COMPARATIVE STUDY BETWEEN UNIFORM AND VARIABLE CODING
USED FOR INFERRING PROSODIC RULES IN AUTOMATIC SPEECH
RECOGNITION EXPERT SYSTEMS

M.K. NASRI, G. CAELEN-HAUMONT, J. CAELEN
ICP/INPG-ENSERG - UNIVERSITE STENDHAL
UNITE ASSOCIEE AU C.N.R.S. N° 368
46, AVENUE FELIX VIALLER
38031 GRENOBLE CEDEX FRANCE

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.

1. 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, communication 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, the 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 Systems):

- 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 units. 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.

\[ E(dB) \]
\[ t \text{ (ms)} \]
\[ \text{V} \quad \text{C} \quad \text{C} \quad \text{V} \quad \text{C} \quad \text{V} \quad \text{C} \quad \text{V} \]
\[ \text{S} \quad \text{S} \quad \text{S} \quad \text{S} \quad \text{S} \]

**Figure 1:** Representation of vocalic nucleus ( NV), pseudo-
syllable (PS) and syllable (S).

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:

\[ \text{Rs}(n) = \frac{(\text{Ds}(n) - \text{Ds}(n-1))}{\text{Ds}(n-1)} \]

where \( \text{Rs}(n) \) = The relative rhythm at the nth syllable.
\( \text{Ds}(n) \) = 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:

\[ \text{Rps}(n) = \frac{(\text{Dps}(n) - \text{Dps}(n-1))}{\text{Dps}(n-1)} \]
where \( Rps(n) \) = the relative rhythm at the \( n \)th vocalic nucleus center.
\( Ds(n) \) = the duration of the \( n \)th pseudo-syllable.

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

2.2 System description

1. Vocalic nuclei detection.
2. Estimation of prosodic parameters over the vocalic nuclei.
3. Derivation of intermediate parameters.
4. Rules.

\[ \text{Process} \rightarrow \text{Prosodic parameters} \]

- \( D \) = Duration.
- \( F0 \) = Fundamental frequency.
- \( E \) = Energy.

Figure 3: General Scheme of prosodic system.

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 mentioned 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 linear adjustment of 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:

\[ NV(n) = \text{the } n \text{th vocalic nucleus.} \]
\[ Ec(n) = \text{energy (coded in 4 levels) for } NV(n). \]
\[ Fp(n) = \text{value of } F0 \text{ for } NV(n). \]
\[ Fc(n) = Fp(n) \text{ for } NV(n) \text{ coded in 4 levels [7].} \]
\[ D(n) = \text{duration between two successive vocalic nucleus centers.} \]
\[ aD(n) = D(n) - D(n-1) \text{ (} aD(n) < 0 \text{ => acceleration / slow down).} \]
\[ DLFc(n) = \text{the difference between the declination line of the duration } DL(n) \text{ and } D(n); \text{ this parameter in used for detecting the duration accent.} \]
\[ DLDc(n) = DLD(n) \text{ coded in 8 levels.} \]
\[ DLFc(n) = \text{the difference between the declination line of } F0 \text{ and the value of } F0(n). \]

The following remarks are mentioned for each one of the curves:

(a) uniform coding: For coding a prosodic parameter curve we divide its variations range uniformly into a number of levels, and then we code every point of this curve according to its position between the levels. Figure 5 shows an example of the fundamental frequency curve coded into four levels.

(b) variable coding: In this kind of coding the division of the parameter range will be variable along the sentence. Figure 6 presents this kind of coding.
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:

ML1. IF \( (\delta D(n) > 0) \) AND \((\delta Fc(n) < 0)\) THEN \( NV(n) < -FM \).

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

3.4. Beginning of lexical word:

ML2. IF \((\delta D(n) < 0)\) AND \((Fc(n) = 3)\) AND \((Fc(n+1) = 4)\) AND \((Ec(n) = 5)\) THEN \( NV(n) = -DM \).

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 the 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>
<thead>
<tr>
<th>Type</th>
<th>MG(m)</th>
<th>MG(b)</th>
<th>ML(m)</th>
<th>ML(b)</th>
<th>ML(t)</th>
<th>ML(q)</th>
<th>ML(c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>words</td>
<td>366</td>
<td>12</td>
<td>307</td>
<td>270</td>
<td>48</td>
<td>18</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 1. Number and type of words in our corpora.

m=1 syllable, b=2 syllables, t=3 syllables, q=4 syllables, c=5 syllables.

<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>150</td>
<td>156</td>
<td>174</td>
<td>720</td>
</tr>
<tr>
<td>Wrong</td>
<td>0</td>
<td>0</td>
<td>24</td>
<td>24</td>
<td>18</td>
<td>66</td>
</tr>
<tr>
<td>percentage</td>
<td>100%</td>
<td>100%</td>
<td>86.2%</td>
<td>86.6%</td>
<td>90.6%</td>
<td>91.6%</td>
</tr>
</tbody>
</table>

Table 2. Result of boundary detection using uniform 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>

Table 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.

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.

6. CONCLUSION

REFERENCES

Contribution à la détermination automatique de la prosodie du français parlé à partir d'une analyse syntaxique. Etablissement d'un modèle de génération. Thèse de Doctorat d'Ingénieur, INP Grenoble.

Meta-stratégie en Reconnaissance dans le projet "DIRA". 17èmes JEP, GALP-SPA, Nancy.