# A Step Forward in Multi-granular Automatic Speech Recognition

Gianpaolo Coro
*coro@na.infn.it*

November 20, 2007

This thesis is a first effort to make a step further in the understanding of speech recognition. The starting point of the ideas here presented goes back to the early language scientific theories, which have been followed in time, by a set of psychoacoustic experiments, models, and technical realization attempts.
An hypothesis will be assumed, which will be called *multi-granular* as the next discussions will better define. A speech signal contains information distributed on different time scales, and humans are able to catch it all. Furthermore, other knowledge sources are used, which are not necessarily linked to the signal, but can even come from semantics or pragmatics.
This work focuses on what can be extracted from the signal, without investigating other knowledge sources.
Human auditory system needs that more parallel cognitive functions operate a *chunking* on the unfolding of the information over time, to catch all information coming from the signal. Humans seem to perform speech recognition successfully also because of a partial parallelization process. The left-to-right speech stream is captured in a multilevel grid in which several linguistic analyses take place simultaneously.
An example of realization for a multi-granular automatic speech recognizer is here presented. Dynamics coming from the signal, which are segmental or supersegmental in nature, are catched in a single model which tries to take the best of them, in order to improve system performances. Each analysis level, set up on a certain scale, is a *grain*. The whole system is so defined as a *grain-set*, which, in this experiment, will only come from signal characteristics. The elements of the set are correlated and cooperate during the speech processing.
Some problems arise in the definition of such a system. Firstly the several

sources of information have to be identified and secondly they have to be coded and modeled in some way. A further problem is in the time span of the events, which are not synchronous to each other.
Despite of its essential usefulness in people interaction, speech recognition is one of the most difficult human feature to model with an automatic system. This work is meant to go towards a better understanding of the problems lying in the gap between human and machines on speech recognition. As usual, the main aim is to *simulate* and not to *emulate* human behaviour.

The recognizer here developed, has been compared to a standard model, with an improvement of absolute *17%* in the task of number recognition in the range *0-999,999*. The results come out from the experiment open several discussions on speech events coding, on the behaviour of the machine learning models employed, and on further developments of the ideas.

# Contents

# List of Algorithms



# List of Tables

# List of Figures

# 1 Introduction

Spoken language recognition in human beings is a natural, robust and performant feature. This means that it is able to function correctly even in uncomfortable situations, in presence of background noise or reverber. According to many perceptual experiments, human speech recognition system acts computations, filters and adaptations to speakers it deals with, and is then able to transform a vocal signal in a succession of words to which associate an interpretation.
The details of how this is achieved goes beyond our current knowledge and, besides many theoretical and experimental models, there are lots of further aspects to discover.
Despite the many problems encountered during the development of scientific models, the science of automatic speech recognition has evolved and achieved success in creating artificial methods which can simulate some human behaviour, but the goal is the complete understanding and catching of such topic and is still very far. Automatic Speech Recognizers (**ASR**) are going to work well enough to meet market requirements and many of them are employed as dictation systems. The design and building issues for such artificial systems presents difficulties because of complexity problems, as for the real-time requests of functioning and robustness requirements, as they have to function almost "everywhere and for everyone". Factors yet studied in linguistics, as speech variability form person to person, environmental noise, words confusion, coarticulation effects, are some of the aspects influencing performances and efficiency of ASRs, increasing the gap to human system simulation.
Modern dictation systems, in which the user can speak into a microphone and phrases are automatically written on an electronic page, can reach high performances, but they need to be trained on a precise speaker, and no effectiveness is granted on another voice. Such systems are generally called *speaker dependent*, and are not involved in the present discussion.
This thesis is about *speaker independent* systems, whose performances are calculated on many speakers, with different genders and dialect inflections. The aim is surely more challenging respect to speaker dependent systems, because it has to take into account the high speech variability between humans.

In ASR building, scientists have always started from hypotheses about human communication. They also state that there exist an atomic unit, which is the *Base Unit* around which human speech recognition is centred.

Many aspects of speech signals have been formalized and lots of units have been investigated, such as syllables, phonemes and so on. Generally a stochastic system is employed, which finds to associate a unit to a piece of signal, while another process decodes the spoken phrase by assembling the best sequence of units. The whole process is governed by a grammar, which contains the probabilities of concatenation between units.
The problem of the choice of the best Base Unit is fascinating, and has risen many debates in psychoacoustics as well as in informatics. The main problems are in the balance between the possibility to represent such units, formalizing their characteristics, and their robustness to environmental and speakers variations.
The choice of a domain of application[1] for an experimental system is also critic. It is not wise to face the problem of building a novel model, by running it directly on a natural language application. The right subset of a language has to be chosen, which must present many of the problems which can be found in large vocabulary applications, and a fast implementation which focuses on the model, rather than on performances. In this experiment a *corpus* has been collected, which is a set of recordings with a related hand transcription, with indications about units present in the speech signal, along with their position. The corpus is necessary because ASRs are machine learning systems that need to be trained on prepared examples. Also performances are calculated on a corpus basis, because hand transcriptions are used as references to test the truth of the automatic productions.
The above explained is the classic approach, which does not involve other aspects intervening in human recognition. Some theories and experiments have highlighted that also long-time span processing is important and that events like rhythm and accent are involved. All these features will be investigated in this thesis on the strand of multi-granular approach, which integrates all those information sources in a complex model for an ASR. As explained in the next sections, such an idea has evolved during last years, bypassing initial problems of complexity and scalability, also thanks to technology evolution.

[1] *Domain of application* means the semantic and syntactic area corresponding to a defined environment for the application of an ASR. For example medical refertation or help desk services put several restrictions on language and vocabulary.

## 1.1 The Idea

At the beginning of the 20th century, Ferdinand de Saussure, the father of modern linguistics, in his *Lectures on General Linguistics* [18] stated that, when facing the problem of describing a "living" language, there is a unique rational method to be used, which consists in

- Collecting the set of elementary sounds on direct observation.
- Putting apart the system of signs which serve to represent, imperfectly, the sounds.

He said that many grammar scholars of his time, were still fond of the erroneous methodology of researching *how* each letter of a language, they wanted to describe, was pronounced. In this way it is not possible to completely represent the phonological system of an idiom. Such set of features has to be distinguished from written language. Speech has its own life and a separation has to be made between spoken language and its representation, the written language, which has a slower evolution and sometimes adapts itself to speech. Furthermore de Saussure stated that, during the reading process, two behaviours can take place. A word which is new or unknown, is read letter by letter, while an usual word is catched in "one shot", independently from the letters which it is made of. The discussion about written and spoken language is concluded by stating that linguists have to limit themselves to $<<$desire that usual writing is free from its biggest absurdities, because, if in language teaching a phonological alphabet can give services, its use has no to be generalized$>>$.
The theory continues with a treatment of spoken language and its role respect to written language, which is *somewhat confusing* because it is only a sort of image of its spoken counterpart. In the end, the author states that speech follows other kinds of grammars, structures and evolutions.
The considerations by de Saussure are important in two ways

- They make us understand that speech has to be considered separately from writing. Also words, syllables and grammars representations have to be changed and to be different from written language ones.
- There exist some forms of *grains*, which are visible in writing, but also in speech. When a word is new, we have to investigate at the fine grain of letters (or phonemes in speech) details. If it is a usual word, we make use

of large grains, which catch the entire word without going to explore finer and, above all, slower details.

These two assumptions are the very starting point of this thesis, which so begins from early studies about language. De Saussure was critic about its colleagues, which focused only on fine phonetic details.
The underlying idea is that many dynamics exist in a speech signal, which can be extracted and used to reach a robust recognition. What is proposed is not something new, but it has always been in speech research, in all the environments. As said above, an evidence can be found in de Saussure's theories, but also in psychoacoustics, linguistics and engineering.
The here presented model will give an organization to the concept of multi-granular speech recognition, by introducing the ideas of speech grains and merging them together.

In the next sections, an introduction to some basic linguistic definitions is firstly depicted, in order to let the reader understand which are the speech units we will refer to. Then psychoacoustic models will be introduced, in order to give evidence about intuitions of multi-granular ideas, in environments which are above informatics. Finally, an overview of the evolution in time of the concept of multi-granular recognition follows, where it will be clear that the theory has always been present since early ASRs models.

## 1.2 Speech Units

### 1.2.1 The Spoken Language

Communication by means of spoken language can be seen like an interaction between an emitter (the speaker) and a receiver (the listener) which is achieved by means of a physical support (air). Aside from a technical point of view, several definitions can be found, which cannot synthesize the complexity contained in the term *Speech*.

Speech is originated by the intention to communicate an idea, and the central nervous system, by means of muscles, is able to transform it in act and to transfer it to a listener. This one captures the signal, under the form of air pressure variations, in the hearing system, processes it and converts it in neural stimuli, which are then interpreted by the central nervous system. The speaker constantly controls production organs, on the basis of acoustic signal hearing [30].

When words are combined in sequences of spoken tokens, the pronunciation of the single segments involved can be subject to changes. Speed and rhythm can be responsible of low volume, deletions, insertions or complete modification of their common characteristics.

Words can be decomposed in syllables, which can assume strong (**stressed**) or weak (**unstressed**) forms. Words which can represent grammar relations in a language, are particularly subjected to these alterations. In fast speech some sounds can be deleted or changed, in spoken Italian this happens when a final vowel meets an initial vowel, and both are not stressed. E.g. the italian phrase *"è un vero amico"* (*he is a good friend*) can sound like *"ènveroamico"*. Other times, between two words a sound can be introduced, for example in presence of disfluences or of false starts. This is the case of embarrassing or uncertainty situations: e.g. *ca-can I invite you?*.

All these factors contribute to make spoken speech a separated world respect to written language. Some problems can be studied and rules can be found to catch them, but it is not always possible. A distinction has to be made between phonological aspects, which are governed by rules, and phonetic aspects, as those illustrated above, which could correspond to unsolvable problems. The hypothesis on which automatic speech recognition is based, is that those phenomena can be studied by means of stochastic methods.

Spontaneous speech presents situations which are very difficult to solve, because of the whole variability spectrum. A good starting point could be the study of

human auditory system, in order to understand all those external factors which act in human listening process.
In the aim to understand which are the entities involved in speech recognition, a discussion about the concept of base unit has to be introduced. This will treat the problem of the existence of a speech unit around which all human speech recognition is based. Such a unit could so be fundamental for an ASR.

### 1.2.2 The Base Unit of Speech

One of the fundamental hypotheses which support the possibility of an automatic speech recognition is that it is possible to collect a sequential type of information in time, and that, when a sufficient amount of is reached, it is processed. A buffer must exist, whose length is related to the concept of Speech Base Unit and to all the information which, at short or long range, contribute to the recognition [39].
Speech Base Unit can be defined as the minimal form of acoustic information around which the most part of human spoken language recognition is organized. Linguists' and psycholinguists' general opinion about this argument is not really clear. Some trends follow the idea that the entity does really exist and has few, distinct, manifestations. Other trends argue that is not possible to identify a single unit, but the manifestation of such a concept has to be searched among combination of atomic units.
All these studies have finally demonstrated that units perception is highly dependent on the context. The recognition process is a real complex analysis involving more than a single scale unit. This thesis will present a model for such idea, which will be furtherly discussed in chapter 6.
Besides its existence in human processes, it could be very useful for an ASR, provided that it is small enough to represent a good variety of manifestations and that it is computationally efficient. Syllables, for example, can include many speech phenomena, and are able to build up all the words of a language. Modelling such phenomena can then result in high performances. Unfortunately this is not the case, because syllables are difficult to formalize in terms of signal characteristics, as well as prosodic units. On the other side, phonemes can present quite well formalizable characteristics, but cannot contain important aspects which are essential to recognition robustness.
In the following sections, the main speech entities will be defined, with a look

also to the possibility of a formal representation.

### 1.2.3 Phonemes and Phones

As de Saussure stated in his theory, in the attempt to formalize a spoken language, the writing can be a good starting point, even if it has to be abandoned when going into the deep details of the study. It could be thought that alphabet letters are the fundamental bricks by means of which also spoken phrases are built up [66]. Alphabetic systems are born in the aim to graphically reproduce the uttered elementary sounds. Unfortunately there exist many languages which use the same alphabet, but associate different pronunciations to the same symbols.

**Example 1**: the French word *chic* and the English word *cheap* present the same initial sequence of characters but different pronunciations.

**Example 2**: the French word *chic* and the English word *ship* have the same initial sounds, but different transcriptions.

In order to avoid those problems in language research, linguists have invented a transcription system of uttered elementary sounds. The elements of this alphabets are called *phonemes*, which so represent classes of sounds and are independent from the language.

