IDENTIFICATION AND AUTOMATIC GENERATION OF PROSODIC CONTOURS FOR A TEXT-TO-SPEECH SYNTHESIS SYSTEM IN FRENCH

S. de Tournemire
France Telecom, CNET (Centre National d'Etudes des Télécommunications)
Technopole Anticipa, 2 avenue Pierre Marzin, 22307 Lannion Cedex
E-mail : detourns@lannion.cnet.fr

ABSTRACT

This paper presents the realisation of an automatically trainable computational prosodic model for French Text-to-Speech Synthesis. The methodology proposes the construction of the model in two steps.

The first step consists in predicting fundamental frequency contours and duration of syllables from abstract prosodic markers using neural networks [17,12].

In this step, the abstract prosodic markers are automatically extracted from the signal by analysing prosodic realisations [2] and identifying a prosodic alphabet and a set of labelling rules.

The second step integrates the model into the CNET Text-to-Speech Synthesis system [7] by using its linguistic levels and predicting abstract prosodic markers from text and linguistic labels.

The system is evaluated by naïve listeners and compared with the actual CNET Text-to-Speech Synthesis system.

INTRODUCTION

In French, as in most Indo-European languages, a sentence can be said with many different prosodic contours. The prosody depends on extra-linguistic phenomena (speaker), paralinguistic phenomena (doubt, happiness,...) and linguistic phenomena (syntax, semantics, pragmatics). Unfortunately, most Text-to-Speech (TTS) synthesis systems do not take into account all these levels of variability. For example, in some TTS synthesis systems, the synthesis of a new voice consists in modifying acoustic levels but excludes modifying the prosodic ones.

Automatic learning techniques offer some solutions for this problem because they allow prosodic regularities to be automatically extracted from a prosodic database of natural speech. Such techniques depend on the construction of a large generally hand-labelled corpus which is extremely time consuming and is an obstacle to rapidly adapting the prosody.

We present a solution which quasi-automatically "captures" a new prosody from a corpus of natural speech. This solution is presented in the first part of the paper. In the second, third and fourth parts, the corpus preparation, the prosodic labelling of the corpus and the automatic learning of prosodic contours are respectively described. Those three stages produce models permitting prosodic contours to be predicted from prosodic labels. In the fifth part, the generation of the prosodic labels from text is described. This allows the contour prediction models to be integrated into the TTS synthesis system. Finally, in part 6, the overall system is evaluated.

1. METHODOLOGY

TTS synthesis includes 4 different processes:

1. **Linguistic processing** which makes a sentential analysis to establish the phonemic transcription, part-of-speech tagging and syntactic structures.
2. **Symbolic prosodic processing** which predicts a symbolic description of the text prosody (break locations, accentuation, ...).
3. **Numeric prosodic processing** which predicts prosodic parameters (fundamental frequency (F0) and segment duration).
4. **Acoustic processing** which generates the speech signal.

In this work, we are interested in the prosodic processing (symbolic and numeric). Linguistic and acoustic processing will be furnished by the CNET French Text-to-Speech synthesis system (CNETVOX [7]; PSOLA [4]).

We first propose to develop the numeric prosodic processing. In this step, prosodic labels used for the automatic learning of prosodic contours are extracted from the recorded corpus. In this way, the prediction model parameters are optimised on well labelled data. The three main stages are: **1) The Corpus preparation** which extracts and models the prosodic parameters from signal. **2) Prosodic labelling from signal** which identifies a "prosodic alphabet" and labels the corpus with this alphabet. **3) Automatic learning of prosodic contours** which trains a model with data resulting from 1) and 2) to learn the parameters of the prosodic contours prediction model.

The symbolic prosodic processing is realised in a second step. It consists in deducting prosodic labels from the CNETVOX linguistic processing.

Finally, the prosodic processing (symbolic, numeric) are integrated into the CNET TTS synthesis system.

2. CORPUS PREPARATION

2.1 Extraction of prosodic parameters

The corpus is composed of 312 utterances of declarative sentences of variable size (4173 words and 6767 syllables) offering a large variety of lexical, syntactic and semantic forms. It has been read by a professional female speaker and automatically segmented into acoustic segments [1]. An F0 calculation is made at each transition between segments, giving 2 F0 values and one duration per segment.

As the syllable is an essential unit in basic auditory grouping [16], we effectuate a syllable based modelling of duration and stylisation of F0 contours.

2.2 Duration modelling

The duration modelling consists in replacing the segment's duration by a syllable elasticity factor *k* [2] using the expression:

$$k = \frac{D_{syll} - \sum \mu_{seg,C}}{\sum \sigma_{seg,C}}$$

where $\mu_{seg,C}$ and $\sigma_{seg,C}$ are the mean and the standard deviation respectively of a particular segment type *seg* in the context *C* and D_{syll} is the syllable duration.

