ISCA Archive
http://www.isca-speech.org/archive

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

AUTOMATIC WORD DEMARCATION BASED ON PROSODY

Paul Munteanu, Bertrand Caillaud, Jean-François Serignat, Geneviève Caelen-Haumont

Laboratoire CLIPS/IMAG, CNRS, Université Joseph Fourier, INPG

38041 Grenoble CEDEX 9, France

Tel : +33 4 76 51 45 26, Fax : +33 4 76 44 66 75

Email : `Paul.Munteanu@imag.fr`

INTRODUCTION

This paper presents a work on the acquisition of the prosodic knowledge that will be incorporated in a Word Prosody agent of a distributed speech understanding system (MICRO). The multi-agent architecture of MICRO, based on wholistic analytic double processing, is first described. MICRO uses prosody with a rather new view. This group of agents quickly produces information that will be used by the analytic pathway (acoustic-phonetic analysis, lexical access, syntactic and semantic analysis, ...) as anchor points or for lexical hypotheses filtering or sorting. We discuss the role of the Word Prosody agent in this architecture and the induced requirements for its design. Then, we present some experiments that were made in order to decipher the prosodic encoding of word boundaries and lexical categories.

1 PROSODY IN SPEECH UNDERSTANDING

It is commonly admitted that prosodic characteristics of utterances (such as intonation and temporal patterns) play an important role in the cognitive processes involved in human speech understanding. However, prosodic information is not yet widely integrated in automatic speech understanding. This may be because the powerful methods used for speech recognition do not apply to this kind of information in a straightforward way. Despite these difficulties, speech researchers have reported the use of prosody in speech recognition, mainly for: semantic interpretation improvement [8], isolated word recognition [15], resolution of syntactic and semantic ambiguities [11] [12], micro-prosody for the acoustic phonetic decoding [14].

In most of these approaches, prosodic information is used in the post-processing of the output of some classical speech recognition methods. Little research has been consecrated to the use of the prosody for the improvement of the lexical access, especially for the French language.

In the MICRO project [4], a multi-agent speech understanding system, we propose an alternative approach : a prosodic group of agents realizes the wholistic processing of speech, besides the classical analytic one (from acoustic-phonetic to linguistic analysis). Since prosody reflects lexical, syntactic, semantic and dialog phenomena, this fast processing provides the analytic pathway with information that is used by its agents at different levels. In this paper we will focus on the automatic detection and classification of lexical boundaries that will be used to improve the lexical access of the analytic pathway.

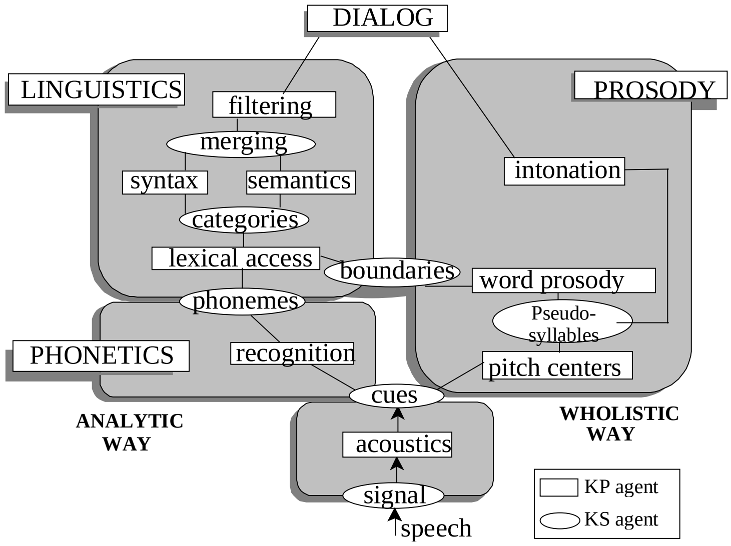

Figure 1 : Software architecture of MICRO

1.1 The Micro architecture

Two main features are necessary for the emergence of a cooperative behavior from a society of agents. On the one hand, each cognitive or computing entity should be independent to build its own opinion on the situation. This independence is a direct consequence of modularity. On the other hand, each module should be aware of the evolution of its environment to adapt its behavior to contextual changes: modules should be interactive.

