ABSTRACT

This paper describes the use of prosodic rules for segmenting speech signal in sentences, words (lexical and grammatical) and putting milestones on their boundaries. The strategy in an automatic speech recognition system uses these milestones for filtering lexical and syntactical hypotheses in order to limit the depth of the lexical and syntactical search.

Prosodic parameters (pitch, energy, duration) are estimated over vocalic nuclei and normalized using two types of coding: a) uniform coding, b) variable coding.

INTRODUCTION

Several factors influence the prosodic contours of a sentence namely: the fundamental frequency variations, the energy variations and the duration of segments. The most important factors are:

- (a) linguistic context: elocution style, conversation type (natural dialogue or reading text mode), (b) textual context (sentence mode: statement or interrogative, sentence structure, etc.) and the interaction between the successive segments (coarticulation effect) -- the presence of a phonological boundary between two phonemes may reduce the amount of the mutual interaction -- (c) speaker variability (socio-linguistic factors, speaking rate) and the strategy used by the speaker (distribution of the accents, distribution of pauses, etc.), (d) pragmatic situation.

Prosodic contours show great variability from one speaker to another for the same situation of a dialogue.

On the other hand, prosody has a certain relation with linguistic levels [3], [8], [11]. For this reason prosody provides information at several levels in an ASRS (Automatic Speech Recognition System):

- Phonetic level: By using micro-prosody effects (for example open vowels are more energetic than closed vowels, the micro-melody of sonorous occlusive and nasal consonant are quite different, VOT (Voice Onset Time) is a possible acoustic cue for stop consonant recognition, etc.);
- Lexical level: The accents in the words do not follow a random distribution, but they depend first on the position of the word in the prosodic group, and second on their syllable number and size.
- Syntactic and Semantic levels: The syntactic markers (pauses) and the semantic markers (degree of emphasis which underlines the meaning) contribute to the oral punctuation of the sentence in the deep structure level [4],[5],[6].

The objective of using prosodic components in an ASRS is to decode the information carried by suprasegmental parameters at the position of stressed syllables in the sentence, and about the boundaries of words and sentences.

Thus several questions can be asked:

1. Do deep prosodic frames exist which are useful for speech recognition and which can be made to appear by direct processing of prosodic parameters?
2. What are the convenient units over which prosodic parameters can be estimated?
3. What are the convenient coding alternatives?
4. What is the best strategy for utilizing the prosody in ASRS?

2. PROSODIC SYSTEM DESCRIPTION

2.1. Prosodic parametrization

Prosodical parameters (pitch, energy, duration) are descriptive, which makes their use very difficult. Consequently, they must be estimated over a discrete unit. For the fundamental frequency and the energy parameters we have chosen the vocalic nuclei because they carry the most significant variations for these two parameters. For the duration parameter we have the choice between two units:

1. the syllable: The syllable is very convenient for representing the linguistic level but it has not an acoustical reality because we cannot exactly detect its boundaries.
2. the pseudo-syllable: We can define the pseudo-syllable as a segment of speech which is delimited by the centers of two successive vocalic nuclei. Figure 1 illustrates the two units.

We remark from Figure 1 that the concept of pseudo-syllable can deviate from the concept of syllable when the syllabic structure becomes complex (for instance in series of CVC or VCC). In order to assess the pseudo-syllable concept we have compared the relative rhythmic curves estimated using the duration of syllables and pseudo-syllables for 60 sentences spoken naturally by three French speakers. This experiment was first realized using sentences manually segmented into syllables; the relative rhythmic curve is then estimated as the difference between the duration of two successive syllable divided by the duration of the first syllable, as indicated in the following relation:

\[ R_{s}(n)=\frac{(D_{s}(n)-D_{s}(n-1))}{D_{s}(n-1)} \]

where \( R_{s}(n) \) = The relative rhythm at the nth syllable.
\( D_{s}(n) \) = The duration of the nth syllable.

Next the relative rhythmic curve of pseudo-syllables was estimated in the same manner using the duration between the centers of two successive vocalic nuclei detected automatically over the sentence:

\[ R_{ps}(n)=\frac{(D_{ps}(n)-D_{ps}(n-1))}{D_{ps}(n-1)} \]
where Rps(n)=The relative rhythm at the nth vocalic nucleus center.
D(n)=The duration of the nth pseudo-syllable.