Given the segment, its context and the elasticity factor of the host syllable, the segment duration D_{seg} is calculated with: $D_{seg} = \mu_{seg,C} + k\ \sigma_{seg,C}$

2.3 F0 Stylisation

The F0 stylisation is based on 4 points of the syllable F0 contour considered as phonologically relevant [5]: the beginning of the syllable, the beginning and the end of the vowel and the end of the syllable. The stylised F0 curve is obtained by linear interpolation between these points.

3. PROSODIC LABELLING FROM SIGNAL

Most methods of automatically learning prosodic contours for TTS synthesis rely on hand labelled training data [12,17]. Systems to automatically label a speech signal with prosodic patterns for speech recognition [2,18,19] also need hand labelled training data. As there is no consensus concerning the prosodic labels in French, we first propose to identify a set of labels (a "prosodic alphabet") and then establish labelling rules using this alphabet. This method permits the recorded corpus to be automatically labelled but the labelling rules may need to be adjusted for a new corpus.

3.1 Prosodic principles

The identification of the prosodic alphabet is based on prosodic principles specifying the main prosodic events and their location in French.

We consider that the main prosodic events (pauses, syllable lengthening, F0 movements) take place at the end of prosodic words (minimal accentuated units, [9]) where final stress is realised. In addition, secondary stress assumes a rhythmic function, preventing large distances between two final and/or emphatic stresses [15].

From these principles, a symbolic description of prosodic parameters is made with 3 types of labels: "break labels" and "F0 shape labels" which are identified at the end of prosodic words, and "accent labels" which are identified within prosodic words.

3.2 Identification of prosodic alphabet

3.2.1 Identification of "Break labels"

We make a distinction between punctuation breaks (full-stop and comma in this corpus), pause breaks and lengthening breaks.

In addition to punctuation pauses, pauses can be realised on syntactic boundaries or between two words with common boundaries constituted of vocalic elements [6]. The analysis of the duration distribution for such pauses allows 3 classes of pause duration to be identified (see frame 1).

A similar analysis is made for lengthening, using CNETVOX lengthening prediction module. We finally obtain the following alphabet:

B0. Full-stop
B1. Comma
B2. Long pause: more than 220 ms
B3. Medium pause: between 120 and 220 ms
B4. Short pause: between 60 and 120 ms
B5. Strong lengthening: elasticity factor is more than 1
B6. Weak lengthening: elasticity factor between 0 and 1
B7. Prosodic word boundary

Frame 1 : Breaks alphabet

3.2.2 Identification of "F0 shapes"

We make the assumption that F0 shapes on the last syllable of a prosodic word are composed of at most 2 elementary shapes (rise, fall or flat). After analysing the combinations of elementary shapes effectively realised on the corpus, we identify the more frequent F0 shapes and the ones rarely realised. We finally obtain the alphabet below:

S0. Flat : F0 variation between -1 and 1 semi-tone
S1. Fall : F0 variation inferior to -1 semi-tone
S2. High rise : F0 variation superior to 6 semi-tones
S3. Rise : F0 variation between 1 and 6 semi-tones
S4. Fall-Rise
S5. Flat-Fall
S6. Rise-Fall

Frame 2 : F0 shapes alphabet

3.2.3 Accent identification

Besides internal accent, there are two other types of accent: emphatic accent characterised by a high F0 rise at the beginning of a word and the secondary accent characterised by a F0 rise on the first or the antepenultimate syllable of a word [11,15]. The comparison between F0 variations on different

syllable locations in a word let us identify the 2 classes below:

A1. Weak accent : F0 rise between 2 and 4 semi-tones
A2. Strong accent : F0 rise superior to 4 semi-tones

Frame 3 : Accents alphabet

An example of prosodic labelling is given here:

Example: "Il fallait avoir /B7-S1/ d'autres motifs, /B1-S2/ comme par exemple, /B1-S2/ la ma/A1/ladie /B7-S0/ de sa mère./B0-S1/." ("One should have other motives, as for example , the illness of ones mother.")

3.3 Automatic labelling

The definition of the prosodic alphabet contains F0, duration and lengthening thresholds that let us write labelling rules. These rules allow us to automatically transcribe the corpus. This labelled corpus is then used during the automatic learning stage.