Words like the English *can* and the Italian *casa* will have a phonetic transcription which begins with the same symbol */k/*.

The most diffuse phonetic alphabet is the International Phonetic Alphabet (**IPA**).

The instances of the abstract classes which consitute the set of phonemes, are called *phones*. E.g. the symbol */t/* indicates a phonemes which represents all the utterable *t* sounds. Such effective sounds are the phones. In speech production, the pronunciations of the phonemes vary from person to person and from word to word. A sequence of phones in a stressed syllable is different from the same sequence when not in presence of stress, even if the reference phonemes are always the same.

In a more formal framework, a phonetic transcription is an operation consisting in representing the phonetic form of a word (or text) in written language.

**Example** : phonetic transcription of a word

*Pietro -> ['p j e t r o]*

The accent is generally placed before the stressed syllable.

### 1.2.4 Syllables

**Syllable:** a phonetic unit constituted by one or more sounds, pronounced with the same emission of voice. It can be formed by a vowel, a diphthong, alone or accompanied by one or more consonants [23].
**Syllable:** a speech unit for which there is not a satisfying definition [37].
**About Syllables:** There are not common borders at which syllables unify, but each is separated and distinct from the other (Aristotel, Categories)

Syllables have been described as pushes of respiration muscles, peaks of sonority, energy impulses, necessary units in mental organization and in speech production, a group of movements in a vocal signal and a base unit of speech.
An adequate definition of syllable does not exist, despite the long discussion about their role in human speech recognition, . From the point of view of an automatic system, based on them, there is a need for such a definition.
Humans seem to possess an intuitive concept of syllable. And that's why also a "non expert" person is able to divide words in syllables, even if they are not always able to say which rule they have used. Its a common opinion that a syllable is constructed around a nucleus, which is the most intense and always present part. Many syllables start with a first part, called onset, with increasing energy, and terminate with a descending one, called coda.

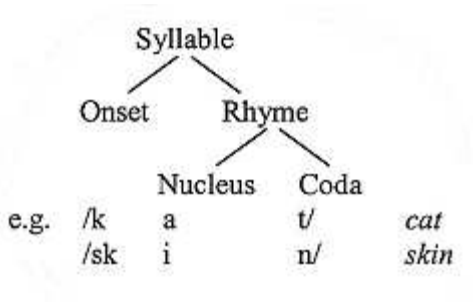



Figure 1: Syllable Structure.

Even if this could seem a good definition of syllables segments, automatic subdivision of words still finds several problem, and the best performance is around *10%* of error rate in placing syllabic markers. This is due to the high variability

of spontaneous speech and to speech rate, which does not let the automatic system to recognize the correct energy islands.
From an abstract point of view, a syllable has to necessarily contain groups of phones and evident acoustic manifestations. Sometimes such phenomena are not detected because people can hear syllables even when they have not been really uttered. This event is called *mirage* and is present in fast speech situations.
From these considerations, researchers have deduced that syllable is a perceptive entity rather than a linguistic one.
Many efforts have been made in representing syllables in terms of acoustic tracts. The overview in the next chapter illustrates some of them. One of the best representation has been tried by Greenberg, with the Modulation Spectrogram [36], followed by some experiments in ASR [17].
As said before, from the point of view of the methodology of modern recognition systems, the ideal base unit of speech should be large enough to incorporate the most part of phonological effects, for example co-articulation and prosodic correlations between phonemes, and it should also have stable and well defined boundaries. If a syllable had a formal definition, it so would result in outstanding performances. Fujimura [21] proposed a work where he depicted the advantages in using syllable rather than acoustic components in speech recognition. He also proposed theories about prosodic structure interpretation in terms of syllabic tracts and obtained interesting results in speech synthesis by means of such units.
The lack for a well defined and accepted definition of syllable, is the main reason for the phonemes to be used in commercial and more diffuse ASRs. Moreover a syllable oriented system is usually strictly linked to an automatic syllabic segmenter or to poorly discriminative features. Finally, speech rate and variability are particularly affecting for syllables, because entire pieces of that structures (onsets or codas) can be deleted.

### 1.2.5 Prosody

Prosody is the part of linguistics which studies the set of phenomena which superpose o accompany the primary articulation of sounds. It is meant to be the set of melodic and rhythmic characteristics of speech. Prosody has been studied deeply as an important source of knowledge for speech understanding. In last years literature, lots of works have aimed to find a model for such phenomenon.

These studies have emphasized the fact that prosody is slightly related to the segmental composition of the vocal signal. It is a *super-segmental* aspect, because it contains information which goes beyond the fine phonetic details.
From an acoustic point of view, the term prosody means intensity, duration, intonation and spectral profile of an utterance [15].
These characteristics let us disambiguate the meaning of some phrases.
**Example** : the following phrases

*Mercy impossible, kill.*

*Mercy, impossible kill.*

differ only in the intonation and distribution of pauses, but the meaning changes strongly.
Prosody can be defined formally by means of several acoustic characteristics, which are not able to describe it completely. Some of these will be largely described in the next chapter.

## 1.3 Multigranular Models in Psychoacoustics

In this section a brief overview is presented about the intuition of multi-granular ideas in psychoacoustics in recent years.
During the search for the Base Unit (**BU**) of Speech in linguistics, experiments have been carried out about establishing if the atomic entity of speech recognition was the phoneme or the syllable[39]. As said in the previous sections, the opinions of the scholars are not coherent, but they almost agree on the fact that such a unit does really exist. The discussion about the identity of such a unit has not come to an end, someone identifies it with the phoneme, other ones with the syllable, but there is another trend supporting the assumption that the Base Unit of Speech has to be found in a combination of many units, segmental and non-segmental in nature, so that it is not an *atomic* phenomenon.
Units like syllables are able to incorporate speech phenomena like co-articulation between phones and other long span features, but the definition of syllable is difficult to catch even in linguistic environment. Researchers have so addressed other units, related to phonemes or syllables, such as diphones, triphone or half-syllables, in order to take a single unit with the characteristics of the searched base unit.

Nygaard et al. [46] have stated that there is not a single BU for all the situations, even if human perceptive system uses few organizing entities. This is a new direction in the research on speech segmentation and identification, and the multi-granular assumption here presented is very close to that.
The most used experiments for base unit detection are the *Monitoring* ones. The main assumption here is that there is a correlation between how fast a human subject is able to recognize and respond to an acoustic stimulus, and how fundamental is the recognition unit, on which the experiment focuses.
Experiments can have many realizations, but in most cases a person is asked to react as fast as possible to the perception of speech signal portions of the length of a phoneme or a syllable. The researchers assume that the correlation between the chosen unit and the reaction time is simple to identify [16].
Experimental results can be divide in three classes:

- The ones who calculate better reaction times for the syllables [39].
- The ones who identify the phoneme as the base unit [2].
- The ones who declare that such results are not significant, because only a single experimental paradigm has been used. Maybe there are more sub-lexical units which are basilar for human speech recognition [35].

Another thread about multi-granularity can be found in neuroscience. Poeppel's research [47] deals with parallel processing of speech. He states that speech signals contain information on different time scales, which are processed bilaterally in superior temporal cortex.
Starting from the above overview, a perceptive model can be described, in which the concept of multiple levels of analysis, is explained and inserted into the general framework of human speech recognition mechanisms. Hawkins et al. [55], in last years, have presented and pursued *Polysp* (POLYsystemic SPeech Understanding), a general framework by which $<<$episodic multimodal sensory experience of speech can be simultaneously processed into different types of linguistic and non-linguistic knowledge at a variety of levels of abstraction$>>$. The main aim is the understanding of the processes which govern the interaction with another person, rather than building a $<<$complete description of a given utterance at successive, obligatory stages of formal linguistic analysis$>>$.

Polysp focuses on how meaning is understood from spoken utterances, without going to only analyze the fine phonetic details of a speech signal.
Hawkins' et al. work explores the contribution of phonetic knowledge to how we understand words, and some implications for what makes a plausible model of spoken word understanding. They show certain types of fine phonetic detail systematically reflect not just the phonemic content, but the wider phonological and grammatical structure of the message and, while some systematic differences in phonetic fine detail are relatively localised in the speech signal, others stretch over several syllables, and that both types can make speech easier to understand. They state that one consequence of focusing only on fine phonetic details in models of spoken word recognition and understanding, is that other processes and stages of analysis may be given inappropriate emphasis, and that this has happened in models which adopt the convenient fiction that the phoneme is the basic input unit to the lexicon. In consequence, no current phonetic or psycholinguistic theory accounts satisfactorily for how normal connected speech is understood.
The approach in Polysp starts from experiments where they explore characteristics of human listeners, that may define the way they make sense of richly informative sensory signals. Amongst these, they emphasize various forms of learning, and some current neuropsychological views about the nature of memory and the organisation of mental categories, both linguistic and non-linguistic. Hawkins declares that $<<$phonetic categories are like all other mental categories: self-organising (emerging from the distribution of incoming sensory information in combination with pre-existing relevant knowledge), multimodal and distributed within the brain, dynamic, and context-sensitive (or relational) and therefore plastic, or labile$>>$.
So, the model main assumption is that the phoneme has dominated thinking in both speech science and psycholinguistic research on spoken word recognition, at the expense of other types of phonological and grammatical structures. Much of the systematic variation in speech that indicates linguistic structure has been ignored. Short-domain spectral temporal events that relate most directly to phoneme identity have dominated perceptual research in speech science, together with a tendency to separate segmental and prosodic information in thinking and in research. Partly for practical reasons, this idea has either been adopted in many computational models of spoken word recognition and lexical access, or it has strongly influenced them.
The traditional view that speech is understood by being organised into *inde-*

*pendent* prosodic and segmental abstract units is rejected, while they suggest that the perceptual correlates of linguistic units are typically complex, often spread over relatively long sections of the signal and simultaneously contribute to more than one linguistic unit. They state that this complexity is a crucial determinant of how we understand speech.
In Hawkins' model, the multi-granular idea is meant to be the basis of speech understanding rather than of simple utterances recognition. They regard each phonetic segment as best described in terms of all of its structural properties, rather than solely or mainly in traditional phonetic terms. Long-domain segmental information, instead, is defined in terms of time and syllables. Such information is defined as a perceptual information extending for at least a syllable, or, somewhat arbitrarily, for about 100 ms or more.
As for the realization of the model they suggest, some hints and guideline are given, while only partial results are reported. The rest of the model will be explored in European projects, as for example S2S [51].
What is defined as multi-granular, in the present thesis, is defined *poly-systemic* in Hawkins et al. [55] work, in that language is seen as a set of interacting systems.

## 1.4 Multigranular Models in the 80s

The 80s have been pioneer years for ASRs, because firsts attempts were made about improving systems performances, in an environment where technology could not support ASR complexity.
Above all the attempts towards the building of systems to be employed in common applications, there are few examples of people who tried to understand if such models could achieve human system simulation. The most famous general framework, the Hearsay II model [50], is based on informatic structures largely employed in the 80s, the *blackboards*.
The Hearsay II uses the concepts of *stimulus* and *response frames* of knowledge source instantiations, competition among alternative responses, goals, and the desirability of a knowledge source instantiation, for the development of a general control mechanism.
Experimental results demonstrate the effectiveness of such a model. Inputs to the system are temporal sequences of sets of acoustic segments and associated hypothesized phonetic labels. Several kinds of speech understanding knowl-

edge engines are encoded in several independent knowledge source modules (**KS**s). Some of the employed domains are: acoustic-phonetic mappings, phone expectation-realization relationships, syllable recognition, word hypothesation and verification, syntax and semantics. The model then addresses not only to signal features, but goes beyond, considering also language knowledge.
The state of the system at any point in time is represented by a global data base (the *blackboard*) which holds, in an integrated manner, all of the current hypothesized elements, including alternative guesses, at the various levels of interpretation (e.g., segmental, syllabic, lexical, and phrasal). In addition, any inferred implicative or confirmatory relationships among various hypotheses are represented on the blackboard by weighted, directed links between associated hypotheses. The weight and direction of a link reflect the degree to which the hypothesis at the tail of the link supports (or confirms) the hypothesis at the head. The blackboard may be viewed as a two-dimensional problem space, where the time and information level of a blackboard hypothesis serve as its coordinates.
Processing consists of additions, alterations, or deletions made to data on the blackboard by the various KSs. Each KS is data-directed, that is it monitors the blackboard for arrival of data matching its precondition pattern. Whenever its precondition is matched, a copy of the KS is instantiated (invoked) to operate separately on each satisfying data pattern. Finally, when the KS is executed, its (arbitrarily complex) logic is evaluated to determine how to modify the data base in the vicinity of the precondition pattern that triggered the invocation. The data pattern matching the precondition of a KS is called the stimulus frame (**SF**) of the invocation, and the changes it makes to the data base are referred to as its response frame (**RF**). Each KS may be schematized as a production rule of the form *precondition* => *response*.
The whole process set up in the blackboard, is parallelized as what is suggested by the psychoacoustic models described in the above sections. Such an approach presents problems in the parallel evaluation of numerous alternatives and in the fact that, at any point in time, a great number of KS applications are warranted by the existence of hypothesized interpretations matching the various KS preconditions. A control process is introduced to schedule the numerous potential activities of the KSs to prevent the intractable combinatorial explosion that would inevitably result from an unconstrained application of KSs.
Going into details, the experimental results presented by the authors are obtained as follows [50]:

- All segmental hypotheses are generated from the parametric representation of the acoustic signal.
- All grammatically feasible sentence-initial and sentence-final words are predicted top-down.
- Possible interior words are predicted bottom-up, based on stressed syllable hypotheses constructed from segmental information.
- These predicted words are then rated, and the most likely words in each time interval are placed on the blackboard.
- The control process is implemented, using thresholds.
- A heuristic word sequence hypothesizer attempts to identify the most probable sequences of word hypotheses (consisting of successive language-adjacent word pairs).
- KSs are invoked to attempt to parse the hypothesized word sequences to determine if they are grammatically coherent, to predict possible time-adjacent grammatical word extensions, to hypothesize and verify new words satisfying these goals, to concatenate grammatical and time-adjacent word sequences, to reject phrases and words, and to generate new word sequence hypotheses.
- Whenever a more valid overall sentence hypothesis is generated, weak hypotheses are deactivated, and associated pending actions are eliminated.