Figure 2 shows that the syllabic and pseudo-syllabic rhythmic curves are highly correlated.

The following remarks are mentioning for each one of the curves:
(a) the relative rhythmic curve is estimated as indicated in relation (2), (b) the linear melodic curve is estimated by the linear adjustment of the melodic curve over the vocalic nuclei, (c) the skeleton of melodic curve is estimated as the mean of the fundamental frequency over vocalic nuclei. The declination line is estimated as the linear adjustment of the points of the melodic skeleton curve, (d) the skeleton of intensity curve is estimated as the curve which joints energy over successive vocalic nuclei, (e) fundamental frequency curve.

2. In the second stage prosodic parameters are coded and other functions are derived from these parameters. These functions are:

NV(n) = the nth vocalic nucleus.
Ec(n) = energy (coded in 4 levels) for NV(n).
Fc(n) = value of F0 for NV(n).
D(n) = duration between two successive vocalic nucleus centers.

where Rps(n)=The relative rhythm at the nth vocalic nucleus center.

D(n)=The duration of the nth pseudo-syllable.

Figure 3 describes the general scheme of this system. The three stages are briefly presented below:

1. In the first stage the detection of vocalic nuclei is accomplished [13], and then prosodic parameters (fundamental frequency, energy, duration) are estimated over vocalic nuclei. Figure 4 shows the output of this stage.

Figure 4 shows the output of this stage.

The following remarks are mentioning for each one of the curves:
(a) the relative rhythmic curve is estimated as indicated in relation (2), (b) the linear melodic curve is estimated by the linear adjustment of the melodic curve over the vocalic nuclei, (c) the skeleton of melodic curve is estimated as the mean of the fundamental frequency over vocalic nuclei. The declination line is estimated as the linear adjustment of the points of the melodic skeleton curve, (d) the skeleton of intensity curve is estimated as the curve which joins energy over successive vocalic nuclei, (e) fundamental frequency curve.

2. In the second stage prosodic parameters are coded and other functions are derived from these parameters. These functions are:

NV(n) = the nth vocalic nucleus.
Ec(n) = energy (coded in 4 levels) for NV(n).
Fc(n) = value of F0 for NV(n).
D(n) = duration between two successive vocalic nucleus centers.

where Rps(n)=The relative rhythm at the nth vocalic nucleus center.

D(n)=The duration of the nth pseudo-syllable.

Figure 3 describes the general scheme of this system. The three stages are briefly presented below:

1. In the first stage the detection of vocalic nuclei is accomplished [13], and then prosodic parameters (fundamental frequency, energy, duration) are estimated over vocalic nuclei. Figure 4 shows the output of this stage.

The proposed prosodic system is divided into three stages Figure 3 describes the general scheme of this system. The three stages are briefly presented below:

1. In the first stage the detection of vocalic nuclei is accomplished [13], and then prosodic parameters (fundamental frequency, energy, duration) are estimated over vocalic nuclei. Figure 4 shows the output of this stage.

The following remarks are mentioning for each one of the curves:
(a) the relative rhythmic curve is estimated as indicated in relation (2), (b) the linear melodic curve is estimated by the linear adjustment of the melodic curve over the vocalic nuclei, (c) the skeleton of melodic curve is estimated as the mean of the fundamental frequency over vocalic nuclei. The declination line is estimated as the linear adjustment of the points of the melodic skeleton curve,(d) the skeleton of intensity curve is estimated as the curve which joins energy over successive vocalic nuclei, (e) fundamental frequency curve.

2. In the second stage prosodic parameters are coded and other functions are derived from these parameters. These functions are:

NV(n) = the nth vocalic nucleus.
Ec(n) = energy (coded in 4 levels) for NV(n).
Fc(n) = value of F0 for NV(n).
D(n) = duration between two successive vocalic nucleus centers.

where Rps(n)=The relative rhythm at the nth vocalic nucleus center.

D(n)=The duration of the nth pseudo-syllable.

Figure 2 shows that the syllabic and pseudo-syllabic rhythmic curves are highly correlated.