- **Modularity -** The modular paradigm describes cognition as the emergence of the global activity of a society of modules working in a cooperative way on their own domain of skill.
- **Interactivity -** In the cognitivist paradigm [7], lower level modules blindly work without considering top-down information. Contextual adaptation is then limited to a filtering of ascending hypotheses. On the opposite, interactive theories [9] militate against such a sequential description. In other words, every cognitive module has a direct access to bottom-up hypotheses as well as upper top-down information.

1.1.1 General description of MICRO

Following the functional description below, MICRO is an heterarchic society of agents developed on MAPS, a software environment dedicated to multi-agents structures design [3]. MAPS is based on the distinction between two kinds of agents: Knowledge Servers (KS), which maintain and transmit figurative knowledge and Knowledge Processors (KP), which handle operative knowledge.

MICRO is made of five groups of agents [4] :

- *Acoustics*: This group corresponds to human auditory system processing. It provides the upper level of analysis with acoustic cues of speech signal.
- *Phonetics*: This group handles a stochastic process of speech recognition that provides lexical hypotheses to the linguistic system.
- *Linguistics*: This group aims at building conceptual representations of sentences in order to allow their interpretation. See [1] for detailed description.
- *Prosody*: This group quickly produces information on temporal demarcations of words and syntagmatic groups of words. It also models speech intonation.
- *Dialog*: This group drives dialog strategies in respect with pragmatic information.

1.2 Prosodic agents in MICRO

Considering the Micro architecture described below, prosodic agents have to produce robust knowledge that will be used by the analytic way. In this paper, we will focus on two agents of the prosodic way: the Pitch Centers Detection agent and, especially, the Word Prosody agent.

1.2.1 "Pitch Centers Detection" Agent

It is generally admitted that the prosodic information parameter can be represented by stylized parameters without loosing important information. Nevertheless, there are numerous kinds of methods to produce these stylized parameters. So, we chose to use the pitch center stylization considering our goal. This kind of treatments makes an important reduction of the data. But perception analysis have shown that the information needed for demarcation and understanding of words and sentences are not suppressed by this stylization method [6].

After this treatment, the pitch center agent associates to each pitch center the following prosodic parameters:

- **Pitch**: average value of pitch on the detected pitch center
- **Energy**: same for energy
- **Duration**: distance (in terms of time) between the current and the previous pitch center
- **Pause**: the duration of the pause (if it exists) just before the current pitch center.

These parameters are linearly normalized in [0,1] interval, for each sentence. In order to attenuate the influence of the well-known phenomenon of average pitch decrease along sentences, pitch normalization is realized around the pitch regression line.

1.2.2 "Word Prosody" Agent

This agent is in charge of fast providing information on word boundaries and lexical categories to the analytic way of the speech understanding system. Its task may be defined as the following classification problem:

Input: Pitch, energy, duration and pause calculated for the current pitch center and its two closest neighbors = *12 numerical features* for each pitch center.

Output: Reliable (even not numerous) information on word boundaries and lexical categories. These word demarcation indices may be: beginning of lexical word (BLW), end of lexical word (ELW), internal syllable of a lexical word (ILW), monosyllabic lexical word (LW), beginning of grammatical word (BGW), end of grammatical word (EGW), monosyllabic grammatical word (GW). The Word Prosody agent has to classify the current pitch center in one of these *7 disjoint classes*, thus realizing both the segmentation and the lexical typing of the obtained segments.

This kind of information can be integrated in the speech recognition process in two different manners:

- as anchorage points. These points will constitute the skeleton of the analytic processing. Even if errors occur during analysis, the system can safely restart the treatment after the next anchorage point, if it has been faithfully detected. Therefore, in order to use their results in this manner, prosodic agents have to produce reliable, even if not numerous, information.
- in order to sort or filter sentence hypotheses produced by the lexical access, by giving more importance to the ones that match with the prosodic boundaries. Prosodic information can be used to filter or sort hypotheses depending on the accuracy of information produced.