4. AUTOMATIC LEARNING OF PROSODIC CONTOURS

Automatic learning techniques have been widely used to generate prosodic parameters (F0 and duration). Probabilistic models [8,14], classification trees [13] and neural networks [17,12] have given conclusive results. Nevertheless, such techniques have not experimented with French for which systems based on rules and the concatenation of predefined prosodic forms are predominant [7]. Since neural networks have proven successful in the automatic learning of F0 contours in German [17] and segment duration in Italian [12], we choose this technique for our system.

Neural network architecture and parameters are determined in an experimental way. They end up in a two-layer fully-connected neural net. The activation functions are sigmoidal with slope respectively 1 and 0.2 for hidden and output layers. The mean square error is used for back-propagation.

For F0 learning, input vectors contain prosodic labels (breaks, F0 shapes and accentuation) for the current and contextual syllables (the 2 previous ones and the 4 next ones). Output vectors contain 4 F0 values of the syllable coded on a logarithmic scale.

For duration learning, input vectors contain prosodic labels (breaks, accentuation and information about syllable composition) for the current and contextual syllables. Output vectors contain the elasticity factor for the syllable.

The resulting models are evaluated when they are integrated into a TTS synthesis system (section 6).

5. GENERATION OF PROSODIC CONTOURS FROM TEXT

We now have a model to generate prosodic contours (F0 and duration) from prosodic labels. We want to integrate this model into a TTS synthesis system. This requires the prosodic labels to be generated from text.

CNETVOX linguistic processing includes text pre-processing, part-of-speech tagging, phonemic transcription, and prosodic break location (determined by 275 syntactic rules). A grouping of prosodic breaks and a mapping with the break alphabet defined in frame 1, locates break labels. Then, an analysis of F0 shapes more frequently realised for each break label gives the F0 shape labels shown in Table 1.

CNETVOX	Break	n°	F0 shape	n°
0	full-stop	B0	fall	S1
1	comma	B1	high rise	S2
2-25	long pause	B2	high rise	S2
25-75	medium pause	B3	high rise	S2
75-100	short pause	B4	rise	S3
100-225	strong lengthening	B5	rise	S3
225-275	weak lengthening	B6	rise	S3
prosodic word	neither pause nor lengthening	B7	fall	S1

Table 1 : Breaks and F0 shapes deducted from CNETVOX.

As we can see in table 1, not all F0 shapes appear amongst the labels and the same shape is always associated with a given break. Some automatic learning techniques [10] have been investigated but they have not yet improved the labelling.

6. EVALUATION OF THE OVERALL SYSTEM

6.1 Objective evaluation

Figure 1 compares generated F0 and duration curves with natural F0 and duration curves for one sentence.

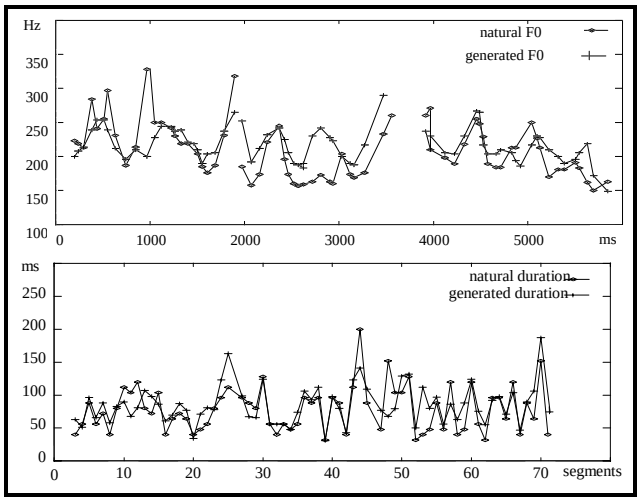

Figure 1 : F0 and duration contours

As we can see, the generated contours respect the general trends of the natural ones.

6.2 Subjective evaluation

6.2.1 Methodology

We made a perceptive multi-criteria evaluation including a quality and an intelligibility test. This methodology has been defined and adopted as "a methodology for evaluating synthetic speech quality" [3]. It consists in evaluating 5 systems: 3 types of

natural speech (natural non-degraded speech, speech with a signal-to-noise ratio of 20dB and speech with a signal-to-noise ratio of 10dB) and 2 types of synthetic speech (one with CNETVOX prosody and the other with automatically generated prosody – pauses are the same for both). All speech material is filtered in the telephonic band and 8kH sampled. The evaluation is made with 16 naïves listeners.

6.2.2 Results

As show in figure 2, for most of the criteria, the synthesis system with generated prosody yields slightly better scores than the synthesis system with CNETVOX prosody. This difference is significant for acceptability criterion in both tests. Both types of synthetic speech are scored between natural speech with a signal-to-noise ratio of 20dB and natural speech with a signal-to-noise ratio of 10dB.