The proposed prosodic system is divided into three stages Figure 3 describes the general scheme of this system. The three stages are briefly presented below:

1. In the first stage the detection of vocalic nuclei is accomplished [13], and then prosodic parameters (fundamental frequency, energy, duration) are estimated over vocalic nuclei. Figure 4 shows the output of this stage.

The following remarks are mentioning for each one of the curves:
(a) the relative rhythmic curve is estimated as indicated in relation (2), (b) the linear melodic curve is estimated by the linear adjustment of the melodic curve over the vocalic nuclei, (c) the skeleton of melodic curve is estimated as the mean of the fundamental frequency over vocalic nuclei. The declination line is estimated as the linear adjustment of the points of the melodic skeleton curve,(d) the skeleton of intensity curve is estimated as the curve which joins energy over successive vocalic nuclei, (e) fundamental frequency curve.

2. In the second stage prosodic parameters are coded and other functions are derived from these parameters. These functions are:

NV(n) = the nth vocalic nucleus.
Ec(n) = energy (coded in 4 levels) for NV(n).
Fc(n) = value of F0 for NV(n).
D(n) = duration between two successive vocalic nucleus centers.

where Rps(n)=The relative rhythm at the nth vocalic nucleus center.

D(n)=The duration of the nth pseudo-syllable.

Figure 3 describes the general scheme of this system. The three stages are briefly presented below:

1. In the first stage the detection of vocalic nuclei is accomplished [13], and then prosodic parameters (fundamental frequency, energy, duration) are estimated over vocalic nuclei. Figure 4 shows the output of this stage.

The following remarks are mentioning for each one of the curves:
(a) the relative rhythmic curve is estimated as indicated in relation (2), (b) the linear melodic curve is estimated by the linear adjustment of the melodic curve over the vocalic nuclei, (c) the skeleton of melodic curve is estimated as the mean of the fundamental frequency over vocalic nuclei. The declination line is estimated as the linear adjustment of the points of the melodic skeleton curve,(d) the skeleton of intensity curve is estimated as the curve which joins energy over successive vocalic nuclei, (e) fundamental frequency curve.

2. In the second stage prosodic parameters are coded and other functions are derived from these parameters. These functions are:

NV(n) = the nth vocalic nucleus.
Ec(n) = energy (coded in 4 levels) for NV(n).
Fc(n) = value of F0 for NV(n).
D(n) = duration between two successive vocalic nucleus centers.

where Rps(n)=The relative rhythm at the nth vocalic nucleus center.

D(n)=The duration of the nth pseudo-syllable.

Figure 3 describes the general scheme of this system. The three stages are briefly presented below:

1. In the first stage the detection of vocalic nuclei is accomplished [13], and then prosodic parameters (fundamental frequency, energy, duration) are estimated over vocalic nuclei. Figure 4 shows the output of this stage.

The following remarks are mentioning for each one of the curves:
(a) the relative rhythmic curve is estimated as indicated in relation (2), (b) the linear melodic curve is estimated by the linear adjustment of the melodic curve over the vocalic nuclei, (c) the skeleton of melodic curve is estimated as the mean of the fundamental frequency over vocalic nuclei. The declination line is estimated as the linear adjustment of the points of the melodic skeleton curve,(d) the skeleton of intensity curve is estimated as the curve which joins energy over successive vocalic nuclei, (e) fundamental frequency curve.

2. In the second stage prosodic parameters are coded and other functions are derived from these parameters. These functions are:

NV(n) = the nth vocalic nucleus.
Ec(n) = energy (coded in 4 levels) for NV(n).
Fc(n) = value of F0 for NV(n).
D(n) = duration between two successive vocalic nucleus centers.

where Rps(n)=The relative rhythm at the nth vocalic nucleus center.

D(n)=The duration of the nth pseudo-syllable.

Figure 3 describes the general scheme of this system. The three stages are briefly presented below:

1. In the first stage the detection of vocalic nuclei is accomplished [13], and then prosodic parameters (fundamental frequency, energy, duration) are estimated over vocalic nuclei. Figure 4 shows the output of this stage.

The following remarks are mentioning for each one of the curves:
(a) the relative rhythmic curve is estimated as indicated in relation (2), (b) the linear melodic curve is estimated by the linear adjustment of the melodic curve over the vocalic nuclei, (c) the skeleton of melodic curve is estimated as the mean of the fundamental frequency over vocalic nuclei. The declination line is estimated as the linear adjustment of the points of the melodic skeleton curve,(d) the skeleton of intensity curve is estimated as the curve which joins energy over successive vocalic nuclei, (e) fundamental frequency curve.

