AUTOMATIC TRANSCRIPTION OF INTONATION USING AN IDENTIFIED PROSODIC ALPHABET

S. de Tournemire

France Télécom, CNET (Centre National d'Etudes des Télécommunications)

ABSTRACT

A solution is proposed for rapidly adapting prosodic models to a new voice or a new application. First, a prosodic alphabet that is supported by linguistic knowledge is identified at the acoustic level. The observation of the realisation of prosodic events on the acoustic corpus allows classes of breaks, F0 shapes and accents to be constructed and automatic transcription rules to be written. Then the transcribed corpus is used in the estimation of the parameters of a prosodic model for French. The good F0 contours and duration generated with the prosodic model verify the agreement of the identified alphabets to describe prosodic phenomena. Finally, the prosodic model is integrated in the CNET standard French Text-to-Speech Synthesis system. The quality of the generated prosody is considered by naïve listeners as equivalent to the handcrafted system. This result verifies the appropriateness of the alphabet as prosodic descriptors.

1. INTRODUCION

In most Indo-European languages, a sentence can be spoken with many different prosodic contours. The prosody depends on many factors (syntax, semantics, pragmatics, and speaker) which are difficult to model by one simple Text-to-Speech (TTS) synthesis system. Linguistic and acoustic levels of the TTS system have to be adapted to the application or the speaker.

Automatic learning techniques offer a solution to this problem of adapting prosodic models to a new voice or a new application, 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 corpus, which is generally hand-labelled. This labelling process is extremely time-consuming and is an obstacle to rapidly adapting the prosody.

Prosodic transcription makes a symbolic representation of prosodic phenomena by associating labels with acoustic realisations. This symbolic representation is between a linguistic (syntax, semantic) and acoustic (F0, phone duration, pauses, energy) description of prosody. It must not be too far from a linguistic level if it is to be predicted from the linguistic descriptors. It must also not be too far from an acoustic level if it is to be automatically extracted from the signal. For this reason, the alphabet must incorporate both linguistic and acoustic information.

First, the prosodic alphabet is identified for two speakers and the corpus is transcribed. Then, the transcribed corpus is used in the estimation of the parameters of a prosodic model for French. The figure 1 resumes these two steps in order to make the methodology explicit.

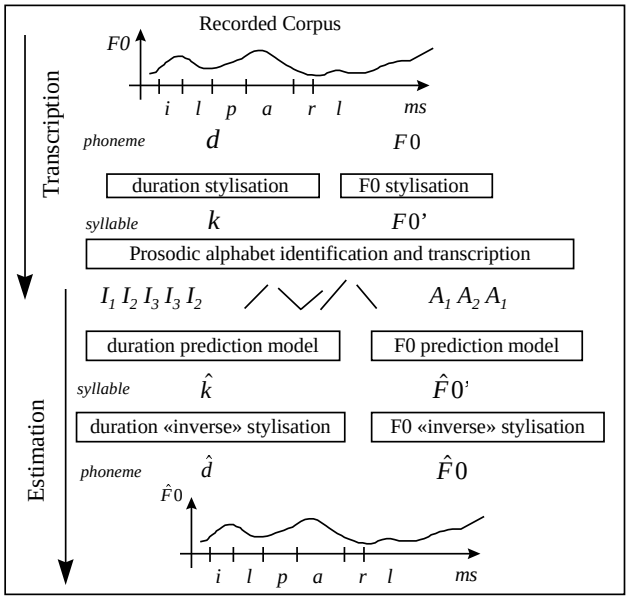

Figure 1: Transcription of the corpus and estimation of prosodic model parameters

Finally, the prosodic model is integrated into the CNET standard French Text-to-Speech Synthesis system by predicting prosodic labels from linguistic descriptors. Results of a subjective evaluation of the overall system are presented.

