepped), in every domain
4. Internal fall : the fall from the H* to the following L starts during the H* in case of AP final, but stays a rise in case of upstep
5. Contrast enhancement : the L before a H final tone is lower than the L before a L final tone

Our introductory discussion on the need of an intermediate phonetic level becomes clear, at the light of these rules which do not give details about F0 values nor phonemic duration but clearly deal with categorical entities.

3. INTERMEDIATE PHONETIC LEVEL

The actual input to a synthesizer generally consists in the duration of each phoneme and the description of a F0 curve using points in the time-frequency space. We call this representation a Numerical Phonemic Plan. This representation has a high perplexity (about $5x10^{16}$ different values per syllable, if considering only 3 F0 points per vowel). Several researchers have proposed models that reduce the number of parameters needed for the description of a F0 curve : for instance INTSINT [4], Tilt [14], a Tonal Perception model [1]. We present here a model based on syllabic categories that facilitates the expression of implementation rules and has been statistically and perceptually validated. This model, which integrates F0 and duration is reversible, linear for an easier mapping and the labels are directly linked to the segmental string.

3.1. Methodology of Development and Validation of an Intermediate Phonetic Level

3.1.1. Description of the corpora. Two corpora have been used for this study. One is publicly available within the MBROLIGN project [9] and consists of 70' (14 898 syllables) or read speech automatically labeled and manually corrected (corpus LE SOIR). The other corpus consists of 9' (980 syllables) of "acted" speech by a professional speaker, which is a rather different speaking style, closer to the one required in an Interactive Vocal System. The corpora composed of NNPs have been split so that short stimuli were obtained. Le Soir has been split into 1287 stimuli, while the other one in 48. The speech signals have not been considered, each representation only dealing with the adjacent one. The problems induced in French by the mute e have been eluded by suppressing all word-final e, and replacing all other schwa by the corresponding phoneme (@->2 in SAMPA codes).

3.1.2. Degradation Category Rating tests. From all the levels of representation considered an analysis/resynthesis procedure was implemented, and the perceptual equivalence of different representations of the same phonemic plan has been evaluated by DCR tests. The test platform has been developed as a web interface, where listeners were confronted to one pair of utterance and the basic five judgement categories. Additionally, listeners could specify which syllables were particularly affected by the procedure. The results of these tests were both used to validate representations and algorithms but also to choose the optimal values of certain parameters.

3.2. Symbolic Coding of Duration

Listeners have the tendency to be more perceptive to changes in syllabic duration than in phonemic ones. Thus, the first level of parameters reduction is (1) to mix the phonemic durations into syllables duration and (2) to compute the phonemic durations from the syllabic one. The first task has been implemented following phonological rules [10], the second using the elasticity principle [3], with a modification of the accommodation algorithm for final syllables, because of a stronger compression of syllables onset in that position observed in the corpus.

Then a more abstract representation has been decided : the syllabic durations have been categorized by a division of the normalized distribution sets (following z-scores). Results are shown in table 3, and DCR tests indicated that listeners found hardly perceptible a categorization between 7 syllabic categories. This optimal number of categories results from two constraints :

perceptual equivalence (DCR score), and reduction of perplexity (minimal number of categories).

3.3. Symbolic Coding of the FO Curve

The simple but efficient method for duration coding has been applied to represent the F0 for each syllable : after normalization, each syllable has been affected to a melodic category.

3.3.1. Normalization. Results (see table 3) showed when stylizing an F0 curve by using only three pitch points per vowel, that listeners do not find it degrading, the F0 of each syllable is then represented by 3 values respectively the pitch value at the start, middle, and end of the vowel expressed in percentage of the maximum and minimum value of the utterance in the semitone scale.

3.3.2. Categorization. We have searched for a categorization of syllable melody knowing that although some syllables are well represented by a single value, some others need to be described as moves. This well known phenomenon expresses that an F0 curve is a sequence of static and dynamic tones. We applied an algorithm of tone detection using glissando thresholds [1], assuming that in French only two tones per syllable were sufficient. Thus a syllable is represented by :

- a melodic category : the subdivision of the normalized space (pitch range between highest and lowest points of the utterance) into N parts
- a tonal category consisting of one of the nine combinations of fall, rise or level tones.