A significant amount of tuning of the focussing parameters has to be made, in order to make the system works properly. Nevertheless, the authors declare it is impossible to determine what the optimal values are.
The results from the 61 test sentences in spontaneous speech achieved a 77% of well recognized phrases. No other tests have been made on other confrontation corpora.
Interesting is a statement about some aspects of their approach: $<<$the relatively small grain size of knowledge representation and fine identification of the type and location of knowledge source contributions, apparently affords great advantages in experimenting with mechanisms to control a large, distributed, knowledge-based understanding system$>>$. So, in authors' opinion, starting

from a fine grain is good for having an overall generalization, even if they conclude the work stating that the analysis of their results $<<$indicates that large cost reductions can be obtained by straightforward realization of the proposed focusing principles, particularly if a moderate grain size (the level of word hypotheses) is chosen as a basis for implementing the notions of current state, competition, and stagnation$>>$.
The results were encouraging, even if there wasn't a comparison with other standard models. Unfortunately there were elements which didn't allow the model to go further. The main problems were that the Hearsay II made highly use of the blackboard structures. Lee Erman, one of the original Hearsay-II designers, stated that the reasons for the neglecting of blackboard technology can be resumed in two observations.

- The advantages of blackboard systems do not scale down to simple problems. They are only worth pursuing for complex applications.
- A blackboard system is useful for prototyping an application, but, once developed and understood, the application can be re-implemented without the blackboard structure or opportunistic control machinery.

Other reasons can also be found around in literature.

- Lack of commercial software designed specifically for building blackboard applications.
- The myth that blackboard applications are too slow or too hard to develop.
- Shortage of application developers with experience building blackboard applications.

## 1.5 Multigranular Models in the 90s

The '90s have been characterized by the development of dictation systems and improvements of standard systems performances. With the spread diffusion of personal computers and telephony technologies, a great stress has been given to automatic speech recognition. In the end of the '90s an increasing interest rose again on multi-granular models.

The most significative example is that by Wu [65], which demonstrates that the integration between information coming from syllabic scale and that coming from a phonetic scale, can improve ASR performances, reducing also the dependency from reverber.

The experiment is tested on English digits from *0* to *999* and presents an idea for integrating syllable and phonemes in a single stochastic model.

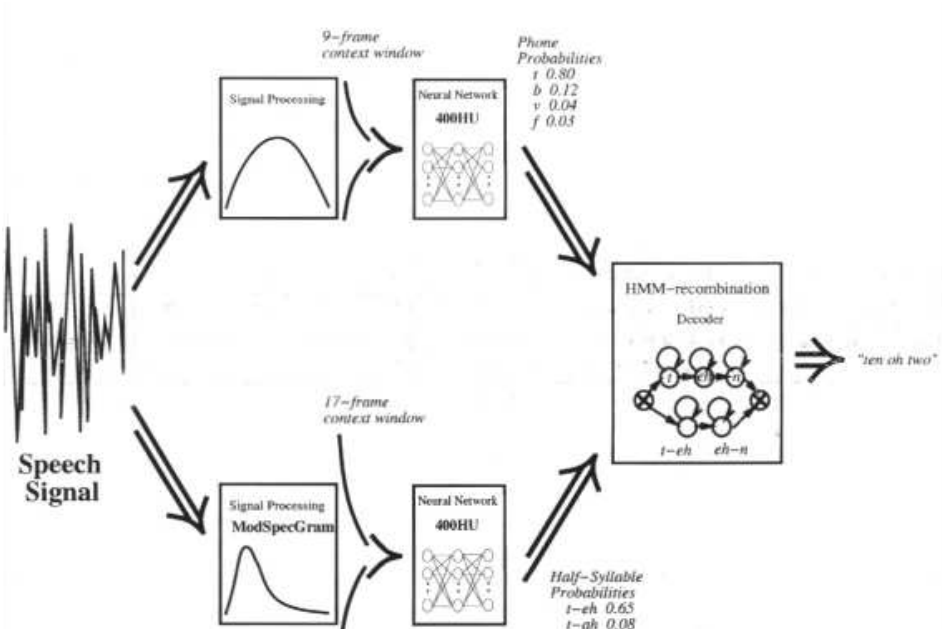



Figure 2: Wu's ASR system.

The best integration is achieved by substituting the two blocks of syllabic and phonetic ASRs, with a single decoding system constituted by a Markov chain in which both syllabic and phonetic pronunciation models are present.

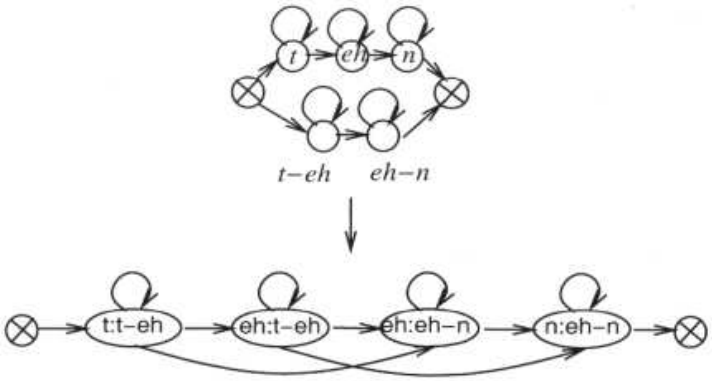



Figure 3: Wu's pronunciation model.

As can be seen from the figure above, from the two separated pronunciation models, a single one is built in which every state represents both a phoneme and a syllable. The whole schema of the recognizers is that in figure 2, where at the exit of Neural Network classification sessions, the probabilities coming out from each of the processing levels are employed together into the decoder. Wu's ASR is based on a syllabic segmentator, developed on Greenberg's Modulation Spectrogram technique [36]. At the end of the essay, results are reported in situation of clear or reverberate speech. Respectively, the Word Error Rate[2] reaches *5.1%* on clean speech and *16.7%* on reverberate speech, while the baseline reference system had reached at best *6.7%* on clean speech and *28,0%* on reverberate speech.

Wu notes how reverber is crucial in speech recognition, and that introducing syllables can partially take care of this problem.

Following Wu's experiment, Ganapathiraju et al. [45], use a purely Markovian approach, employing only HMMs. They try to demonstrate that a markovian model based on syllabic acoustic units, can perform as well as standard systems based phonetic units. They test their model on the same framework of Wu, where a *6.3%* of Word Error Rate is reached by the syllabic model, compared to *5.4%* of phonetic models. Even if the performances by the phonetic model are

[2] defined as $1 - \frac{no.\,substitutions + no.\,deletions + no.\,insertion}{total\ no.\,words\ in\ the\ correct\ sentence} \times 100\%$

better, they try to merge together the two levels, by combining the utterances scores coming from the two models. The performance test is made on the Switchboard corpus, where an absolute increase of *12%* in performance respect to the simple phonetic model is calculated.
Many other experiments can be quoted about the rescoring of phonetic recognizers by means of prosodic features. In their work, Hirose et al. [52], after a recognition process, generate pitch contours for recognition candidates using a speech synthesis scheme, and compare them with the observed countours. The system, tested on the task of detecting phrase boundaries (on the ATR continuous speech corpus) gains an absolute improvement of *5%* respect to a baseline standard system.
Another work following this idea can be quoted. King et al. [58] face the problem of coarticulation modelling by means of syllable modelling. Phonetic tracts (voicedness, tenseness, etc.) are automatically extracted and a segmental recognition is performed. A Hidden Markov Model acts on syllabic units using those feature instead of standard coding. The system shows an overall performance of *36.5%* accuracy on the TIMIT corpus, but the same score is calculated for a standard HMM model of comparison, trained on standard features directly. Their conclusion is that the same kind of information in carried out by standard coding and phonetic "tracts".
Veilleux et al. [44] work is an example of *bottom-up* approach in prosody integration with phonetic information. Speech recognition rescoring is obtained by prosodic profile classification. In order to compute the score of a candidate word sequence and associated parse, automatic break detection is firstly used and the parse is encoded as a sequence of decision trees.
The parse score of a word sequence is then given by

$$S = \frac{1}{n} \sum_{i=1}^{n} \log p(b_i|t_i)$$

Where $b_i$ and $t_i$ correspond to boundaries after the $i$-th word, $t_i$ is a "terminal node" and $p(b_i|t_i)$ is the distribution associated to the terminal node $t_i$. The factor $1/n$ accounts for differences in word length in comparing sentence hypotheses. Each recognized spoken utterance has at least a sequence of break indices, which are scored according to the above expression. Veilleux et al. choose the most probable parse as the intended interpretation. This procedure leads to good performances improvements in phrase disambiguation. The underlying idea is that prosody introduces a further kind of information which is

more linked to speech than to written text, because it is in prosody that the power of interpretation disambiguation lies.

All these works, at the end of the '90s, underline the importance of introducing multi-granularity in speech processing as a crucial step for outperforming ASR design and implementation. This is the starting point of the work described in this thesis.

## 1.6 Multigranular Models Today

The experiments illustrated in the previous section, have been a fundamental layer for the models of the last years, which have largely investigated the integration of multiple acoustic information with different time span.
Following Greenberg's trend [28], and Wu's experiments, an approach to multi-granularity representation is made by Chang [9]. He builds up a multi-tier model (cfr. figure 4) where speech is organized as a sequence of syllables, in contrast to the conventional phonetic-segment based structure assumed in most ASR systems. It differs from the standard syllabic models in representing single syllables as sets of acoustic cues instead of a succession of phonetic features. Acoustic cues refer to the phonetic structure of the syllable in terms of manner, place of articulation, vowels, etc., and also to some non-segmental information supporting the recognition of a syllable in a word. Word templates made up of a succession of such syllabic-phonetic features are introduced and the possibility of mutation for these descriptions is associated to pronounce variability. Chang's system performance reach about *10%* of Word Error Rate on numbers recognition ranging from *0* to *999* (taken from the Switchboard corpus) .

