A Model to Computational Speech Understanding

Daniel Nehme Müller, Mozart Lemos de Siqueira, and Philippe O.A. Navaux

The Federal University of Rio Grande do Sul,
Porto Alegre, Rio Grande do Sul, Brazil
{danielnm, mozart, navaux}@inf.ufrgs.br

Abstract. We propose a speech comprehension software architecture to represent the flow of the natural processing of auditory sentences. The computational implementation applies wavelets transforms to speech signal codification and data prosodic extraction, and connectionist models to syntactic parsing and prosodic-semantic mapping.

1 Introduction

This work argues that it is possible to unify several computational systems to represent the speech understanding process. Thus, we propose the SUM, a Speech Understanding Model, based on a neurocognitive model of auditory sentence (section 2). Through SUM, we search a computational representation for speech signal codification, prosody, syntactic and semantic analysis. The SUM is illustrated in the figure 1.

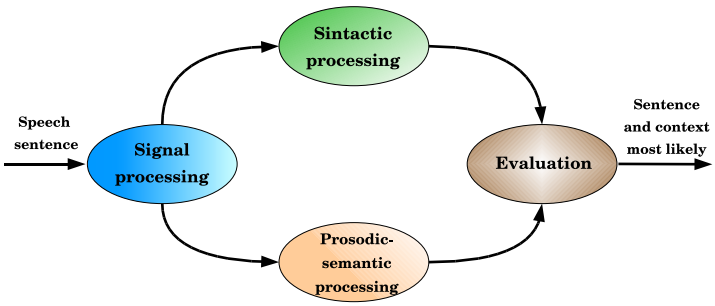

Fig. 1. The Speech Understanding Model - SUM

2 Neurocognitive Model

Angela Friederici [1] proposes a neurocognitive model of auditory sentence processing that identified which parts of the brain were activated at the time, given the different applied tests. She divided the processing of the auditory sentences in four large phases [1][2]. Indeed, the most recent research indicates that the

R. Vieira et al. (Eds.): PROPOR 2006, LNAI 3960, pp. 216–219, 2006.

prosody processing description must be added to the neurocognitive model [3]. In the first phase, it is done an acoustic characteristic extraction and codification. Thus, the prosodic characteristics, defined by the pitch variation, determine the processing segmentation. The linguistic characteristics will be analyzed at the syntactic level by the right hemisphere of the brain during the second phase [2]. The second phase performs the syntactic analysis and it occurs only in the left hemisphere of the brain. The semantic analysis is performed in the third phase and apparently awaits the syntactic analysis output in order to solve interpretation problems, brought about mainly by the words' categories contextualization. In the fourth and last phase the integration among syntax, semantics and prosody, necessary to revisit problems not resolved in the previous phases takes place. The syntax structure correction is necessary when there are lexical terms organization problems [2].

3 The Speech Understanding Model - SUM

From the four described phases in the neurocognitive model, we propose the architecture of SUM, illustrated in the figure 1. In SUM, the first phase extracts the coefficients from speech signal. These coefficients provide the information about the fundamental wave (F0) and they are used in the following phases. The second computational phase is the application of coefficients to realize the syntactic parsing. In the third phase the coefficients are used to semantic contexts definition. The fourth phase receives the analyses from second and third phases outputs. To each analyzed sentence the most likely context is indicated.

In the first phase of the computational model, the signal is processed by the application of wavelet transform. The second computational phase is the application of the wavelet coefficients to generation of temporal registers and parsing trees through the system SARDSRN-RAAM previously developed by Mayberry and Miikkulainen [4]. In the third phase, semantic and prosodic maps are applied using the Self-Organizing Map (SOM) [5]. The fourth computational phase performs the reception and the analysis of the output of the second and third phases. In this phase, the model indicates the most likely written sentence for a given speech sentence. The wavelet transform can be seen as a signal filter, making it possible to build filterbanks through them, and, thus, enabling a multiresolution analysis [6]. In this work we use the multiresolution analysis to speech signal codification. This process was split in phonetic and prosodic approaches. The phonetic approach is obtained from a single decomposition of wavelet coefficients (phonetic coefficients). The prosodic way is extracted from F0 variation (pitch). According to [7], to acquire information on the variations of the F0 speech it is necessary to detect the wavelet maximum points, which correspond to the glottal closure instants (GCI). If the maximum points are obtained, we attain the F0 estimation. The coefficients achieved (prosodic coefficients) will be sent to linguistic parsing system.