The first series of results are summarized in table 3. Here is a partial representation of a syllabic categorized plan using 7 duration categories, 4 melodic categories (L low, M middle, H high, . same) and 9 tonal categories (RR, FF, RF, FR, CF, FC, RC, CR, .) that has been automatically obtained from a NPP (compare the perplexity of 4x9x7 to the $5x10^{16}$ of the lowest representation) :

Syllable	Duration	Melodic	Tonal
Top=93	Bottom=81 (ST)		
vo	4	L	.
tRoR	3	.	.
di	3	M	.
na	4	.	.
t9R	6	H	RR
_	1		
n2	2	L	.
s2	4	.	.
tRuv	5	H	CF
pa	4	L	.
ko~	4	H	RR
fi	4	.	.
gy	4	M	.
Re	6	.	CR
_	5		

Table 2. Example of automatically extracted Intermediate Phonetic Representation “Votre ordinateur ne se trouve pas configuré…” *(Your computer is not configured…)*

3.4. Results

Table 3 shows both statistical and perceptual evaluation of the analysis/resynthesis procedure, between various representations and the original representation, adjacent scores being omitted. The first level of representation corresponds to the accommodation of syllable duration over phonemes duration. The second level is the use of 7 syllabic categories to reduce perplexity. Results for these levels concerns phoneme duration. The third level is a normalization of F0 curve in a semitone scale using three points per vowel. The fourth level is a categorization of the normalized F0 curve in 4 melodic categories and 9 tonal categories. The MBROLA [5] synthesizer has been used. The statistical results are obtained from the corpus Le Soir (N=1287). For the Accommodation and the Duration Categorization, the RMS (R) of phoneme duration difference was measured as well as the correlation coefficient (C). For the F0 normalization and categorization, the RMS and the correlation of F0 values has been computed. For the statistical evaluation of F0 coding, original durations were used. For the perceptual evaluation, the duration categorization was effective. Those scores were similar to the small corpus, not presented here. However, the perceptual evaluation is presented only for the small corpus (48 utterances). The number of pairs the subjects listened to is indicated by (N). A rating of 5 is “no difference perceived”, a rating of 4 corresponds to “small differences non disturbing”, the average rating is indicated by D.

	Level 1 *ms*	Level 2 *Ms*	Level 3 *Hz*	Level 4 *Hz*
Statistical Evaluation	R=4.29 C=0.93	R=5.89 C=0.87	R=2.94 C=0.92	**R=12.1** **C=0.79**
Perceptual Evaluation	D=4.55 N=60	D=4.42 N=60	D=4.20 N=60	**D=4.05** **N=150**

Table 3. Results of the analysis/resynthesis using 7 duration categories, 4 melodic categories, and 9 tonal categories

3.5. Conclusion

We finally reached a representation perceptually equivalent to the original representation, which has a considerably low perplexity and has an expression proximity with the phonological representation : duration categories can be specified from metrical strength (phrasing rules of the model), complex tonal categories occur at accented syllables or at boundary tones, melodic extreme categories can easily be mapped from H or L phonological level tones, and intermediate phonetic categories can express downstep rules or H and L variations.

4. PHONETIC AND PHONOLOGIC MAPPING

4.1. Corpus Labeling and Resynthesis

The small corpus has been labeled according to the phonological model. This task has been done, with auditory and visual display of F0 curve. The extracted phonetic representation, and the resynthesized speech have not been used. Phonetic rules have been implemented in a straightforward way : a duration category was assigned to each final syllable of each domain, L tones before H tones received the lowest melodic category, L tones

before final L tones received the median melodic category. H* were assigned to the highest category, which was decreased from AP to AP, with a reset at each ip beginning and a higher reset (to top category) for a new IP. Tonal categories were used to enhance boundary tones towards the high level (H- received RR) or low (L- and L% received FF) while, internal rises (H*) received RF tonal category.

4.2. Results

Two methods of resynthesis were used. In the first one both duration and F0 were affected, while in the second the original phonemes duration were preserved so that only the melodic part of the model is evaluated. The F0 RMS and correlation coefficient have been obtained extracting F0 every 10 ms and removing information of unvoiced segments. The scores when keeping original durations (21.4 and 0.67) are similar to other F0 generation methods [2] (although observed on biggest corpora) . The DCR average ratings range between 4 “perceptible but non disturbing differences” and 3 “perceptible and little disturbing differences”.