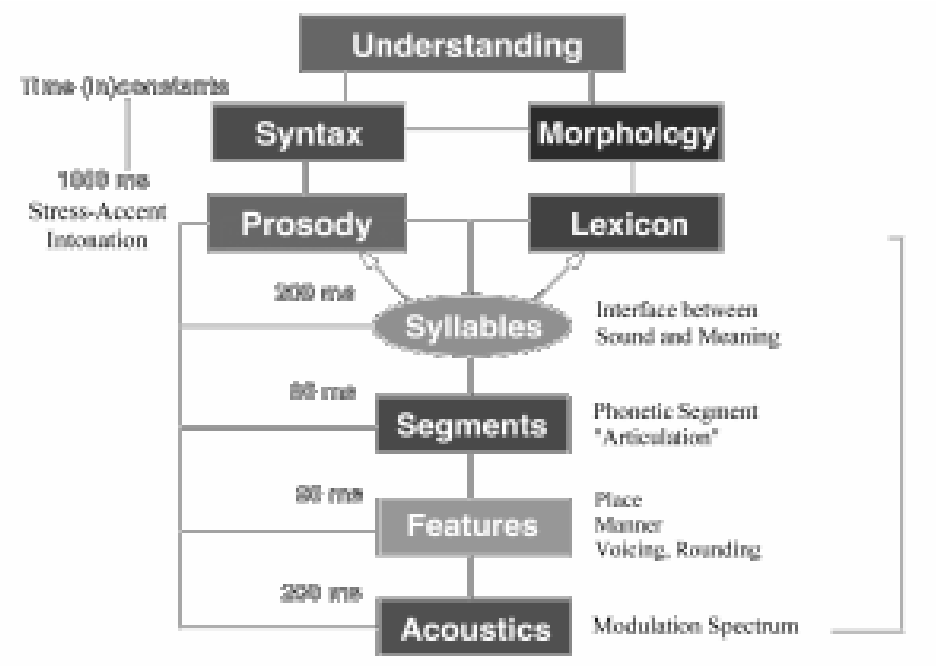



Figure 4: Chang's ASR model.

The above approach, as well as that in Wu [65] or in King et al. [58], can be classified as *explicit* models. They address the problem to understand the perceptive phenomena which lie under a speech signal and model them in an explicit way. E.g. King et al. experiment is an attempt to catch an explicit representation of an hidden feature like coarticulation. Chang [9] tries to find a description or a definition for a syllable and Wu addresses the explicit model of the interaction between syllabic spanning information and phonetic characteristics. These examples can be compared to other kinds of approaches, which start from the basic assumption that the so called hidden features cannot be explicitly described, but a good machine learning system could extract them automatically from a standard description.

Wang et al. [53] explore the concept of stress modelling by means of prosodic features and standard models. They add stress markers to a speech recognizer in order to improve performances. The underlying idea is that stressed syllables provide islands of phonetical reliability, and this information can help an ASR. The model tries to abstract the concept of stressed syllable from a well

defined set of features, like energy, pitch, duration etc. Experiments are made to estabilish the best combination of features which can describe lexical stress, and a *5.3%* improvement in relative accuracy, respect to a standard system, is calculated on the JUPITER corpus.
Kocharov et al. [42] follow an *implicit* modelling approach on multiple acoustic features combination. The spectrum derivative is introduced and combined with the standard MFCCs and voicedness markers. Linear Discriminant Analysis is applied to find the optimal combination of the different acoustic features. Experiments, performed on german continuous digit strings, recorded over telephone line, reach a relative improvement of *20%* in accuracy respect to standard baseline models, while a *4%* relative increment is reached in german large-vocabulary conversational speech (VerbMobil II corpus).
Livescu et al. [3] investigate the use of Dynamic Bayesian Networks to catch hidden features from phonetic characteristics. In previous works the representation has typically been implicit, relying on a single hidden state to represent a combination of features.

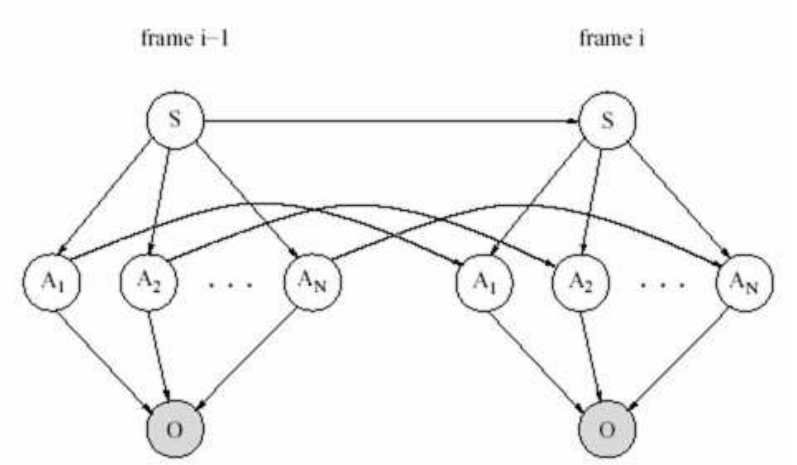



Figure 5: Dynamic Bayesian Network for hidden features.

In figure 5 the model is depicted. In each frame, there are $N$ *hidden features* $A_1 A_2 ... A_N$ , each depending on the current phonetic state $S$ and on its own value in the previous frame. $O$ is the vector of observations (i.e. acoustic

features), which depends on the current $A_i$ features. The intuition for this structure is that, at any instant, each feature $A_i$ is at the target value for the current phoneme, but it is also affected by its own value in other frames because of inertia and continuity constraints. The model is applied to words recognition and performances gain a *25%* of relative accuracy on clean speech in connected digits recognition (on Aurora 2.0 corpus).

Prosody is a key feature because it introduces *super-segmental* information. The reasons for using prosody in ASR has been investigated by several authors. The integration between multi-granular levels has followed mostly a bottom-up process where prosody has been used to rescore the results of phonetic or syllabic recognizers. In some cases a top-down approach is used, where segmental recognizers act on parameters modified by a prosodic analysis.
In this thesis the concept of multi-granularity is investigated towards the integration of the phonetic, syllabic and prosodic levels.
Vergyri et al. [54] analyze prosody in ASR by integrating this kind of knowledge source into a state-of-the-art large vocabulary recognizer. According to them, prosody manifests itself on different levels in the speech signal: within the words as a change in phone durations and pitch, inbetween the words as a variation in the pause length, and beyond the words, correlating with higher linguistic structures and non-lexical phenomena. They investigate three models, each one corresponding to a prosodic model, and eventually merge them. Experiments on the Switchboard corpus show word accuracy improvement adding each prosodic knowledge source. A further improvement is observed with the combination of all the models, demonstrating that each of them captures somewhat different prosodic characteristics of the speech signal.
The first model addresses *word* duration. For each word a duration feature is a vector comprising the durations of the individual phones in the word. For example, the word *"that"*, represented as the phone sequence *dh+ae+t*, is associated to the vector *(10.0 8.0 4.0)*, where the three values represent the durations of the three phones *dh*, *ae*, and *t*, respectively. The duration models are used to rescore the recognition hypotheses in an $N$-best list. In this way, the standard acoustic features $O_A$ are accompanied by the word-duration features $O_D$ in words representation.
The second model introduces pauses into the language model $N$-grams. Probabilities associated to the transition between words are conditioned by the length of the pauses following the words.

The third model deals with *hidden prosodic events*. Prosody correlates also with linguistic structures beyond the words themselves, and includes cues other than durations. Some higher-level phenomena, such as sentence boundaries and speech disfluencies, manifest themselves prosodically and can be thought of as *hidden pseudo-words*. Word sequences are tagged as alternations of word and prosodic events $W_1E_1W_2E_2..W_nE_n$. During testing, the events are unknown, and the model for the situation becomes equivalent to a HMM, whose states are *(word,event)* pairs.
The merged model is obtained from the integration between all the levels. The best sequence of words and pauses $WS^*$ is calculated by the following approximation, in which a product between the probabilities from the single models is performed

$$WS^* \simeq argmax_{ws}(\sum_E P(W,E)P(F,E))P(S,W)P(O_A|W,S)P(O_D|W,S)$$

$O_A$ are the standard acoustic features, $O_D$ the word-duration features, $E$ is the sequence of hidden prosodic events $E_1E_2..E_n$, $F$ is the set of acoustic features for the $E$ events, $W$ is the word sequence and $S$ is the inter-word pauses sequence. Improvements in performance are tested on the Switchboard corpus (the NIST Hub-5 benchmarks), where an absolute increase of *1%* is found respect to a baseline standard HMM system.

Shriberg et al. [57] face the problem of the integration between prosody and language models in speech recognition. The aim, as usual is to calculate the joint probability *P(W,S)* of a sequence of words $W$ and target classes $S$. A prosody model is defined as a framework in which the probability *P(S|F,W)* is calculated, where $F$ is a set of prosodic features. After a phone level alignment of the training set (taken from Switchboard and Broadcast News corpora) they provide duration of pauses, syllables, rhyme, vowel duration and speaking rate to a decision tree which acts a classification. On another side they introduce a language model *P(S|W)* used during the extimation of *P(W,S)*. The probability *P(S|W)* is calculated to predict the possible classes given the words. Finally a merging phase follows, in which they suggest three methods for language and prosodic model integration

- Posterior interpolation. Conditional probability *P(S|F,W)* is computed via the prosodic model. Also *P(S|W)* is calculated and then a linear

combination of the two is performed.

- Posteriors as features. *P(S|W)* is calculated via the language model and this posterior estimate is used as an additional feature for a prosodic classifier.

- HMM-based integration. Likelihoods *P(F|S,W)* are obtained from the prosodic model and used as observation probabilities in a HMM associated to the language model. The HMM then calculates *P(S|F,W)* exploiting both the kinds of knowledge.

Their experiments, using the third approach, demonstrate an improvement of relative *2%* in speech recognition accuracy respect to a simple N-gram baseline model on the Switchboard corpus. Better results are showed in disfluencies detection. They think the weak point of the application to speech recognition is in the integration method, which should be more sophisticated. Shiberg et al. [57] experiment is an example of top-down approach because prosodic analysis comes before phonetic recognition.

## 1.7 Discussion

So far a panorama of *multi-granular* models has been depicted. The idea belongs to early studies about language. The concept has been developed during last decades and several approaches are born, which aimed to catch the multiple dynamics lying in a speech signal. A vocal signal appears to be a concurrence of several acoustic events with different time spans, which act together in order to make human recognition robust and efficient. Psychoacoustic models and experiments support the hypothesis.
A border has been traced in modern approaches. Someone uses pure mathematical models to embed the acoustic phenomena, as they have the role to extract such layers from the signal features. Other ones prefer to explicitly model those events in order to discover what is the nature of such dynamics. All the results are in agreement with the fact that using multiple sources of information is fundamental for ASRs performances emprovement. All the approaches are not meant to emulate human speech recognition system, but to partly simulate it.
The model presented in this thesis will follow an hybrid approach, respect to the ones described above. There will be two layers, the former exploiting a

mixture of syllabic and phonetic models, which makes use of a mathematical model aiming to automatically extract phonetic and syllabic dynamics from a sequence of acoustic features. This model follows what has been defined as an *implicit* modelling approach. The latter will be an *explicit* model of the prosodic information lying in the speech signal. The two models will be merged together and results will show an evident improvement in performances respect to a standard baseline system. The merging technique will be largely discussed as it rises deep questions about the limits of acoustic information representation and words structure.

The discussion will always focus on features coming from signal characteristics, without going to explore other aspects, like semantics or pragmatics, which could be equally fundamental. This choice as been made in order to understand what kind of useful information lies in the pure speech signal and how to exploit it at best.

Chapter II will deal with the difference between segmental and non-segmental information, with a catalogue of several techniques to extract features for speech units representation.

Chapter III illustrates the model chosen as the baseline for performances comparison. This model has been built with the standard ASR structure.

Chapter IV presents the ASR that has been taken as the layer for *segmental multi-granular* recognition.

Chapter V introduces the model for prosodic representation and recognition.

Chapter VI focuses on the merging phase between the models.

Chapter VII summarizes the experimental results on the chosen corpus.

Chapter VIII analyzes the results and their meaning.

# 2 Chapter II: Segmental and Non-Segmental Speech Features

## 2.1 Introduction

Phenomena representation and coding is a big informatics branch, and the term *feature* generally refers to a code related to some signal characteristics. The term *segmental feature* generally indicates a characteristic of an acoustic signal which acts in a well defined temporal length of a speech segment. Examples can be found for the *phonemes*, because they generally include 10 ms stationary speech portions with discriminant characteristics, which can be automatically extracted. All the other phenomena associated to the articulatory structure or to characteristics for which no borders can be marked, are defined *non-segmental*. Examples for such events are the "*segment internal temporal structures*" which does not necessarily directly depend to segment or sub-segment boundaries [10]. The so called *super-segmental* features, are included in the class of non-segmental characteristics, as they refer to signal properties which "add" to segmental units and typically have longer spans and a not well defined periodicity. An example is prosody, which is a very useful information in speech understanding and adds its information to the segmental layer.

The reason to make a difference between segmental and non-segmental information, is that while the former is often associated to speech units with well defined spectral characteristics, the latter refers to features which are not always formalizable. Many automatic speech recognizers make use of stochastic classification models set up on speech units. Such methods associate a speech unit to a sequence of acoustic features, or are able to calculate the likelihood of a speech unit to such sequence. Acoustic features associated to segmental information univocally correspond to a speech unit, but this is not the case with non-segmental features. This is the first problem to face when building a speech recognizer addressing to such information.

