Multigranular Scale Speech Recognizers: Technological and Cognitive View

Francesco Cutugno, Gianpaolo Coro, and Massimo Petrillo

Department of Physics, University Federico II, Naples, Italy
{cutugno, coro, massimo.petrillo}@na.infn.it

Abstract. We propose a Multigranular Automatic Speech Recognizer. The hypothesis is that speech signal contains information distributed on more different time scales. Many works from various scientific fields ranging from neurobiology to speech technologies, seem to concord on this assumption. In a broad sense, it seems that speech recognition in human is optimal because of a partial parallelization process according to which the left-to-right stream of speech is captured in a multilevel grid in which several linguistic analyses take place contemporarily. Our investigation aims, in this view, to apply these new ideas to the project of more robust and efficient recognizers.

1 Introduction

Many works available from various scientific fields ranging from neurobiology to experimental phonetics seem to concord on the idea that speech signal contains information distributed on more different time scales and that, in order to process it properly, it is necessary that more parallel cognitive functions operate a chunking on the unfolding of the information over time. It seems that speech recognition in human can success because of a partial parallelization process according to which the left-to-right stream of speech is captured in a multilevel grid in which several linguistic analyses take place contemporarily. Evidence of parallelized speech processing can be seen in many authors like Poeppel [1]. In recent speech perception theories (Hawkins, Smith [2]) the existence of a multimodal sensory experience is stated being processed and transformed into different type of linguistic and non linguistic knowledge. These ideas have rapidly influenced many researchers involved in ASR (Automatic Speech Recognition) projects (Wu [3], Chang [4], Greenberg [5]). Newer ideas like "syllabic pre-segmentation", "word n-gram statistical combination", "parallel and multiscale speech coding" have been introduced in speech processing (similar concepts were also present, in Erman et al. [6]).

2 Multigranular Automatic Recognition

Modern approaches to Automatic Speech Recognition are typically classified on the base of different identification of the so called "Base Unit" of Speech (the minimal form of acoustic and linguistic information around which human speech recognition is organized). Supported by perceptive experimental results and application efficiency,

S. Bandini and S. Manzoni (Eds.): AI*IA 2005, LNAI 3673, pp. 327 – 330, 2005.

the most common approach is the "phonetic" one: it is hypothesized that a sequence of phones is sufficient to recognize a word. Hereby we mean by phone one or more acoustic instances of the abstract classes of speech sound known as "phonemes".

A possible alternative is to identify the Base Unit with the syllable. Even in this case a number of perceptive experiments give support [7]. We refer here a syllable as a group of phones strongly connected each other by dynamic constraints and by temporal evolution of the articulatory apparatus.

In the rest of this article we propose a third approach, following the idea of a "multigranular" recognizer and refusing to use a single type of Base Unit.We will attempt a first design of a framework in which two o more linguistic units, directly connected to different time scaled processes, could generate a multilevel lattice taking into account all the information available in speech signal during the speech recognition process: phones and syllable could constitute the first two levels of analysis directly followed by words and other events. Collateral to the theoretical discussion on the Base Unit, is the problem of the choice of the "technical" instruments for the recognition. The statistical approach is the most used, but some alternatives must be evaluated: the redundancy, present in the acoustic signals, has to be "swindled" in order to reduce complexity, and, furthermore, we also want control the speech recognition process at every step.

Experimental evidence brings us to think that systems like Hidden Markov Models (HMMs) can extract recognition-useful information with less variables than other systems, while many authors prefer a "hybrid approach", that makes use of a Neural Network for the recognition of single linguistic units, followed by a lexicon for word decoding.

Assume now we have a stochastic recognizer (HMM, Neural Network etc.) for every linguistic unit (phone, syllable, word etc.) and let's see how we can melt together the single "grain" recognitions, in order to get the most probable spoken word.

We could build a "lattice" linking linguistic units of the same type, that would represent any possible pronunciation of the words in a reference dictionary.

For example we could think about a "phone lattice": every node will represent a single phone and every arc the probability of moving from a phone to the successive (this is the "classic approach" lexicon). By means of the sequence of recognitions, performed by the previous statistical system, we will get an "optimal" walk through the lattice and so the most probable phrase or word. At the same way, we can imagine a syllable based lattice or a word based lattice.

Arc weights are chosen on the basis of a statistical analysis on a reference corpus and correspond to the succession frequency between two linguistic units.

The number of arcs and their topology in the lattice are clearly different in this three models: while in the phones lattice each node has at least, in principle, a connection with any other unit in the set, not every syllable (or word) can be put before any other.

Statistics on English showed that only few syllable from all possible ones are most used during a natural speech conversation [8], so this should be also the case in Italian, were we know bi-syllable as most used words. This kind of analysis could result in a pruning of the possible combination of syllables or words and a loss of complexity.

From a theoretical point of view a multigranular recognizer should take into account all the three lattices, acting on the basis of their behaviour. The lattices should also be able to communicate each other. A first, rough, idea could be that all the three lattices acted in parallel with the decision basing on the "best scoring" lattice.