2. In the second stage prosodic parameters are coded and other functions are derived from these parameters. These functions are:

NV(n) = the nth vocalic nucleus.
Ec(n) = energy (coded in 4 levels) for NV(n).
Fc(n) = value of F0 for NV(n).
D(n) = duration between two successive vocalic nucleus centers.

where Rps(n)=The relative rhythm at the nth vocalic nucleus center.

D(n)=The duration of the nth pseudo-syllable.

Figure 3 describes the general scheme of this system. The three stages are briefly presented below:

1. In the first stage the detection of vocalic nuclei is accomplished [13], and then prosodic parameters (fundamental frequency, energy, duration) are estimated over vocalic nuclei. Figure 4 shows the output of this stage.

The following remarks are mentioning for each one of the curves:
(a) the relative rhythmic curve is estimated as indicated in relation (2), (b) the linear melodic curve is estimated by the linear adjustment of the melodic curve over the vocalic nuclei, (c) the skeleton of melodic curve is estimated as the mean of the fundamental frequency over vocalic nuclei. The declination line is estimated as the linear adjustment of the points of the melodic skeleton curve,(d) the skeleton of intensity curve is estimated as the curve which joins energy over successive vocalic nuclei, (e) fundamental frequency curve.

2. In the second stage prosodic parameters are coded and other functions are derived from these parameters. These functions are:

NV(n) = the nth vocalic nucleus.
Ec(n) = energy (coded in 4 levels) for NV(n).
Fc(n) = value of F0 for NV(n).
D(n) = duration between two successive vocalic nucleus centers.

where Rps(n)=The relative rhythm at the nth vocalic nucleus center.

D(n)=The duration of the nth pseudo-syllable.
3. The third stage of our prosodic system consist of the application of prosodic rules for labeling speech signal with the following labels:

\[ \begin{align*}
  \text{DP} &= \text{beginning of the sentence.} \\
  \text{FP} &= \text{end of the sentence.} \\
  \text{MG} &= \text{grammatical word.} \\
  \text{DM} &= \text{beginning of a lexical word.} \\
  \text{FM} &= \text{end of a lexical word.}
\end{align*} \]

Prospodic rules are inferred using the coded prosodic parameters at the output of the second stage by a statistical study of these parameters over vocalic nuclei using a manually segmented corpora and then are assessed and are filtered linguistically.

3. PROSODIC RULES

We have inferred 15 prosodic rules; we present below five of them:

3.1. Beginning and end of a sentence:

\[ \begin{align*}
  \text{Ph1. IF} & \ \ (\partial D(n) \leq 500 \text{ms}) \text{ OR } (\text{NV}(n) = \text{EOF}) \\
  \text{THEN} & \ \ \text{NV}(n) \leftarrow \text{DP}; \\
  & \ \ \text{NV}(n-1) \leftarrow \text{FP}+\text{FM}.
\end{align*} \]

**Comment:** A large increase in the duration between two vocalic nuclei indicates the beginning of a new sentence for the nth vocalic nucleus, and the end of the previous sentence for the n-1 vocalic nucleus. Generally, the end of a sentence corresponds to the end of a lexical word.

\[ \begin{align*}
  \text{Ph2. IF} & \ \ (\partial D(n) \geq 300 \text{ms}) \text{ AND } (Fc(n-1) < 2) \\
  & \ \ \text{AND } (Ec(n-1) < 2) \\
  \text{THEN} & \ \ \text{NV}(n) \leftarrow \text{DP}'; \\
  & \ \ \text{NV}(n-1) \leftarrow \text{FP}+'\text{FM}'.
\end{align*} \]

**Comment:** For this rule we accept a less important increase in the duration but we impose a lower energy level and a lower fundamental frequency level.