A more detailed explanation about the differences in information representation can be given by showing phonemes and syllables biological production. As stated in the previous chapter, syllables are units which have not well defined boundaries, sometimes they are purely perceptive events and it is not possible to detect a well defined set of features which is able to precisely discriminate among them. They can be considered on the borderline between non-segmental and segmental phenomena.

Speech is produced by air-pressure waves emanating from the mouth and the nostrils of a speaker. In most of the world languages, the inventory of phonemes, as discussed in the previous chapter, can be split into two basic classes:

- Consonants - articulated in presence of constrictions in the throat or obstructions in the mouth (tongue, teeth, lips) as we speak
- Vowels - articulated without major constrictions and obstructions

The sounds can be further partitioned into subgroups based on certain articulatory properties. These properties derive from the anatomy of a handful of important articulators and the places where they touch the boundaries of the human vocal tract. Additionally, a large number of muscles contribute to articulatory positioning and motion [30].
The most fundamental distinction between sound types in speech is the **voiced / voiceless** distinction. Voiced sounds, including vowels, have in their time and frequency structure, a roughly regular pattern that voiceless sounds, such as consonants like *s*, lack. Voiced sounds typically have more energy as shown in figure 6.

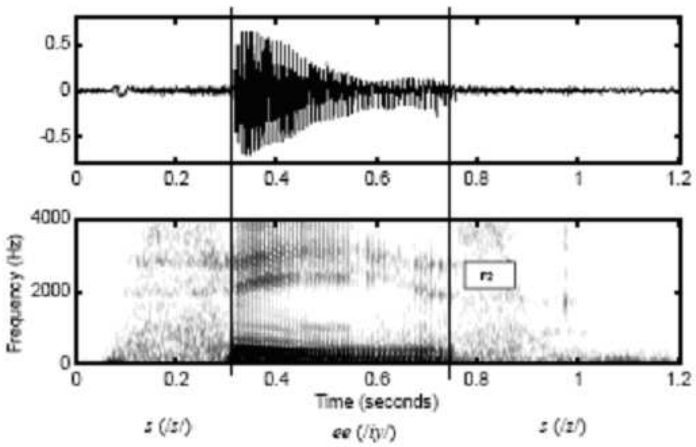



Figure 6: Spectrogram representation of the word *sees*.

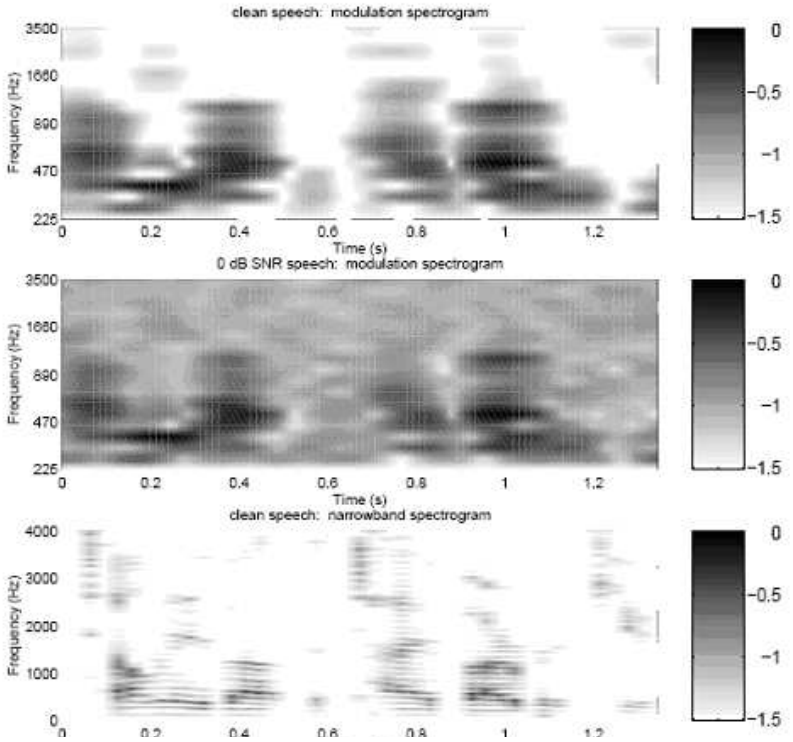



Figure 7: Modulation Spectrogram representation of a vocal signal [36].

It can be seen that the waveform of the word *sees*, consists of three phonemes: an unvoiced consonant /s/, a vowel /iy/ and, a voiced consonant /z/.
When the vocal folds vibrate during phoneme articulation, the phoneme is considered voiced, otherwise it is unvoiced. Vowels are voiced throughout their duration. The distinct vowel timbres are created by using the tongue and lips to shape the main oral resonance cavity in different ways. As can be seen from figure 6, and can be argued from the discussion above, phonemes can present well defined characteristics and can be inscribed in 20 ms segments. The representation of the signal by means of the spectrogram emphasizes this aspect.
Differently from the segmental situation, a syllable is not well discriminable by means of a precise scale analysis. They need about 100-200 ms segments in order to be catched properly, but the lenght is highly variable in that range. As will be largely discussed in the next section, a rough version of the spectrogram can be

associated to such a unit. This representation, called *Modulation Spectrogram* makes clear two aspects:

- The syllables are not as well classifiable by an automatic system as phonemes
- Such units are more robust to reverber or environmental variability

This points are clear by looking at figure 7, where in presence of reverber, the spectrogram is completely altered, while the modulation spectrogram is still recognizable.
The next section shows an overview of the most used techniques for features extraction, either segmental or non-segmental, where a particular stress will be given to difficulties in units coding.

## 2.2 Signal Representation

### 2.2.1 Segmental features

As yet explained, segmental features address to well temporally defined characteristics of a speech signal. Most of the techniques used to extract such information, are aware about the identity of entities to search for. The following sub-sections will present two methods for phonetic features extraction. The first, the *LPC* method, is born as an attempt to find an information representation which was univocally associated to the formant frequencies of phonemes, but also be robust to little environmental changes. The second, the *MFCC* method, is an evolution of the previous technique, which introduces different focuses and scales of analysis referring to human speech perception. Such an innovation is to make the segmental features more robust, and to extract that part of the phonetic information which is most invariant among different speakers.

**LPC** *LPC* is the most classic method for phonetic features extraction. The aim is to catch the formant frequencies of the phones constituting a vocal signal. It is not an accurate or robust model, in that the technique produces values which are strictly dependent on noise and recording modality.
The reported description has an historical motivation, because this is the starting point for all the successive methodologies addressing to phonemes representation.

LPC is based on the *Source-Filter* model for speech production. In human phonatory apparatus, air goes from lungs to larynx, where it can find resistance by vocal cords. It then proceeds to the pharynx and mouth in the oral cave.
If vocal folds activity is present, then a sound is produced like a succession of impulses which vary air pressure. From larynx on, the sound is amplified by the reflections on the oral cave walls.
Only some of the frequency components of the signal are emphasized by resonance, depending on the mouth and pharynx shape.
According to this predictive model, the production of consonants or vowels in presence of vocal cords activity, can be seen like a train of impulses followed by resonances. Such sounds are called *voiced* or *sonorant*. When there is no activity by the vocal cords, then the sound is called *voiceless*, this is the case of many consonants or few vowels in spoken language. The signal in this case is similar to modulated noise. The oral cave can be represented by a single dimension system, which can be dicretized into a succession of $N$ filters.
In vowels or sonorant consonants, there is a succession of impulses which propagates from the glottis to the lips. The system shape will depend on the sound produced.
In voiceless consonants a simple air flow passes through the succession of filters.
Figure 10 summarizes the whole process. The switch represents the choice between voiced or voiceless sounds.

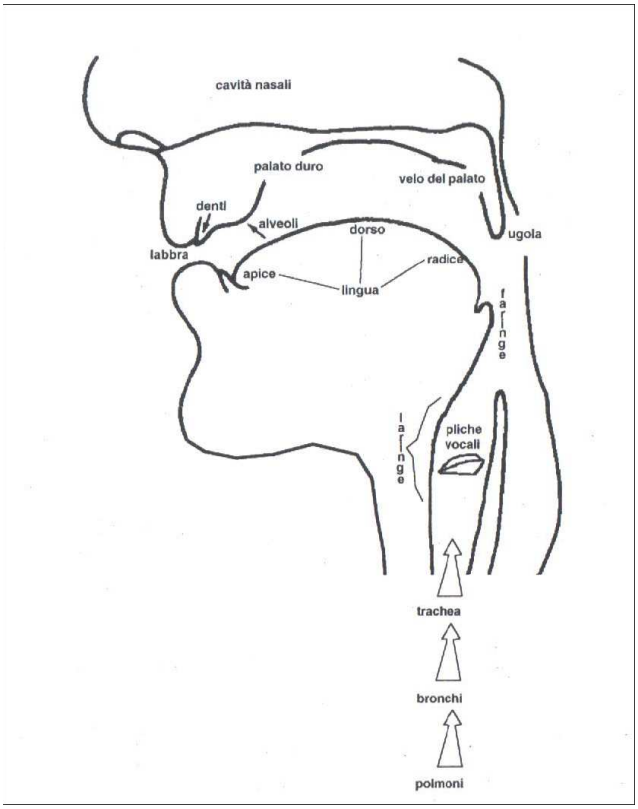



Figure 8: Representation of the air route form the lungs to the lips.

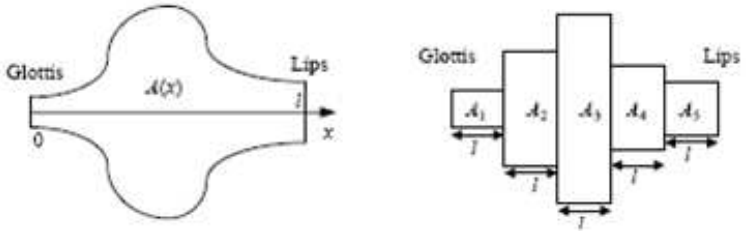



Figure 9: Source-Filter model representation.

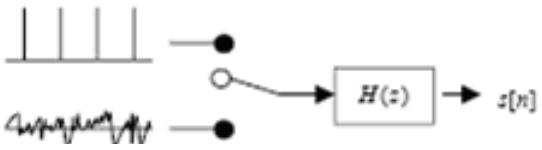



Figure 10: *Source-Filter* model schema.

From the above schema it can be argued that, ones the exciting impulse train and the filter set has been detected, the signal is univocally determined.
The LPC technique tries to go up to filters identities, and to the resonance frequencies they produce.
In common applications the oral cave is approximated by an all poles filter [43]. In order to have a perfect approximation, those poles should be infinite in number.
If $E$(z) is the $Z$-transform of the glottis excitation, $H$(z) is that of the filter impulse response and $S$(z) is that of the exit signal, for the properties of such systems it results that

$$S(z) = H(z) * E(z)$$

Where

$$H(z) = 1\frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}$$

Is the all pole filter representing the oral cave.
In practical cases, we start from the signal and reconstruct the filter. So the inverse relation is used

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + e[n]$$

Where $e$[n] is the impulse train by the glottis.
The LPC technique takes its name from the fact that the $n$-th signal sample can be predicted by a linear combination of $p$ coefficients, where $p$ is the order of the approximation.
The idea of the recognition systems based on LPC, is that the coefficients $a_k$ are univocally associated to the phonemes characteristics. The equation above can be solved in order to calculate such coefficients and use them as acoustic features in phonetic models. The signal analyzed is typically a 20 ms segment of speech and a vector of $p$ elements is extracted, which represents the characteristics of the phone lying in that piece.

Typically a good approximation for $p$ is 12. The reason for such choice is that a low value results in a rough approximation, while an high value leads to a confusing filter, which models all the frequencies in the spectrum and not only the formants.
There are standard methods for calculating the $a_k$ coefficients from the equation above. Among them the most famous methods, which will not be discussed here, are [30]

- The covariance
- The autocorrelation
- The lattice formulation

**MFCC** *Mel Frequency Cepstral Coefficients* is a short time coding technique addressing 20 ms pieces of signal. Like LPC it refers to the source-filter model, but with an innovative idea. The frequency analysis is conducted on a filterbank rather than on a single filter. This happens because a different focus is given to several regions of frequencies, according to human acoustic perception.
Suppose $x$[n] to be the vocal signal and $X[k]$ its associated Discrete Fourier Trasform [43].