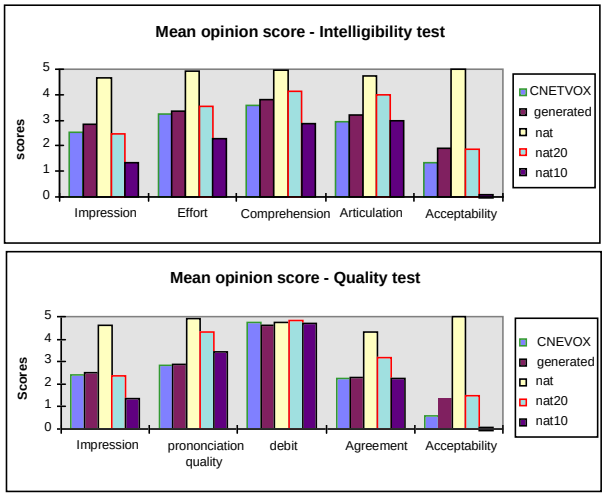

Figure 2 : Subjective evaluation results

CONCLUSION

We propose a methodology for constructing a prosodic contour generation model. This methodology allows the system to be automatically adapted to a new corpus (consisted of a new voice or a specific application). The model is realised and integrated into the CNET TTS synthesis system for French. The overall system is evaluated with objective and subjective criteria. The results show that the automatically trainable system is perceived as good as the hand-crafted CNETVOX system, and better under some acceptability criteria.

REFERENCES

[1] Boëffard, O., (1993), "Segmentation automatique d'unités acoustiques pour la synthèse de la parole", Thesis, Université de Rennes I.

[2] Campbell, W.N., (1993), "Detecting prosodic boundaries in a speech signal", ATR Research Activities of the Speech Processing Department, Jan-march 1993.

[3] Cartier, M., Emerard, F., Pascal, D., Combescure, P., & Soubigou, A., (1992), "Une méthode d'évaluation multicritère de sorties vocales. Application au test de 4 systèmes de synthèse à partir du texte", J.E.P., Bruxelles, pp. 117-122.

[4] Charpentier, F., & Moulines, E., (1989), "Nouvelles techniques de synthèse de la parole", L'écho des recherches N°137.

[5] De Tournemire, S., (1994), "Recherche d'une stylisation extrême des contours de F0 en vue de leur apprentissage automatique", J.E.P., Trégastel.

[6] Emerard, F., (1977), "Synthèse par diphones et traitement automatique de la prosodie", Thesis, Université de Grenoble, France.

[7] Larreur, D., Emerard, F., & Marty, F., (1989), "Linguistic and prosodic processing for a text-to-speech synthesis system", Eurospeech, pp. 510-513, Paris.

[8] Ljolje, A., & Fallside, F. (1986), "Synthesis of Natural Sounding Pitch Contours in Isolated Utterances Using Hidden Markov Models", IEEE Trans. on Acoust., Speech, and Signal Proc., 34, pp. 1074-1080.

[9] Martin, P., (1974), "Eléments pour une théorie de l'intonation", Rapport d'activité de l'Institut phonétique, Bruxelles, n° 9/1, 97-126.

[10] Ostendorf, M., & Veilleux, N., (1994), "A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary location", Computational Linguistics, Volume 20, Number 1.

[11] Pasdeloup, V. (1988), "Essai d'analyse du système accentuel du français: distribution de l'accent secondaire", 17ème J.E.P. Nancy.

[12] Quazza, S., (1995), "Predicting durations by means of automatic learning algorithms", IV Workshop of the Experimental Phonetics Group, Turin, Italy.

[13] Riley, M.D., (1992), "Tree-based modelling of segmental duration", Talking Machines: Theories, models, and Designs, G. Bailly, C. Benoit, T.R. Sawallis (Editors) Elsevier Science Publishers B.V.

[14] Ross, K. N., (1995), "Modeling of intonation for speech synthesis", PhD dissertation, Boston University.

[15] Rossi, M., (1985), "L'intonation et l'organisation de l'énoncé", Phonética 42: pp. 135-153.

[16] Tatter, V.C., (1975), "Selective adaptation of acoustic and phonetic detectors", MA Thesis, Brown University.

[17] Traber, C., (1992), "Fo generation with a data base of natural F0 patterns and with a neural network", Talking Machines : Theory, Models and Designs, pp. 287-304.

[18] Vaissière, J., (1989). "On automatic extraction of prosodic information for automatic speech recognition system.", Eurospeech, Vol. 1. , pp. 202-205.

[19] Wang, M.Q., & Hirschberg, J., (1992), "Automatic classification of intonational phrase boundaries", Computer Speech and Language, 6, pp. 175-196.