2. PROSODIC ALPHABET AND PROSODIC TRANSCRIPTION

Prosodic transcription assigns prosodic symbols (for instance break indices) to the speech signal. In a study of French, Mertens [11] distinguishes three levels of representation: the acoustic, perceptual, and linguistic levels.

- Acoustic level transcription makes a symbolic description of prosodic parameters. It is used for speech recognition and for automatic transcription of prosodic databases [15,16].
- Perceptual level transcription makes a symbolic description of what is perceived. It includes pitch level, accentuation, syllable lengthening, pause and respiration. It is used for studying intonation and is done manually because there is no tool that can provide a systematic transformation of the acoustic F0 data into an estimate of the perceived pitch.

- Linguistic or morphological level is derived from the perceptual level or predicted from text. It is an abstract representation of intonational units.

Depending of what needs to be modelled or studied, one or another of these levels is used. In this work concerning prediction of F0 contours and duration from text, just two levels of description of prosody can be used: the linguistic and acoustic levels. The perceptual level is necessary only for evaluation but can be excluded from the process of predicting prosodic parameters from text. In this way the transcription process is simplified and can be automated.

The proposed solution identifies a prosodic alphabet at the acoustic level but considers linguistic knowledge. The symbols of this alphabet are not very different from those already proposed for French (Martin [10], Hirst [8], Mertens [11], for instance). The main difference is that they are found automatically at the acoustic level and are dependent on the corpus from which they are identified. The following paragraphs will show how adequately chosen symbols lead to good prosody when associated with automatic training of the F0 and duration prediction module.

2.1. Prosodic Principles

The identification of the prosodic alphabet is based on prosodic principles specifying the main prosodic events and their location in French. The main prosodic events (pauses, syllable lengthening, F0 movements) take place at the end of prosodic words (minimal accentuated units) where final stress is realised. In addition, secondary stress assumes a rhythmic function, preventing large distances between two final and/or emphatic stresses.

From these principles, a symbolic description of prosodic parameters is made with three 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.

2.2. Corpus Preparation

The corpus is composed of 312 utterances of declarative sentences of variable size (4173 words, 6767 syllables) offering a large variety of lexical, syntactic and semantic forms. The context of utilisation of speech synthesis justifies that the corpus is not issued from spontaneous speech (services which have to give precise information without any hesitation and prepared in advance). The corpus has been read by two professional speakers and automatically segmented into acoustic segments [2]. An F0 calculation is made at each transition between segments, giving two F0 values and one duration per segment.

Speaker		Mean	Standard-deviation
1 (man)	*F0*	217Hz	39Hz
	Duration	82ms	41ms
2 (woman)	*F0*	104Hz	24Hz
	Duration	85ms	41ms

Table 1: Acoustic description of the corpus.

Duration modelling

The modelling of prosodic parameters with a unit bigger than the phoneme enables variations due to micro prosody or intrinsic segmental duration of the speaker to be separated from those due to the prosodic organisation of the text. This separation allows the parameters to be optimised for two simple models [3] instead of for a very complex one [1]. As the syllable is an essential unit in basic auditory grouping, a syllable based modelling of duration and stylisation of F0 contours is done.

The duration modelling consists of replacing the duration of the segment by a syllable elasticity factor *k* [3] 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 estimated elasticity factor $\hat{k}$ of the host syllable, the segment duration $\hat{D}_{seg}$ is given by:

$$\hat{D}_{seg} = \mu_{seg,C} + \hat{k} \cdot \sigma_{seg,C}$$

The mean modelling error is 8ms and 11ms for the two speakers. This error is strongly correlated with the number of samples for each class of phoneme. The CNET PSOLA synthesis system [7] is used to calculate a new speech signal from the modelled duration. During an informal subjective evaluation, expert auditors did not perceive any differences between the original and modelled duration.

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 end of the vowel, and the end of the syllable. The stylised F0 curve is obtained by linear interpolation between these points. The mean error is 1Hz for both speakers and no difference is easily perceived between the stylised contour and the original one. In fact, this stylisation is very close to a stylisation by two points per phoneme.