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi nk/N}, \quad 0 \leq k < N$$

A filterbank of $M$ filters is introduced, where the $m$-th filter is

$$H_m[k] = \begin{array}{cc} 0 & k < f[m-1] \\ \frac{2(k - f[m-1])}{(f[m+1] - f[m-1])(f[m] - f[m-1])} & f[m-1] \leq k \leq f[m] \\ \frac{2(f[m+1] - k)}{(f[m+1] - f[m-1])(f[m+1] - f[m])} & f[m] \leq k \leq f[m+1] \\ 0 & k > f[m+1] \end{array}$$

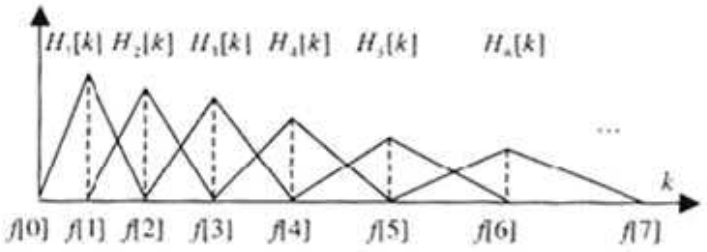



Figure 11: Representation of the filterbank.

Such filters emphasize the spectrum around certain frequencies, whose values are obtained according to *mel* bands. Lower frequencies will be analyzed in detail, while higher frequencies will have a lower focus.

$$f[m] = (\frac{N}{F_s})B^{-1}(B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1})$$

Where $f_l$ and $f_h$ are the lower and higher boundaries of the $m$-th filter, $F_s$ is the sampling frequency of the signal, $M$ the total numer of filters and $N$ the FFT samples [43].
$B$ is the mel scale transformation function for frequencies. It is defined as

$$B(f) = 1125ln(1 + f/700)$$

$B^{-1}$ is its inverse, defined as

$$B^{-1}(f) = 700exp((f/1125) - 1)$$

The number of filters, $M$ , is the coding order, that is the length of the vector which will represent a 20 ms signal. From each filter a single coefficient is calculated.

Each Mel Frequency Cepstral Coefficient is obtained by the following transformation of the signal and filters.

$$c[n] = \sum_{m=0}^{M-1} S[m] cos(\pi n(m - 1/2)/M), \quad 0 \leq n < M$$

Where

$$S[m] = ln[\sum_{k=0}^{N-1} |X_e[k]|^2 H_m[k], \quad 0 < m \leq M$$

The coding vector is so the discrete cosinus transform of the production of the $M$ filters.
Common uses of such technique, set $M$ to a value of 13 on about 20 ms speech segments overlapped by 10 ms.
Many experiments [30] have demonstrated that the nature of the introduced filterbank, in combination with the cosinus transform, makes these coefficients more robust to noise respect to LPC coefficients. This is the most used technique in automatic speech recognizers based on phonetic base units of speech.

### 2.2.2 Non-Segmental features

The following is an overview of *non-segmental* features extraction. While in the previous case, the methods searched for a specific form of information, e.g. formant frequencies, now some events or cues are investigated, which are non-segmental features, but it is not always clear which are the units addressed to.
The presentation in this section, will start from methods for syllabic features extraction. Such techniques are not able to describe them completely, and this is one of the most important problems in speech recognizers based on those units. Other techniques will deal with prosodic information extraction. Even in this case the list of features will not be sufficient to completely represent the phenomenon.

**Modulation Spectrogram** The Modulation Spectrogram technique is born to catch syllabic information, to be employed into automatic syllables segmentators. It has been introduced by Greenberg [36] and applied in speech recognition in [65].
Referring to figure 12, the passages of information extraction are the following:

- The vocal signal is passed through a FIR filterbank of trapezoidal shape, with a relative superposition depending on the filter frequencies. The process, as in the MFCC analysis, tries to simulate the sensibility variation of the human auditory system to different frequencies. Each exit of the filterbank represents the signal, filtered according to a different frequency band. The number of filters usually employed in common applications is 20
- The signal is cut in the negative part and then enveloped (with a lowpass filter at 28 kHz), to better emphasize the units

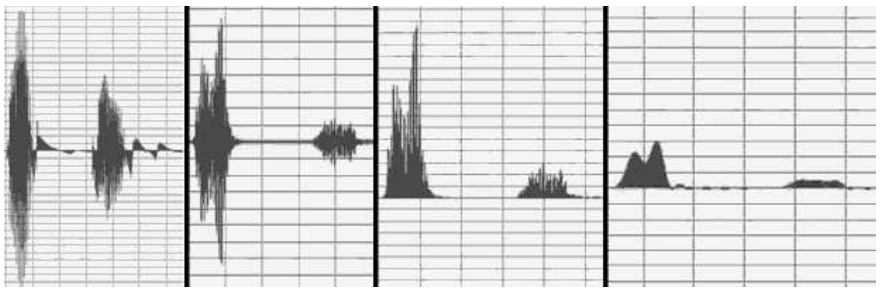


Figure 12: Representation of the first two steps.

- The enveloped signal is downsampled to reduce processing complexity
- The Fourier Trasform is calculated on 250 ms windows, overlapped by 25 ms and the components at 4 Hz are recorded

The process, for each exit of the filterbank, returns a succession of the spectrum amplitude at 4 Hz for segments of 250 ms overlapped by 25 ms. These values represent the spectral components of events with 250 ms periodicities, that should correspond to a syllable. According to Greenberg [36], the analysis of 4

Hz modulations are associated to low speech variations, which can be syllabic events.
Psychoacoustic experiments [39] have showed that the slow modulations over 16 Hz are not necessary to human speech recognition and an acoustic signal can be still understandable even if only the modulations up to 6 Hz are preserved. From such results it can be guessed that long analysis segments can be much robust to interferences or noise, even if they loose fine aspects of speech structure [41]. That means the modulation spectrogram is not performant, if used as a recognition feature, but it can be useful in syllable segmentation.

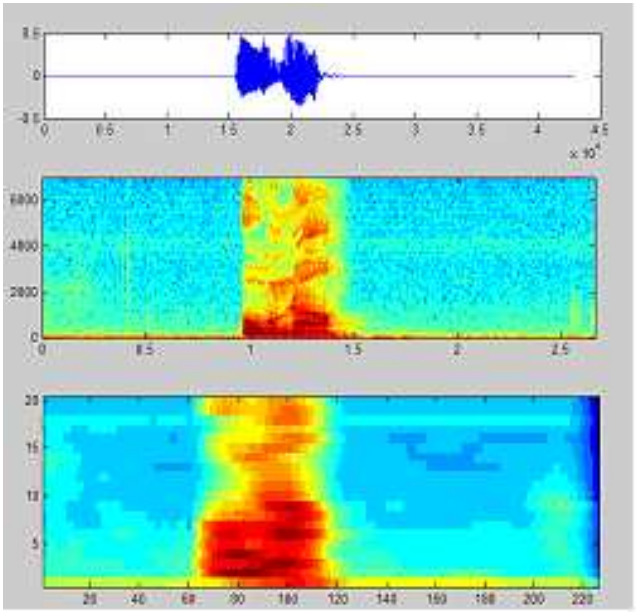


Figure 13: The modulation spectrogram of the word *cinque*, compared to the classic spectrogram.

In figure 14 the behaviour of the technique in presence of white noise is depicted. The disturb destroyes both the spectrogram and the modulation spectrogram, even than the latter is affected in a weaker way.

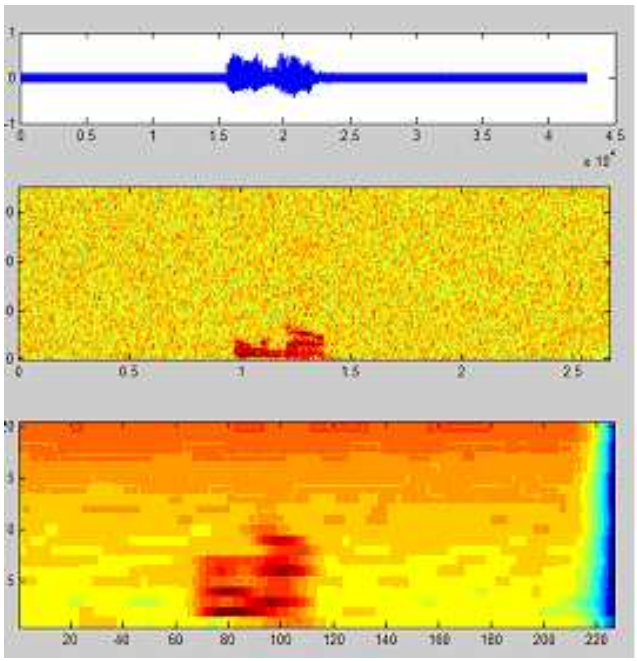


Figure 14: The modulation spectrogram of the word *cinque*, in presence of white noise.

Figure 15 shows the case of reverbered signal. It is evident that the aspect of the spectrogram changes strongly, while the Modulation Spectrogram is still recognizable. That is because reverber duplicates the formant frequencies, while it affects the slow variations in a minor way.

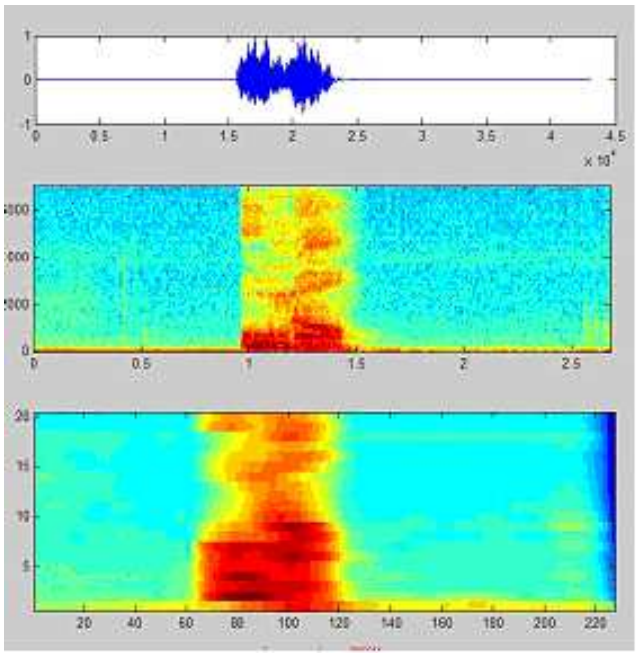


Figure 15: The modulation spectrogram of the word *cinque*, in presence reverber.

**Pitch, Energy and Duration** Fundamental prosodic aspects are defined here, referring to the intonation and emphasis of an utterance:

- In the production of a speech signal, in a portion of about 20 ms, where signal is supposed to be stationary, the fundamental frequency or *pitch* is the frequency of oscillation of the vocal folds. Pitch is what makes people perceive sounds as acute or grave. Other frequencies intervene in a signal, which are related to resonances in the vocal tract. A segment in which pitch presence is detected is called *voiced*, otherwise it is called *unvoiced* or *voicedless*. The process which automatically extracts the pitch is usually the autocorrelation procedure [30] which will not be described in this frame. The general procedure uses short-term analysis techniques, which calculate the signal autocorrelation value $f(T|x_m)$ for every frame $x_m$ of length about 10-20 ms, where $T$ is a possible pitch period. The choice of the best pitch period in the segment, is taken by evaluating

$$T_{best} = argmax_{|T}(f(T|x_m))$$

- The *energy* of a speech segment is the intensity by means of which that segment has been produced. It is calculated as

$$E = \sqrt{\frac{\sum x_i^2}{N}}$$

$N$ indicates the number of elements in the speech portion, while $x_i$ is a single sample. Generally the energy trend of a signal is extracted calculating the energy of 10-20 ms segments every 5 ms.

- The *duration* of a speech segment is simply its time length. This aspect is responsible for speech rate.

Duration is a key feature in speech representation, because it is related to stress and metrics. The succession of long and short duration syllables is responsible for the emphasis of some parts of the utterance as well as to the stress and the rhythm.

**Differenciated Energy** The energy value for detecting the presence of a fricative consonant in a speech segment, is called *differenciated energy*. Fricatives are characterized by the spectral presence of noise with high varying frequency and weak formantic structure at low frequencies. If the signal is high-pass filtered, with a cut-off frequency of 1100 Hz, then the obtained energy trend will be very different from the previous one, in the regions where a fricative consonant is present.