2 LEARNING OF THE PROSODIC LEXICAL LABELS

To create the "word prosody" agent described before we decided to use machine learning methods. As we said before the task of this agent is a classification task : it associates a set of numerical features to a class indicating the kind of boundary. In order to avoid the inadequacy of a particular learning bias for this classification task, we experimented several methods which use different search spaces and search heuristics. We will present the results of this comparison before the results of the classification itself. The data used for all experiments is a speech corpus of read text (5 sentences), pronounced by 9 different speakers (6 males and 3 females) and already labeled with phonetic annotations.

2.1 Method Comparison

We decides to compare for this task three different kinds of methods : machine learning, neural and statistical algorithms. We also used different algorithms in each field. We only present in this paper one in each field :

- PMBC [5], [10]: our approach, a rule induction algorithm based on successive generalizations of examples
- NOPT [2]: a connectionist algorithm based on conjugate gradient optimization, successfully used in other speech processing tasks. In this experiment we used a 3-layered neural network.
- CGL [13] : gaussian linear classifier, we also used the quadratic classifier but the linear one obtained better results for this task.

The comparison was made using the following protocol for each of the seven tasks (Classification of BLW, ELW, LW, ILW, BGW, EGW, GW). Each algorithm was trained and tested 50 times on a random split of the data. For each task, groups where made with the t-test statistical criterion. For instance, if an algorithm is in the first group this indicates that no other algorithm is better than it with a significant difference.

Figure 2 presents the classification of each algorithm in the first two groups.

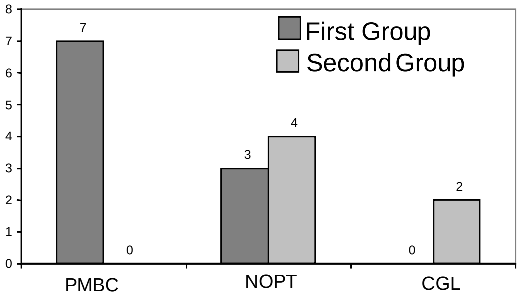

Figure 2 : Results for the method comparison

These results clearly indicate that PMBC (the machine learning algorithm) is better than NOPT and CGL for this kind of task, this advantage is more clear with CGL but is also significant with NOPT. Considering that we used PMBC for the following experiments.

2.2 Classification Results

Results presented below correspond to 50 tests performed on the reference set for each class. We provide two performance indices which are estimated as follows:

$$\text{detection rate} = \frac{P_p}{P} \quad \text{reliability} = \frac{P_p}{P_p + N_p},$$

where P is the number of examples of one pitch center class, N is the number of counter-examples (all examples which do not belong to the learned class), P_p is the numbers of well classified examples, N_p is the numbers of misclassified counter-examples.

As the different classes are oddly distributed in our reference test, the global accuracy is not a suitable parameter for comparing classification performances between classes (our algorithm easily achieves a global accuracy of 90% for BGW and EGW by simply ignoring the less frequent plurisyllabic grammatical words).

Task	Reliability		Detection Rate	
	Mean	±	**Mean**	±
BGW	**92,00**	27,41	**3,65**	15,17
EGW	**94,00**	23,99	**2,90**	14,49
GW	**73,92**	4,89	**52,69**	5,51
BLW	**59,45**	9,76	**22,05**	6,32
ILW	**66,05**	25,30	**12,47**	5,94
ELW	**73,20**	8,55	**46,69**	6,69
LW	**56,74**	19,18	**13,03**	5,69

Table 1 : Classification result for the 7 basic tasks

Task	Reliability		Detection Rate	
	Mean	±	**Mean**	±
LW’	**84,49**	3,58	**70,59**	5,05
GW’	**76,04**	5,74	**56,04**	5,72

Table 2 : Classification results for the derived tasks

Comments on results presented in tables 1 and 2

– Prosodic information seems to be oddly distributed over the lexical categories. Furthermore, by grouping together some of the initially defined categories (ELW+LW=LW’, BGW+GW=GW’) we achieve better performances as shown on table 2. This observation seems to indicate that different lexical classes share common prosodic encoding: monosyllabic lexical words behave like the ends of plurisyllabic words, while the prosody of monosyllabic grammatical words seems rather close of the prosody of the beginnings of plurisyllabic grammatical words.