2.3. Identification of Prosodic Alphabet

Identification of “Break labels ”

A distinction is made between final punctuation breaks (full stop in this corpus), pause breaks (including comma breaks) and lengthening breaks.

Pauses can be realised on punctuation, syntactic boundaries or between two words with common boundaries consisting of vocalic elements. The analysis of the duration distribution for such pauses allows three classes of pause duration to be identified.

A similar analysis is made for lengthening, using the CNETVOX lengthening prediction module [9].

Identification of "F0 shapes"

It is assumed that F0 shapes on the last syllable of a prosodic word are composed of at most two elementary shapes (rise, fall or flat). After analysing the combinations of elementary shapes effectively realised on the corpus, the more frequent F0 shapes and the ones rarely realised are identified.

Accent identification

Besides internal accent, there are two other types of accent in French. The emphatic accent is characterised by a high F0 rise at the beginning of a word. The secondary accent is characterised by a F0 rise on the first or the antepenultimate syllable of a word [12]. The comparison between F0 variations on different syllable locations in a word enables the two classes given in table 2 to be identified.

Table 2 gives the prosodic alphabets identified for the two speakers. In the table, k is syllable lengthening factor, a is F0 amplitude (semitones) and d is pause duration (milliseconds).

Type	Thresholds Speaker 1	Thresholds Speaker 2
B0 : full stop		
B1. Long Pause	$d \geq$ 220ms	$d \geq$ 450ms
B2. Medium Pause	$120 \leq d < 220ms$	$96 \leq d < 450ms$
B3. Small Pause	$d <$ 120ms	$d <$ 96 ms
B4. Strong lengthening	$k \geq 1$	$k \geq 1$
B5. Weak lengthening	$0 < k < 1$	$0 < k < 1$
B6. No lengthening	$k \leq 0$, prosodic word frontier	$k \leq 0$, prosodic word frontier
S0. Flat	$a \leq 1$	$a \leq 1$
S1. High fall	No	$a > 6$
S2. Fall	$1 < a \leq 6$	$1 < a \leq 6$
S3. High rise	$a > 6$	$a > 6$
S4. Rise	$1 < a \leq 6$	$1 < a \leq 6$
S5. Fall-rise	a > 1	a > 1
S6. Rise-fall	a > 1	a > 1
A1. Weak accent (secondary accent)	F0 rise $2 < a \leq 4$	F0 rise $4 < a \leq 7$
A2. Strong accent (insistence *accent*)	F0 rise $a > 4$	F0 rise $a > 7$

Table 2: Prosodic alphabets.

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.")

2.4. Automatic Transcription

The definition of the prosodic alphabets contains F0, pause duration and syllable lengthening thresholds that permit labelling rules to be written. These rules allow the corpus to be automatically transcribed. This labelled corpus is then used during the automatic learning stage.

3. AUTOMATIC LEARNING AND GENERATION OF PROSODIC CONTOURS FROM TEXT

Automatic learning techniques have been widely used to generate prosodic parameters (F0 and duration). Probabilistic models, classification trees, and neural networks 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 [10]. Since neural networks have proven successful in the automatic learning of F0 contours in German [15] and segment duration in Italian [14], this technique as been chosen for the system described here.

3.1. Automatic Learning

Neural network architecture and parameters are determined in an experimental way. They end up in a two-layer fully connected neural network trained by a back propagation algorithm. The input vectors have been chosen using linguistic knowledge and mutual information. For F0 learning, input vectors contain prosodic labels (breaks, F0 shapes and accentuation) for the current and contextual syllables [6]. Output vectors contain four 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 lengthening factor for the syllable.

The mean errors resulting from the test database for the two speakers are 20Hz and 16Hz per phoneme for F0 and 17ms and 16ms per phoneme for duration. The F0 contours and duration predicted by the neural network have been considered by three expert listeners as the same quality as the natural ones.