The syntactic analysis is allowed by the phonetic codification of words, extracted from wavelet transform, is structuralized through the RAAM, whose

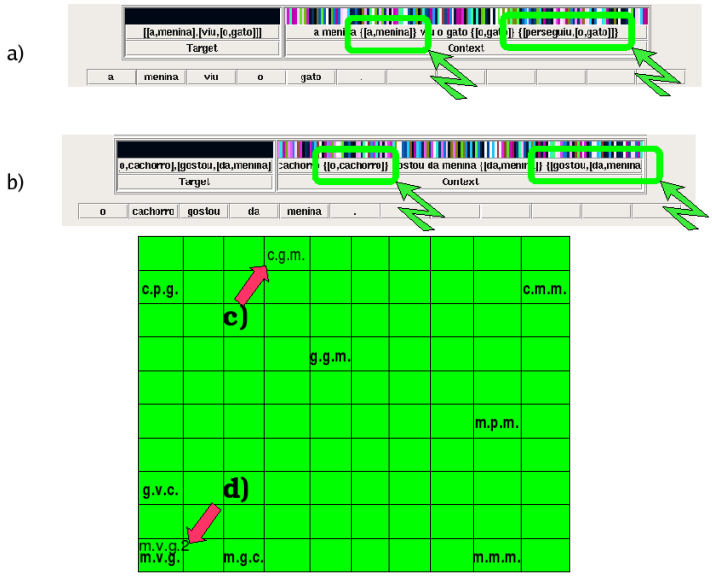

Fig. 2. Sentences recognition in SARDSRN and sentences map: the sentences a) and c) correspond to the c.g.m and b) and d) are the m.v.g.2

activation allows the sequencing of the words in the phrase by the SARDSRN-RAAM. Afterward, the temporal sequence of the component words is initiated, and the pattern presented in the input layer is distributed to the hidden layer and the SARDNET. This net, in turn, also feeds the hidden layer. Parallel to this hidden layer, there is a context layer, characterizing the SRN in the SARDSRN-RAAM. Finally, the output layer generates a pattern sentence that is decoded by the RAAM net. A relevant characteristic of the SARDSRN-RAAM is its great capacity to generate parsing sequences that will allow recognition of multiple parsing trees compressed in RAAM net. The semantic processing is composed by four chained SOM nets. In the first SOM net, the *prosodic map* groups the words according to signals derived from the analysis of variations in the F0 (prosodic coefficients). In the second SOM net, the *phonetic map* is obtained from the relations of phonetic characteristics of each word, extracted by wavelet transform. The net that forms the *semantic map* uses the output information on the activated neuron in the phonetic map plus the activation in the prosodic map. Finally, the last map is responsible for grouping sentences that are informed by the user. The composition of the output of semantic map for each word is the input of the *sentences map*. The recognition of speech patterns is performed by the sentences map, which indicates the most likely sentence. After syntactic and semantic processing, the systems' output are evaluated. The SARDSRN-RAAM system indicates an error rate (¿=0.5) and the semantic maps system points to the winner neuron in the sentences map. If the syntactic processing has a high rate, we can do an approximation by semantic processing, and vice versa.

As illustration of the functionality of the system, two speech sentences not trained had been elaborated: m.v.g.2 - *a menina viu o gato* (the girl saw the cat) and c.g.m. - *o cachorro gostou da menina* (the dog liked the girl). The sentence m.v.g.2, to be presented to the syntax subsystem, resulted in the trained sentence *a menina perseguiu o gato* (the girl chased the cat) as an answer, thus pointing the error of the recognition (fig. 2a). In the sentences map, the identical positioning to the trained sentence m.v.g. - *o menino viu o gato* (the boy saw the cat) was obtained (fig. 2d). In the sentence c.g.m, the great distance (¿2) from trained patterns in the sentence map indicates failure in recognition (fig. 2c). On the other hand, the syntactic system returned the exact written sentence, although it had *not* been trained in it (fig. 2b). These two examples mean that the first sentence corresponded to sentence that had more phonetic representations in common in the trained construction, and in the second sentence the system did not guarantee the semantic recognition, but would identify in syntactic system.

4 Conclusion

The resultant codification demonstrates that there is an interface between existent linguistic parsing connexionists systems to text analysis and the speech. This opens a new method to implementation of systems for written language with speech as input. The use of artificial neural nets in the syntactic and prosodic-semantic processing was presented as a facilitator in the language modeling process. The computational prototype, that demonstrates the processing of the SUM, resulted in a system of analysis by compensation. Therefore, when the syntactic analysis does not offer a good reliable level, it is possible to evaluate prosodic-semantic analysis, such as in human speech understanding.

References

1. Angela D. Friederici, "Towards a neural basis of auditory sentence processing," *Trends in Cognitive Sciences*, vol. 6, pp. 78–84, 2002.
2. Angela D. Friederici and Kai Alter, "Lateralization of auditory language functions: A dynamic dual pathway model," *Brain and Language*, vol. 89, pp. 267–276, 2004.
3. Korinna Eckstein and Angela D. Friederici, "Late interaction of syntactic and prosodic processes in sentence comprehension as revealed by erps," *Cognitive Brain Research*, vol. 25, pp. 130–143, 2005.
4. M. R. Mayberry III and Risto Miikkulainen, "SARDSRN: a neural network shift-reduce parser," in *Proceedings of IJCAI-99*, pp. 820–825. Kaufmann, 1999.
5. Teuvo Kohonen, *Self-Organization and Associative Memory*, Springer-Verlag, 1984.
6. S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pat. Anal. Mach. Intell.*, vol. 11, pp. 674–693, July 1989.
7. S. Kadambe and G.F. Boudreaux-Bartels, "A comparison of a wavelet transform event detection pitch detector with classical pitch detectors," *Twenty-Fourth Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 1073–1078, 1990.