– The induction algorithms are able to find reliable classification rules only for a rather small proportion of the pitch centers, even for the classes with best scores. This observation seems to indicate that the lexical category and the boundaries of words are not always prosodically marked.

The experimental results are rather encouraging. The most reliable classes (end of lexical words and beginning of grammatical words) cover 68% of the total number of pitch centers. Considering their detection rates, about 45% of all pitch centers are rather reliably classified. Therefore, some of these results may be already used in sorting the lexical hypotheses. In order to estimate their suitability as anchor points, further experiments, combining prosody and lexical access, are necessary.

3 CONCLUSION

The results presented in this paper seem to indicate an uneven distribution of the prosodic information over the pitch center classes and suggest that some groupings of these classes may improve the performance of the classification task. The use of prosodic information as anchor points for the analytic pathway needs further experimentation and, probably, an improvement in reliability. However some of the results we obtained may be already useful to help lexical access of a large vocabulary speech understanding system.

4 BIBLIOGRAPHY

[1] Antoine J.Y., 1994, Coopération syntaxe-sémantique pour la compréhension automatique de la parole, PhD thesis, INPG, Grenoble, France.

[2] Barnard E.; Cole R., 1989, A neural-net training program based on conjugate-gradient optimization, *Technical Report CSE 89-014*, Oregon Graduate Institution, Oregon, USA.

[3] Baujard, 0.; Garbay, C., 1990, A programming environment for distributed expert system design, *Proceedings of ExperSys90*, 27-32.

[4] Caillaud B, 1996, Apprentissage de connaissances prosodiques pour la reconnaissance automatique de la parole, PhD thesis, INPG, Grenoble, France.

[5] Caillaud, B.; Munteanu, P.; Serignat, J.F.; Caelen, J., Prosodic Knowledge Acquisition for Lexical Access Improvement, Int. Journal for the Integrated Study of Artificial Intelligence, Cognitive Science and Applied Epistemology (CC AI), Communication and Cognition, Belgium, to appear.

[6] De Tournemire S., 1994, Recherche d’une stylisation extrême des contours de F0 en vue de leur apprentissage automatique, *Actes des 20-èmes Journées d'Étude sur la Parole*, S.F.A., pp.75-80, June, Tregastel, France.

[7] Fodor J.A., 1983, *The modularity of Mind*, MIT Press, Cambridge, Mass.

[8] Lea, W., 1980, Prosodic Aids to Speech Recognition, *Trends in Speech Recognition*, pp. 166-205, Prentice-Hall Inc., Englewood Cliffs, USA.

[9] McClelland, J.L.; Rumelhart, D.E.; and the PDP Group, 1986, *Parallel Distributed Processing: explorations in the microstructure of cognition*, MIT Press/ Bradford books, Cambridge, USA.

[10] Munteanu, P.; Caillaud, B.; Serignat, J.F., 1994, PMBC-a similarity-based rule induction algorithm, *Technical Report*, Institut de la Communication Parlée, Grenoble, France (available from the authors on demand).

[11] Ostendorf, M.; Wightman, C.W.; Veilleux, N.M., 1993, Parse scoring with prosodic information: an analysis/synthesis approach, *Computer speech & language*, 7(3), pp.193-210.

[12] Price, P.J.; Ostendorf, M.; Shattuck-Huffnagel S.; Fong C., 1991, The Use of Prosody in Syntactic Disambiguation, *JASA*, vol. 90, no. 6, pp 2956-2970.

[13] T.W. Rauber, M.M. Barata, A.S. Steiger-Garção, 1993, "A toolbox for analysis and visualization of sensor data in supervision", actes International Conference on Fault Diagnosis, Toulouse.

[14] Vaissiere, J., 1988, *The use of prosodic parameters in automatic speech recognition*, Recent Advances in Speech Understanding and Dialog Systems, NATO ASI Series, Springer-Verlag, Berlin, Germany.

[15] Waibel, A.,1988, *Prosody and Speech Recognition*, Morgan Kaufman Publishers Inc., San Mateo, USA.