THE USE OF PROSODIC AGENTS IN A COOPERATIVE AUTOMATIC SPEECH RECOGNITION SYSTEM

Ph. Langlais and J.L. Cochard
IDIAP, CP 592, CH-1920 MARTIGNY
e-mail: langlais@idiap.ch cochard@idiap.ch

ABSTRACT

Two prosodic agents of a cooperative speech recognition system (namely $\mathsf{ETC}_{\mathsf{vérif}}$) will be presented in this paper. The first agent is processing information available in micro-prosodic variations. The second agent is dealing with linguistically-motivated aspects of prosody which are exceedingly useful to constraint solution space in a recognition task.

1. INTRODUCTION

It is often argued that prosodic can be used in numerous benefit ways in an automatic speech recognition process (ASR) [1]. Prosody is first involved in phonetically conditioned aspects (intrinsic values and coarticulation effects). It is well known for example that high vowels (such /*i*/) have an intrinsically lower duration than low vowels (such /*a*/). A recent study [2] measured a slight improvement of a large vocabulary speech recognition system by inserting a micro-prosodic model. Among all prosodic functions, one is of utmost importance from an ASR point of view: the grouping of related words in so-called prosodic words. Prosodic structuring can be useful for verifying and predicting linguistic organization proposed by other agents (syntactic or/and semantic ones). Numerous studies deal with this function and recent ones report interesting results for specific tasks such as disambiguation [3, 4]. Prosody is also reported to be useful in a speech understanding system specially in dialog situation where events such as repairs [5] and interrupts occur quite frequently [6]. This area is however far from the scope of this paper which will detail the first two enunciated points.

2. MICRO-PROSODIC AGENT

This agent is processing information available in intrinsic and co-intrinsic variations of fundamental frequency, intensity and duration parameters. It provides specially some weighted hypothesis to $\mathsf{ETC}_{\mathsf{vérif}}$ such as voice/voiceless diacritic recognition and voiced obstruent/non-obstruent consonant distinction. In this section we will just sum up special points that are described in depth, from a lexical access point of view, in [7].

Duration cues

We studied, on several corpora of French isolated words (ranging from 500 to 1000 words) uttered by several speakers, the vowels intrinsic durations and the right consonant effect (voice/voiceless and occlusive/constrictive) on the preceeding vowel. Durations were automatically obtained based on two different techniques; the first one is using the duration given by a lexical access module and the second one is based on non-contextual phonemic Markov models. We report hereafter major conclusions for this studies.

Even if high vowels durations are on average smaller than the ones of low vowels, intrinsic vowels durations are not reliable enough to be used in our system. In fact, only oral/nasal vowel distinction can be done (at least partially) with low error probability (see fig. 1).

Contextual effects can be observed on the average values but seem to be too fragile for classification techniques (error probability close to 0.4).

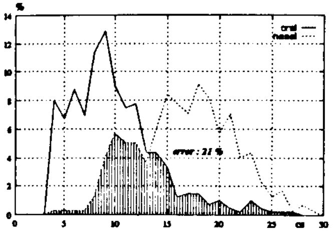

Figure 1. *Distributions of oral and nasal vowel durations on a corpus of 2 speakers' utterances of 800 tri-syllabic words.*

Intensity cues

We studied on the same corpora, the distributions of intensity values (measured by a classical raw power intensity) and we can conclude that local discrimination between vowels like /*a*/ and /*i*/ (see fig. 2) can be achieved with a reasonable probability error (at least for non final vowels), while pre-vocalic consonantic distinction can not.

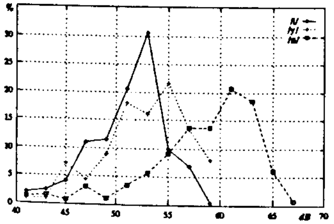

Figure 2. */i/, /y/ and /a/ distributions measured on initial vowels of tri-syllabic words*

Fundamental frequency cues

We measured this parameter with an algorithm implementing the *amdf* technique with satisfactory results. It provides a voice/voiceless decision based on the shape of *amdf* curve computed for each frame of signal. This method has been found very suitable for lexical filtering: more than 60% of a large lexicon was removed from potential candidates by the only mean of this decision with low error rate (less than 3%) on a task of 500 words recognition, each one uttered by 6 different speakers. In a top-down approach, the voice/voiceless distinction was useful too to re-rank lexical hypothesis with an average gain of 3 places.

As studied for duration and intensity, intrinsic and co-intrinsic frequency values have been considered and no robust information was discovered, except the obstruent/non-obstruent consonants distinction that can be achieved, at least partially, with reasonable error probability. The major cue of this distinction is the concave shape usually observed on non liquids intervocalic consonants (see fig. 3).

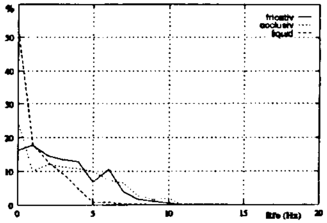

Figure 3. *Distributions of a concavity measure Rfo on different consonants classes.*

3. SUPRASEGMENTAL AGENT

Even if linguistically-motivated aspects of prosody gave rise to a lot of studies, there is not yet a unified model on which all researchers can agree. This is mainly due to the fact that prosodic phenomena depend on many distinct levels of linguistic representation. For example, even if it is well known that prosodic cues occur more frequently at syntactic constituents boundaries; this rule can be denied by other constraints such as rhythmic ones. Thus, learning by example seems to be an efficient way to solve this conflicting situation. We report here first results obtained by a suprasegmental agent based on this technique.