This approach is very expensive, in terms of computational complexity, so it's better finding alternative methods.

A possible choice, that we propose in this paper, sees that the system operating always in a single level (starting for example from the less complex) and, only depending upon heuristic evaluations, foresees the passage to another level until another change or the end of the recognition (Fig. 1) is reached.

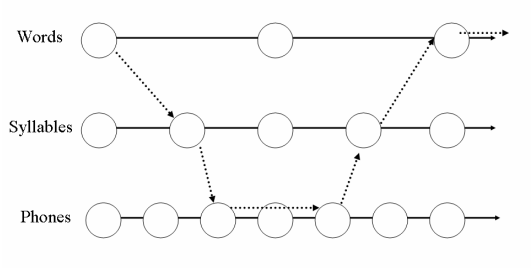

Fig. 1. Recognition process of an ideal Multigranular Recognizer

The heuristic evaluation could be based upon parameters as noise level, or complexity of the level lattice, letting the system to rise or fall also in a "multigranular" scale, according to the belief, explained above, that different granular-levels can vary lattice complexity, and that there are more convenient levels, given a particular situation, that let us recognize a word without the classification of each single phone or syllable. Think about a situation in which, based on the knowledge retrieved from a corpus, a grammar tells us that there is a particular word, into the vocabulary, that starts with the yet recognized phones or syllable and that frequently follows the previously recognized words. On the basis of a heuristic evaluation, the system could infer the correct word without exploring the not-yet recognized phones or syllables, as usually the Viterbi algorithm does in this type of lattices.

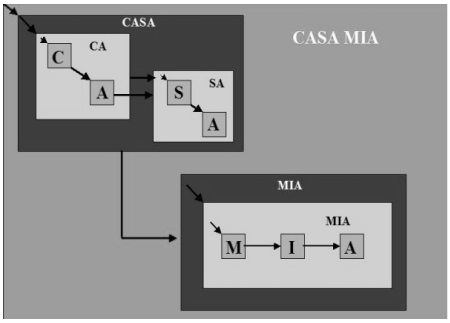

Fig. 2. Statechart Multigranular ASR

The idea we propose here involves the use of a special language: the Statechart [9]. Statechart extends the classical notion of Finite Automata by means of the following concepts: **Hierarchy**, **Concurrency** and **Transmitted Communication** Every state of an Automata is allowed to include other finite automata (**OR** States) and more automata can act in parallel (**AND** states) communicating with the exchange of messages . We want to stress indeed that this is not the only possible implementation of the model, alternatives are still in study. Starting from the statechart concept we show an implementation scheme of a multigranular ASR. Hierachy is central in the model and this is achieved by nested OR states (Fig. 2). This model exploits the varying complexity and performances potentials of several recognizers, in order to take advantage from the best combination. A big weight is given to the heuristic function that has to choose the way, and hierarchy allows a better management of lattices than the parallel model.

3 Discussion

The necessity of a multigranular model derives from a gap between human and machine spontaneous speech recognition. Modern ASR are not able to emulate human auditory system, moreover they are usually based on a single linguistic unit and completely separate from the perceptual behaviour. Models yet developed in this way let us think that the whole system could result in a more robust to interferences one (Wu [3]). We have proposed a new theoretical model, though also an idea of implementation have been discussed, towards a speech multigranular processing.

References

1. Poeppel, D.: The Analysis of Speech in Different Temporal Integration Windows: Cerebral Lateralization as 'Asymmetric Sampling in Time'. Speech Communication 41 (2003) 245-255
2. Hawkins, S., Smith, R.: Polysp: a Polysystemic, Phonetically-Rich Approach to Speech Understanding. Rivista di Linguistica, (2001) 99-189
3. Wu., S. L.: Incorporating Information from Syllable-Length Time Scales into Automatic Speech Recognition. ICSI PhD. Thesis (1998)
4. Chang, S.: A Syllable, Articulatory-Feature and Stress-Accent model of Speech Recognition. PhD. Dissertation, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley (2002)
5. Greenberg, S.: Understanding Speech Understanding: Towards a Unified Theory of Speech Perception. ESCA Workshop on Auditory Basis of Speech Perception, (1996) 1-8
6. Erman L.D., Hayes-Roth F., Lesser V.R., Reddy R. The Hearsay-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty. ACM Computing Surveys 12(2) (1980)
7. Dominic W. Massaro. Preperceptual images, processing time and perceptual units in auditory perception. Psychological Review, 79(2) (1972) 124-145
8. Greenberg, S.: On the Origins of Speech Intelligibility. ESCA Workshop for Robust Speech Recognition for Unknown Communication Channels, (1997) 23-32
9. Maggiolo-Schettini, A., Peron, A., Tini, S.: A Comparison of Step-Semantics of Statecharts. Theoretical Computer Science (2003) 465-498