	Full Model	Original Duration
Syl. duration RMS	15.6	0
Syl. Duration Correl.	0.69	1
F0 RMS	25.8	21.4
F0 Correl	0.48	0.67
DCR	3.43	3.94

Table 4. Comparison ratings between original utterances and synthesized utterances from the PSML phonologic representation

Another advantage of the intermediate phonetic model is to permit an evaluation in terms of category shifts that give a good idea of the accuracy of the phonologic to phonetic mapping (table 5). Further work should be done in the perceptual weighting of category shifts, because sometimes a shift of one category has been perceived while often not. It should be noted that the shift of one and two melodic categories are due to the downstep rule that do not apply systematically The second reason is the absence of L* in the model, that we have generally labeled H*+L-, but sometimes the L- boundary tone was not possible. In these cases, a L* pitch accent would have generated the correct melodic category. The scores of the tonal categories shift are due to the difference between late or early pitch movements and full pitch movements (e.g. CR, RC and RR) that were not modeled.

Shift	0	1	2	More
Duration	36.3%	42.9%	16.5%	4.3%
Melodic	34.6%	39.2%	18.0%	8.2%
Tonal	67.9%	16.7%	13.5%	1.9%

Table 5. Category shift between the observed and generated phonetic representations (980 syllables corpus)

5. CONCLUSION

The synthesized method presented here generates speech from a phonological representation. A phonetic intermediate model has been introduced so that the expression of phonetic rules is easier and straightforward. Perceptual tests showed that the synthetic prosody could match a natural one. However, the approach chosen clearly shows the limits of a rule-driven method and bigger corpus should be used to perform statistical training and possibly help refining phonetic rules for theoretical purposes. Further, we believe that the close links between the phonetic and phonologic representations make feasible the automatic extraction of phonological labels from speech.

ACKNOWLEDGMENTS

We would like to thank Sun-Ah Jun for the precision on her model, Farbice Malfrère for the MBROLIGN project and the MBROLA team for providing phonetic synthesizer.

REFERENCES

[1] d’Alessandro C., Mertens P. 1995. Automatic Pitch Contour Sylization unsing a Model of Tonal Perception, *Computer Speech and Language* 9(3), pp 257-288

[2] Black, A. and Hunt, A. (1996). Generating FO contours from ToBI labels using linear regression Proceedings of ICSLP 96, vol 3, pp 1385-1388, Philadelphia, Penn.

[3] Campbell N. 1991. Segment duration in a syllable frame, *J. Phon.*, Vol 19, 1991

[4] Campione E., Flachaire E., Hirst D., Véronis. J. 1997. Stylization and Symbolic Coding of F0 : a Quantitative Model, *Proceedings of ESCA Workshop on Intonation*, Athens, pp 71-74

[5] Dutoit T., Pagel V.,Pierret N., Bataille F., van der Vrecken O. 1996. The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes, *Proceedings of ICSLP'96*, Philadelphia, vol. 3, pp. 1393-1396.

[6] Hirst D. 1988. Tonal Units as Constituents of Prosodic Structure: The Evidence from English and French Intonation, in Hulst, H van der and Smith (eds), Autosegmental Studies on Pitch Accent. Dordrecht : Foris. pp 151-165

[7] Jun S.A., Fougeron C. 1995. The Accentual Phrase and the Prosodic Structure of French, *Proceedings of the XIIIth ICPhS, Stockholm*, pp 722-724

[8] Ladd D.R. 1996 *Intonational Phonology*, Cambridge University Press

[9] Malfrère F., Dutoit T., 1997. High Quality Speech Synthesis for Phonetic Speech Segmentation, *Proceedings of the European Conference On Speech Communication and Technology*, pp. 2631-2634, 1997

[10] Mertens P. 1993. Intonational Grouping, Boundaries and Syntactic Structure in French, House and Touati (eds), *Proceedings of an ESCA Workshop on Prosody*, Lund, pp 156-159

[11] Pérennou G. 1978. Note interne CERFIA sur la syllabation, Toulouse. France

[12] de Pijper J.R, 1983. *Modeling British English Intonation*, Dordrecht : Foris

[13] Post B. 1993 *A Phonological Analysis of French Intonation*, MA Thesis. Nijmegue. The Netherlands

[14] Taylor P. 1998. The Tilt Intonation Model. *Proceedings of the ICSLP Conference, Sydney*