3.2. Grammatical words:

\[ \begin{align*}
  \text{MG. IF} & \ \ (Fc(n)-Fc(n-1) < 2) \text{ AND } (\partial D(n) < 0) \\
  \text{AND} & \ \ (\partial D(n)-\partial D(n-1) \leq 50 \text{ms}) \\
  \text{THEN} & \ \ \text{NV}(n) \leftarrow \text{MG}; \\
  & \ \ \text{NV}(n-1) \leftarrow \text{FM}.
\end{align*} \]

**Comment:** In this rule the fundamental frequency falls dramatically (skipping at least 2 levels) with a minimum acceleration of 50ms, this rule labels the nth vocalic nucleus by the MG and the n-1 vocalic nucleus by FM.

3.3. End of lexical words:

\[ \begin{align*}
  \text{ML1. IF} & \ \ (\partial D(n) > 0) \text{ AND } (\text{DLF}(n) < 0) \\
  & \ \ \text{AND } (Fc(n)=4) \text{ AND } (Ec(n) \geq 3) \\
  \text{THEN} & \ \ \text{NV}(n) \leftarrow \text{FM}.
\end{align*} \]

**Comment:** An increase of the duration parameter \(\partial D(n) > 0\) accompanied with a melodic accent and a strong energy in the nth vocalic nucleus indicates the end of a lexical word.

3.4. Beginning of lexical word:

\[ \begin{align*}
  \text{ML2. IF} & \ \ (\partial D(n) = 0) \text{ AND } (Fc(n)=3) \text{ AND } \\
  & \ \ (Fc(n+1)=4) \text{ AND } (Ec(n)=5) \\
  \text{THEN} & \ \ \text{NV}(n) \leftarrow \text{DM}.
\end{align*} \]

**Comment:** An acceleration accompanied by an increase in the fundamental frequency and with a strong energy in the nth vocalic nucleus indicates the beginning of a lexical word.

4. PROSODIC ROLE IN THE DIRA ASRS

DIRA is a multi-expert system organized around a blackboard [2], these experts (Acoustic phonetic decoder, lexical analyzer, syntactic semantic analyzer, understanding model) are controlled by a supervisor.

In the DIRA ASRS we have attributed a very important role to the prosodic labels, this role can be distributed to the two following levels:

a. The lexical level: In this level, we have attributed to every word in the lexical knowledge two pointers which indicate the place of the first and the last vocalic nuclei in the word. The lexical analyzer will filter the developed lexical words according to the coincidence of the prosodic labels at the output of prosodic analyzer with the prosodic pointers in developed words.

b. The syntactical level: In this level, grammatical knowledge is represented by an automate in which the possible beginning or end of a sentence are indicated. Prosodic labels will aid to confirm the end or the beginning of the sentence and then to limit the number of hypotheses developed by this level.

5. RESULTS AND DISCUSSION

We have experimented two sets of prosodic rules, which are inferred using uniform and variable coding, over a total number of 120 sentences spoken by three speakers, tables 1, 2, 3 present respectively (the type of word existed in the corpora, the detection results using uniform coding, the detection results using variable coding):
Table 3. Result of boundary detection using variable coding.

<table>
<thead>
<tr>
<th>Label</th>
<th>FP+FM</th>
<th>DP</th>
<th>MG</th>
<th>DM</th>
<th>FM</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>correct</td>
<td>120</td>
<td>120</td>
<td>155</td>
<td>160</td>
<td>183</td>
<td>738</td>
</tr>
<tr>
<td>Wrong</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>12</td>
<td>6</td>
<td>32</td>
</tr>
<tr>
<td>percentage</td>
<td>100%</td>
<td>100%</td>
<td>91.7%</td>
<td>93%</td>
<td>96.8%</td>
<td>95.8%</td>
</tr>
</tbody>
</table>

DP=beginning of a sentence, FP=end of a sentence, MG=grammatical word, DM=beginning of a word, FM=end of a word.

Tables 2 and 3 show that variable coding improves the performance of prosodic rule because it codes prosodic parameter curves correctly along the sentence.

Figure 7 presents an example of labeling two sentences using the above rules.

6. CONCLUSION

In this paper we have presented a method of using unweighted prosodic parameters, (pitch, energy, duration) coded in two ways (uniform, variable), in prosodic rules in an ASRS.

The method proposed has the following benefits:
- It is a bottom-up method.
- It provides milestones which are regularly spread-out over the sentence.
- The number of syllables between two milestones is not very big.
- The milestones are reliably detected.
- The milestones are used for reducing the number of hypotheses in the lexical and grammatical levels.

REFERENCES