This agent makes use of an identification system of prosodic labels which points out, in a sentence, the occurrences of some particular prosodic cues (two-way emergence of a vowel fundamental frequency, lengthening of its duration, ...). The output of this treatment: a prosodic lattice, as well as the syntactic decomposition of the sentence, and its phonetic alignment (obtained by an automatic Viterbi alignment of allophones models) feed a statistical module which updates a knowledge source (KS). This KS quantifies — for a given corpus — the correlations between some syntactic, rhythmic and prosodic units.

Information is organized into a non-connected oriented graph. Each edge bears a syntactic and/or rhythmic constraint automatically derived from learning observations. Each vertex N (called a 'P-node') contains information such as the number of times it has been visited when processing learning data, the number of occurrences of each different prosodic label attached to observations and a tree structure $Crit(N)$ (called a 'SR-structure') describing the syntactic-rhythmic organization modeled by N. Each leaf node of the SR-structure holds all prosodic labels observed at the current constituent boundaries.

An observation O is a SR-structure with a fully described tree (*i.e.* syntactic tree with each node containing the right number of vowels). We denote p_o the depth of the tree and define d as the number of consecutive levels (beginning at root node) with fully instanciated numbers of vowels ($p \in [0, p_o]$ and $p \leq p_o$). $O_{p,d}$ is the SR-structure got from O by filtering out the p first levels with rhythmic constraints of the d first levels (*ex.* : $O_{p_o,0}$ is the syntactic structure of the observation, $O_{1,1}$ holds the number of vowels in the observation).

The graph grows automatically by updating each node holding a SR-structure that can be unified with $O_{p,d}$ and by creating missing nodes using the following algorithm:

```
$explore(N, O, p, d)$
  $if\,(p < p_o)$
    $if\,\exists\, N'$ : P-node / $Crit(N') = O_{p+1,d}$
      $then\,update\,N'$
      $else\,create\,N'$ : son of $N$
    $explore(N', O, p+1, j)$
  $if\,(d < p_o)$
    $if\,\exists\, N'$ : P-node / $Crit(N') = O_{p,d}$
      $then\,update\,N'$
      $else\,create\,N'$ : son of $N$
    $explore(N', O, p, d+1)$
```

An observation O can at most generate $\frac{p \times (p+3)}{2}$ P-nodes but in general factorization (depending on the application) significantly cuts down the graph expansion. This organization allows a user to easily query the system on particular syntactic-rhythmic structures, such as for example: NB(N1-999999(5).VIRG(2).N1-999(4)), that describes a number made up of three distinct groups with respectively 5, 2 and 4 vowels. The system answers by displaying figures of prosodic parameter contours (see fig. 4) modeled by the selected P-node with unifiable information and by providing a matrix describing the frequency of each prosodic label observed at this node.

This KS also provides a convenient way of scoring the adequacy between the measured prosodic cues and the syntactic-rhythmical structure that could be partially (internal node of the tree) or entirely (leaves

of the description tree) defined in order to give weighted hypothesis for a specific input. We report hereafter (see fig. 5) results of a number recognition task. We feed the system with 500 numbers uttered by 70 speakers on a telephone line.

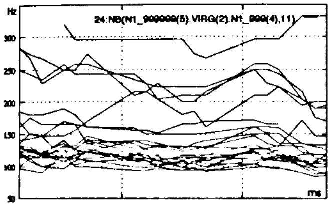

Figure 4. *Example of fo contours proposed by the system. Only initial, middle and final values of each group are taken into account and linearly smoothed.*

For each number, the system is provided with a syntactic structure, a phonetic alignment and a prosodic lattice (above 30 different labels). All these data are automatically computed from orthographic transcription. The system's objective is to predict the syntactic-rhythmic structure of a 100 numbers test data set. The score of a specific P-leaf is given by the maximum mark assigned to each possible path from the root to it. The score of a path is the average scoring of each of its P-nodes and a specific node in a path is scored by a distance between its local prosodic matrix and the input one. The figure 5 reports the ranking rate of the 100 observations.

4. DISCUSSION

This study demonstrates that most of intrinsic and co-intrinsic phenomena are difficult to handle, and only few cues seem to be useful for a recognition process. On this point, this study confirms Dumouchel's conclusions [2]. We propose a user-friendly and efficient system for scoring and/or predicting structural linguistic hypothesis that seems very promising for further investigation on prosody.

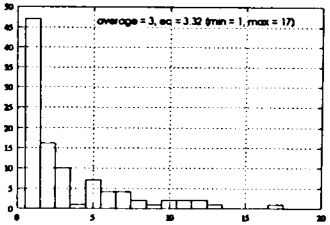

Figure 5. *Ranking of incoming candidates from an average number of possible ranks of 20.*

REFERENCES

[1] J. Vaissière. The use of prosodic parameters in automatic speech recognition. In *Recent Advances in Speech Understanding and Dialog Systems.* NATO ASI Series, 1988.

[2] P. Dumouchel. Suprasegmental features and continuous speech recognition. In *ICASSP*, pages 177-180, 1994.

[3] Andrew Hunt. A generalised model for utilising prosodic information in continuous speech recognition. In *ICASSP*, pages 169-172, 1994.

[4] Price and al. The use of prosody in syntactic disambiguation. In *DARPA workshop on Speech and Natural Language*, pages 372-377, Pacific Grove, Februar 1991.

[5] Nakatani and Hirschberg. A corpus-based study of repair cues in spontaneous speech. *Acoustical Society of America*, 95(3):1603-1616, March 1994.

[6] Kompe and al. Prosody takes over: Towards a prosodically guided dialog system. In *Speech Communication*, pages 157-167, 1994.

[7] Béchet and al. Lexical filtrering by means of prosodic information. In *XIIIth ICPhS*, Stockholm, Sweden, 13-19 august 1995.