3.2. Generation of Prosodic Contours from Text

The integration of the model into a TTS synthesis system 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 [9]. A grouping of prosodic breaks and a mapping with the break alphabet defined in table 2 locates break labels on the text. Then, an analysis of F0 shapes more frequently realised for each break label gives the F0 shape labels.

3.3. Evaluation of the Overall System

A perceptive multi-criteria evaluation including a quality and an intelligibility test [4] has been made. In this methodology, five systems are evaluated: three 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 two 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 8kHz sampled. The evaluation is made by 16 naïves listeners.

For both speakers, the synthesis system with generated prosody and the synthesis system with CNETVOX prosody yield nearly equivalent scores (see figure 2). 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.

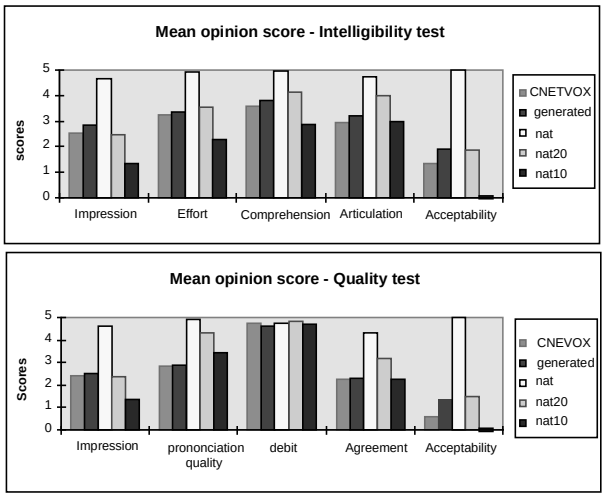

Figure 2 : Subjective evaluation results

This result is very encouraging because CNETVOX results from many years of expertise at CNET. As this system offers very good quality of prosody in French, the main objective in this work was to obtain a quality equivalent replacing laborious craftsmanship by automatic methodology.

4. CONCLUSION

A solution is presented which quasi-automatically "captures" a new prosody from a corpus of natural speech. The methodology of construction of the prosodic model includes three main steps. Firstly, abstract prosodic markers are automatically extracted from the signal by analysing prosodic events and identifying a prosodic alphabet and a set of labelling rules. Secondly, neural networks predict fundamental frequency contours and syllable duration from abstract prosodic markers. In this way, the prediction model parameters are established from well-labelled data. Finally, the model is integrated into the CNET Text-to-Speech Synthesis system by using its linguistic levels and predicting abstract prosodic markers from text and linguistic labels. The evaluation results show that the automatically trainable system is perceived as good as the handcrafted CNETVOX system, and better under some acceptability criteria.

REFERENCES

1. Bartkova, K., & Sorin, C., (1987), "A model of segmental duration for speech synthesis in French ", Speech Communication, 6, North-Holland, pp. 245-260.

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

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

4. 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., Brussels, pp. 117-122.

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. De Tournemire, S., (1998), "Identification et génération automatique de contours prosodiques pour la synthèse vocale à partir du texte", Thèse de doctorat. Ecole Nationale supérieure des télécommunication, Paris.

7. Hamon, C., Moulines, E., & Charpentier, F., (1989), "A diphone synthesis system based on time-domain modifications of speech", Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. 238-241, Glasgow.

8. Hirst, D., & Di Cristo, A. (1986), "Unités tonales et unités rythmiques dans la représentation de l'intonation", Actes des 15èmes Journées d'Etude sur la Parole, Aix en Provence, pp. 93-95.

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

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

11. Mertens, P., (1987), "L'intonation du Français. De la description linguistique à la reconnaissance automatique", Doctoral dissertation, Catholic University of Leuven.

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

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

14. 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.

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

16. Wightman, C. W., & Ostendorf, M., (1994), "Automatic labeling of Prosodic Patterns", IEEE Transactions on Speech and Audio Processing, vol. 2, n° 4.