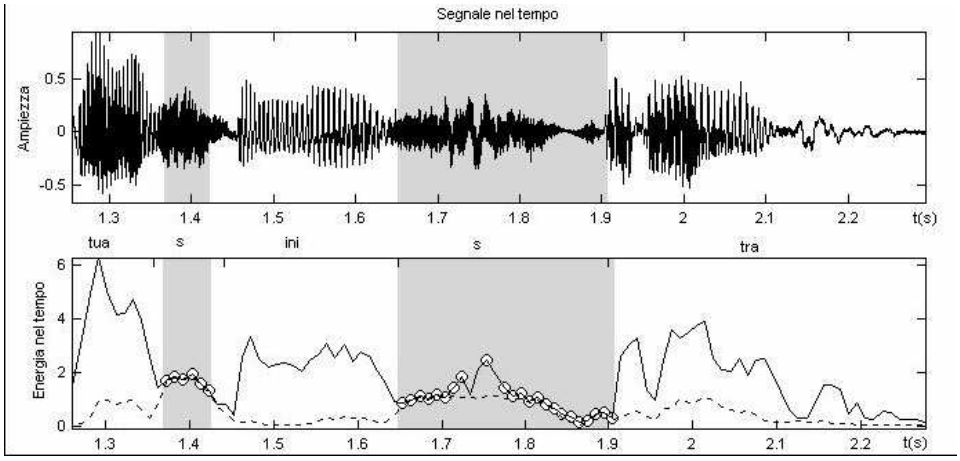



Figure 15.1: Differenciated energy representation and comparison.

**Melodic Accents of Words** Melodic accent is the accent related to intonation and the height of a note. It can be defined at the level of words or phrases, because for each scale such characteristics can be detected.
More formally, a *melodic accent* is defined as the variation of pitch in a time unit. In human communication, it is used to give importance to a part of the dialogue and to mark some parts of a phrase. It has not to be confused with the rhythmic accent, especially because it has not a precise place of the word in which to fall.

The melodic accent is usually calculated by means of Fujisaki accent components (ref. 2.2.2), which will be discussed in further detail later.

It can be represented as a train of rectangular impulses (ref. figure 16). The interesting thing about this feature, is that, for short words, the configuration of the impulse trains is always the same, even if speaker changes.

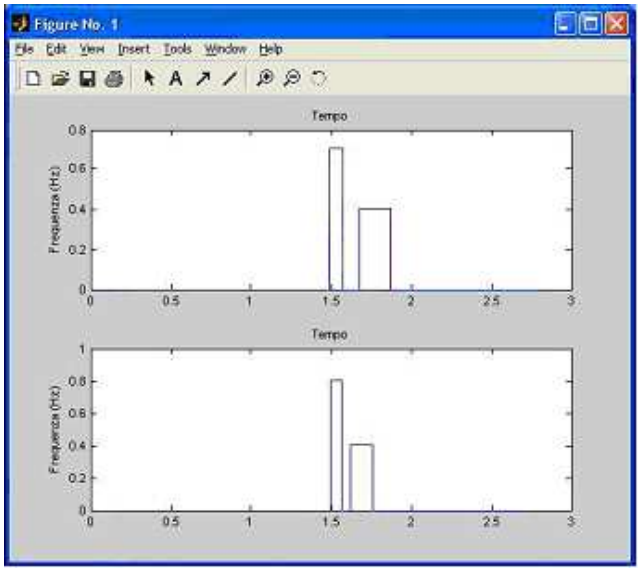



Figure 16: Melodic Accent representation, for the word *quattro*, pronounced by two different speakers. Notice that there is always an high rectangle, followed by a shorter one.

**Melodic Accents of Phrases** Melodic accents of phrases are formally defined as the variation of the fundamental frequency in the domain of the entire utterance, rather than in that of a single word.

Such dynamic is what is commonly defined the *intonation profile*¸ which is responsible for the difference between interrogative and declarative phrases perception.

As can be noticed from figure 17, the speaker utters the phrase "*la mamma mangia la mela*" with a declarative intonation. The profile presents a main peak on the second syllable of the word *mamma*, and a light peak on the second syllable of the word *mangia*, even if the important information lies in the rise and descent of the pitch trend.

The second utterance, with an interrogative intonation, presents a smaller peak on the second syllable, but another one on the first syllable of the word *mangia* (ref. figure 18).

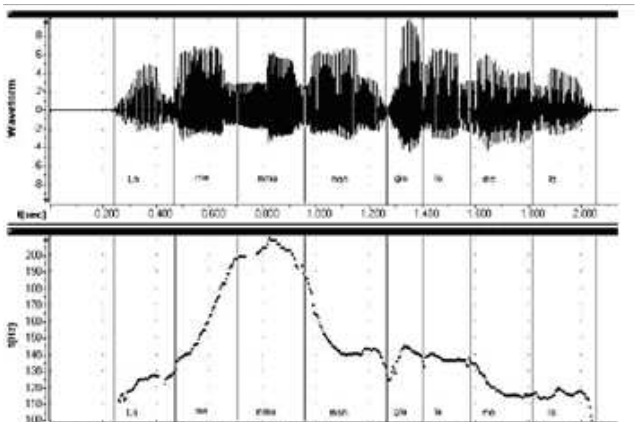


Figure 17: Pitch contour for the utterance *la mamma mangia la mela.*

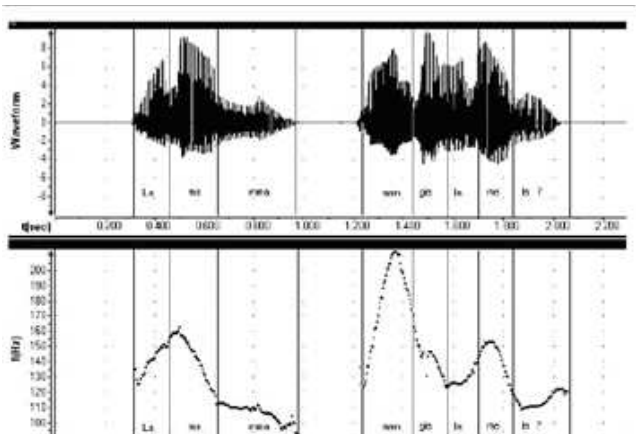


Figure 18: Pitch contour for the utterance *la mamma mangia la mela ?*

**Rhythmic Accents** *Rhythmic accents* can be seen as the set of features which emphasize a syllables in a complex contex [3]. For example the accent of a phrase in a poem, which gives a verse its cadence. Rhythmic accents accompany the principal tonic accent. The Italian word "*indiscutibilmente*" has eight syllables, where only one of them has a principal stress. Furthermore, from a rhythmic point of view, it has also stresses on the first and fifth and seventh syllables.
Such characteristics are linked to the alternation of long and short syllables. Languages tend to avoid two adjacent accented syllables, while regularity is searched.
Stress is crucial in dialogues structure and listener attention. More formally, the presence of a stress on a syllable depends on the syllables energy variation and duration. The Silipo-Greenberg procedure [27] describes how to calculate such feature, but it will not be discussed in this section.

**Vowels Extraction** Information coming from energy and pitch could be sufficient to distinguish words like *uno* and *due*. Unfortunately such parameters are not sufficient to discriminate more complex words as for example *diciassette* and *diciotto*.
In order for prosodic information to be able to manage such cases in a speech recognizer, it is necessary that more information is added.
Vowel configuration is a good candidate for such an aim. An automatic vowel extractor has been introduced in [20] for automatic speaker identification. The vowel configuration is in fact strictly linked to a speaker identity, because it includes formant frequencies information, which depend on the mouth and oral cave shape.

As showed in figure 19 the process acts in the following steps

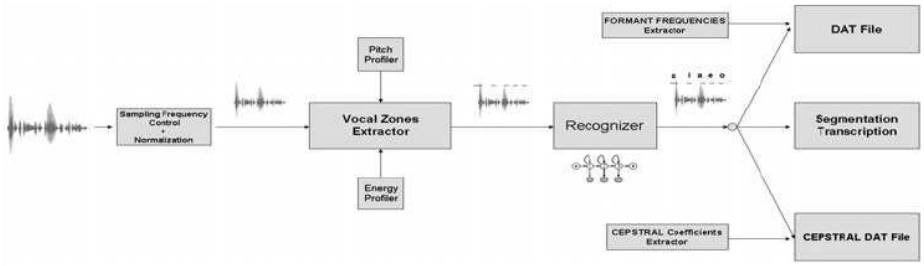



Figure 19: The Vowels extraction procedure schema.

- The signal is downsampled at 8kHz for computational benefits
- A pitch and energy synchronized analysis is performed, where *stable* and *prominent* zones are extracted. The term *stable* refers to a segment where pitch varies slowly, while *prominent* means that zone has even an high energy value
- Each extracted piece of signal is then passed to a HMM model for vowels classification

The model is *not* really able to get all the vowels in the signal, but it catches most of the prominent ones.

**Fujisaki rectangles and impulses** The Fujisaki-Hirose model [29] is a complex prosody analyzer. According to such model, the melodic trend of a phrase is made of two components:

- The *phrase component*, which is linked to intonative syntagms[3], characterized by a fast ascending phase followed by a slow descending trend. Such feature describes the pitch contour of an utterance

[3] A *syntagm* is a string of sounds which have the same logic function in a phrase, according to a syntactic structure.

- The *accent component*, which describes the modulations introduced by the speaker to mark a particular melodic istant

The phrase and accent components are obtained by filtering two signals coming out from two linear systems.
The first signal is a Dirac impulses train, while the second is a rectangular impulse train. Given a signal $s$(t) with a pitch curve described by the function $p$(t), two signals $x_1(t)$ and $x_2(t)$ are extracted which represent the phrase and accent components.

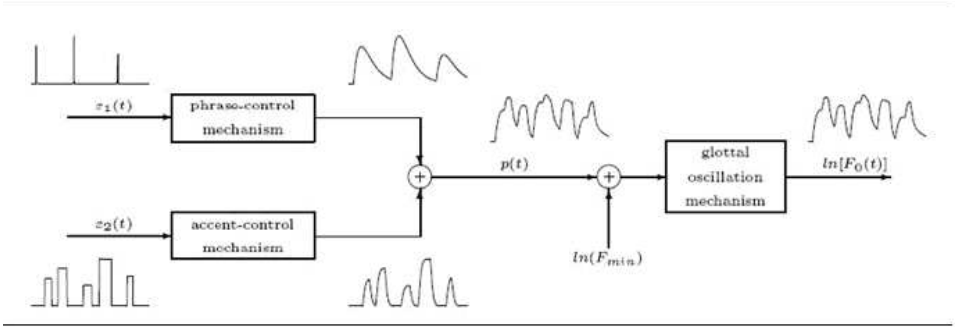



Figure 20: Representation of the Fujisaki-Hirose model.

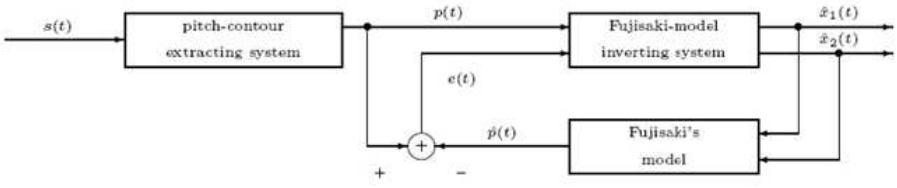



Figure 21: Representation of the prosody components extraction procedure.

Referring to figure 21 a phase of pitch contour extraction, in which the pitch trend is calculated and filtered to reduce noise effects, is followed by the calculation of the error $e(t)$ between the pitch contour and the system exit. The result is achieved by means of iterated passages.

### 2.2.3 Discussion

So far a presentation of the concepts of *segmental*, *non-segmental* and *super-segmental* acoustic information has been made in order to understand which are the most used speech units and events representations. A list of methods has been illustrated, which will be used in the rest of the work.
It is clear there are no features which can completely describe non-segmental or super-segmental information like prosody, or even speech units like syllables, which are on the border line between segmental and non-segmental acoustic events.

# 3 Chapter III: ASR Techniques Overview

## 3.1 Introduction

This section presents an overview of modern Automatic Speech Recognizers (**ASR**) building techniques. As previously said, humans can recognize speech by means of a complex interactions between multiple levels of processing, by using syntactic and semantic information, in combination with powerful processing and classification tools. The sophisticate algorithms developed nowadays, are not sufficient to hold the confrontation with what happens in the central nervous system, and speech is only one of the occasions in which the limitations of technology are evident.

Many knowledge types exist (e.g. linguistic, semantic, pragmatic), which could be integrated into an ASR, unfortunately the identity of all these aspects is not completely clear. The construction of an automatic speech recognizer having very high performances is a hard problem, which could also be not solvable at all.

ASR building has seen many realizations, which came from Artificial Intelligence, as in the case of the blackboards systems [63] [50] or expert systems based on human experience [7]. Some other a pure mathematical approach, as explained in Chapter I, using stochastic models like Neural Networks or Markovian models.

In expert systems much weight is given to euristic knowledge, while in mathematical models formally defined characteristics are searched.

Automatic speech recognizers are interesting even from a commercial point of view, in that such structures have been applied to automatic dictation systems. The structure of these systems is said to be "*speaker dependent*", because they address to a particular person, and no good performances are granted for other speakers. Other applications which use speaker independent systems are rising, especially in telephony applications, even if there are few examples of natural language recognition services.

In this section, the most used structure for speaker independent ASRs will be presented.

In the previous chapters the importance of the base unit of speech has been discussed, with an accent to many units integration. The starting point of common systems is the simplest one: the base unit of speech is assumed to be the phoneme. That's because the phoneme has the most identifiable structure.

The kind of signal which such systems address to, is the connected speech. In informal dialogues many pronunciation errors are made, noisy pauses, unguessable environmental alterations are present, which completely disturb the signal spectrogram. The detection of word separations, noise reduction and speech alterations is not an easy task and constitute a fundamental goal for such systems.
In the next subsection, a general overview of the architecture of classic ASRs is presented in detail, with a description of the mathematical models employed. A particular focus will be given to the Viterbi algorithm in order to emphasize the differences with a novel decoding strategy which will be described in 4.2.5.

## 3.2 General Architecture

The building process of an automatic speech recognizer is made up of two phases. The first regards system training, the second is, instead, the recognition stage. Figure 22 depicts the schema for the training phase. As can be noticed, this is a modular system which is able to train some stochastic models on the basis of prepared examples.

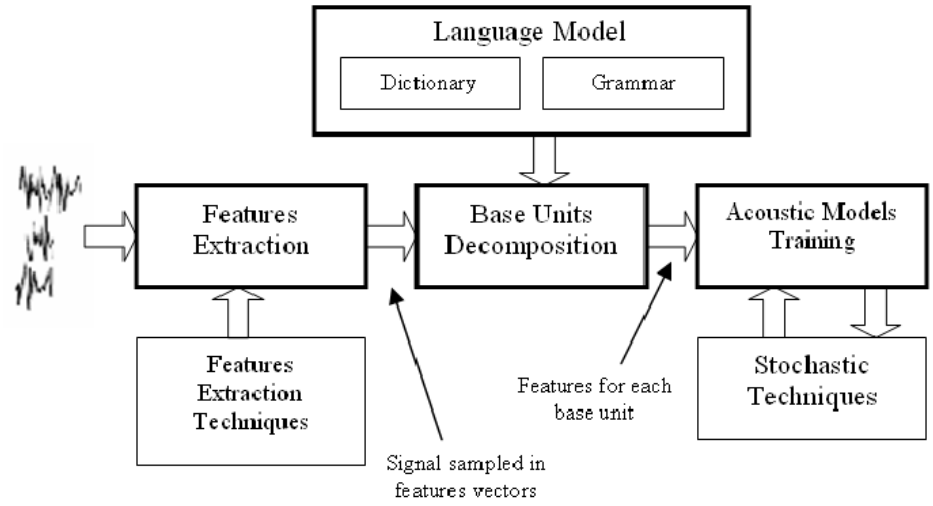



Figure 22: A classic ASR schema for the training phase.

The training procedure can be divided into the following steps

- Features extraction
- Language decomposition in speech units
- Acoustic models training

The next sections will describe the blocks in detail. Roughly speaking, the first module extracts the acoustic features referring to the chosen base unit. In the case of classic ASRs these are phonetic features like *MFCC* or *LPC* (ref. 2.2.1). The second phase divides the dictionary of all the recognizable words into a representation in terms of basic units concatenation. The third phase sets the stochastic models parameters to fit the training data, in order to prepare them for the recognition phase.
The recognition procedure, depicted in figure 23, can be organized as follows

- Features extraction
- Words decoding

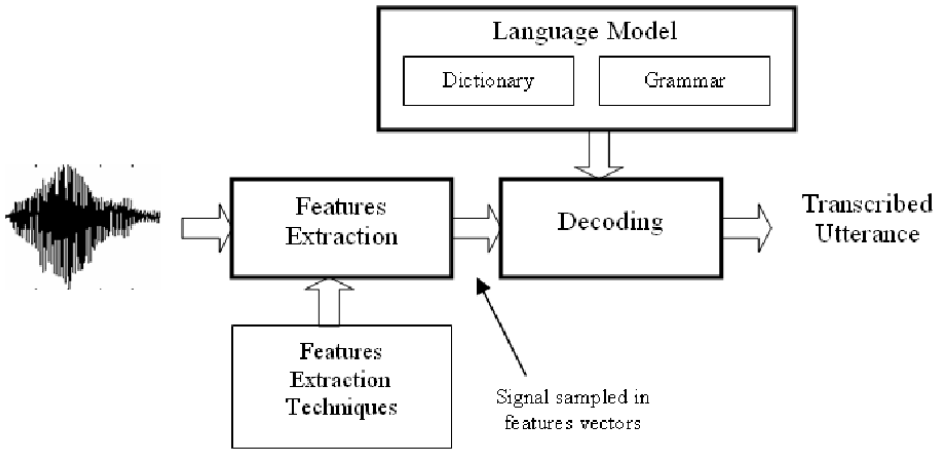



Figure 23: A classic ASR schema for the recognition phase.

Generally speaking, the first block extracts the units features from a speech signal, while the second phase combines stochastic models for units recognition and concatenation in order to reconstruct the uttered phrase.
The following sections will describe, in detail, the blocks and the difficulty of the problems they have to solve.

### 3.2.1 Features Extraction Module

This module deals with signal segments codification. Such problem has been largely discussed in the previous chapter. As classic ASRs need phonetic information, they make use of segmental techniques like *MFCC* or *LPC* (ref. 2.2.1), which address to phonetic characteristics, related to fundamental and formant frequencies.
The coding problem is crucial to an ASR, because the possibility to discriminate a speech unit from another depends on this phase. A concise and informative coding is the best possible, because it leads to a good recognition with the minimum training. The search for this kind of information is crucial for signal analysis techniques, but it is not easy to achieve.
A good unit representation is so a fundamental step to get high performances. Segmental techniques have been consolidated in time, so that, for phonetic ASRs, much of the focus has been set on the stochastic models of the back-end blocks.

### 3.2.2 Decoding Module

In this phase, units are concatenated to reconstruct the uttered phrase. This module is based on information coming from acoustic and language models. The first are the stochastic systems which try to associate a unit to a piece of signal, given a sequence of features, or, as in Bayesian models, try to calculate the likelihood of a sequence of features, given the model. Language model instead, deals with units concatenation probabilities .
The whole process can be so divided into two sub-modules

- The probability estimators for acoustic and language models
- The decoder

Both the steps are fundamental for ASR performances. The first calculates probabilities associated to pieces of signal and to units concatenations, the second is responsible for the combination of these two probabilities. The main aim of the process is to find the best sequence of units, according to concatenation probabilities and signal fitting estimations.
Acoustic models are generally based on Markov models (ref. 3.2.4) or Neural Networks [4].
The performances of such systems depend on

- The quality of the features
- The mathematical model employed to catch the dynamics of speech units
- The number of examples which are supplied in the training phase

The decoder is fundamental in equal measure. It makes use of a grammar, that is a specification about the probabilities of words concatenations, and of pronunciation models, that is a model which contemplates all the possible variations from the standard pronunciation.
In the next section the most common acoustic models, the *Hidden Markov Models*, will be described in details, while in section 3.2.6, the most used decoding technique, the *Viterbi algorithm*, will be illustrated.

### 3.2.3 Acoustic Models

Acoustic models are defined as models which are able to associate a sequence of features vectors to a speech unit, or which can calculate the probability of that sequence, given a model.
In ASR building, generally a mathematical model is realized which is able to classify an event. Formally, this means that the system has to be able to evaluate the probability that the event belongs to one of the possible units, and classify on the basis of the highest score.
The discriminative speech analysis techniques, which try to distinguish a speech event from another starting from the coding, have the aim to evaluate the *maximum a posteriori* probability for the event.
Such methods can be divided into *discriminative* and *non-discriminative*. The former calculate the *maximum a posteriori* (**MAP**) directly, that is

$$m^* = argmax_{|m} P(m|w)$$

Where $m$ is one of the classes involved and $w$ is the event to recognize, e.g. a sequence of features.
The latter, instead, face the problem by calculating the *maximum likelihood* (**ML**), that is the maximum probability $P(w|m)$ of the event, given the model.
The $MAP$ and the $ML$ are linked by the Bayes rule

$$P(m|w) = \frac{P(w|m)P(m)}{P(w)}$$

Neural Networks are generally employed in **discriminative** techniques, in that the representation power of such instruments allows the direct modelling of the MAP. In a Neural Network the output neurons represent speech units, while their outputs can simulate the probabilities that the network inputs belong to the class.
HMMs are generally employed in **non-discriminative** techniques. The algorithms for such systems evaluate the ML directly.
The vantage in using Neural Networks is in the higher control of the training and recognition phase, because an explicit calculation of the gap between the simulated function and the network and can be given.
The HMM approach is less controllable, in the fact that there is no error function which can be calculated. On the other side, the training phase is faster and generally more performant.
The fundamental difference is in the fact that while HMMs can exploit the dynamic process of features extraction, because the processing is synchronous to the production, Neural Networks generally address to yet collected feature vectors, as a picture had been made of a succession of speech segments.
Better performances in speech recognition have been given by hybrid approaches, which exploit the characteristics of Neural Networks in combination with HMMs [14]. Going slightly into details, Neural Networks are employed in modelling HMM states emission probabilities.
Figure 24 depicts such schema. The Viterbi decoder has the role to combine linguistic probabilities integrated into a HMM, with estimation probabilities coming from Neural Networks.

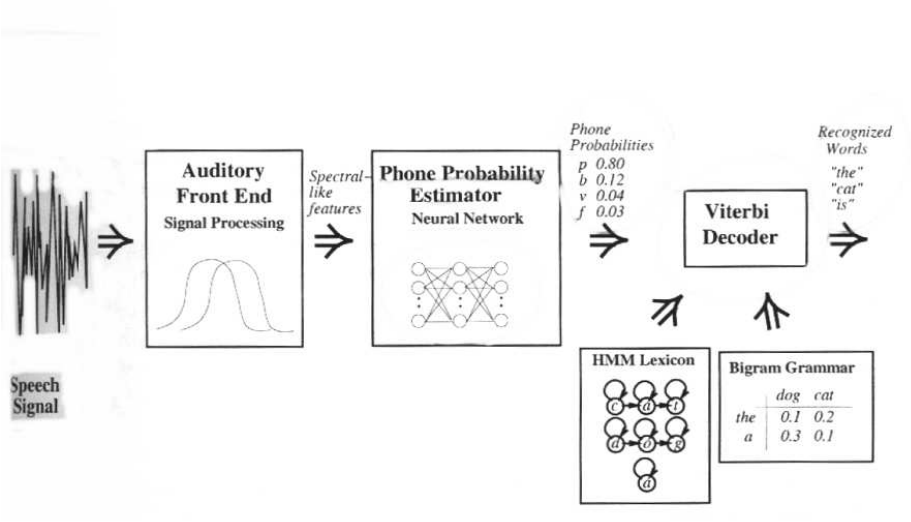



Figure 24: An hybrid ASR schema.

### 3.2.4 Hidden Markov Models

A Hidden Markov Model is a double stochastic model defined by the following elements

- A finite set of states, $S = \{s_1, s_2, .., s_N\}$
- A set of observable features which can be discrete or continue
- A transition matrix, $A = \{a_{ij}\}$, in which $a_{ij}$ is the probability of transition from the state $i$ to the state $j$. More formally $a_{ij}$ is defined as

$$a_{ij} = P(s_t = j | s_{t-1} = i),\ 1 \leq i, j \leq N$$

- An emission probability distributions set, $B = \{b_i(O_t)\}$, which is associated to the states set $S$. This is related to the probability that the $i$-th state emits the observed vector of features $O_t$ at the time $t$

- A set of initial probabilities $\Pi = \{\pi_i\}$where

$$\pi_i = P(s_1 = i), \quad 1 \leq i \leq N$$

  is the probability distribution for $i$ to be the initial state

Furthermore, the following relations must be valid,

$$a_{ij} \geq 0, \quad b_i(k) \geq 0, \quad \pi_i \geq 0, \quad \forall i, j, k$$

$$\sum_{j=1}^{N} a_{ij} = 1, \quad 1 \leq i \leq N$$

$$\sum_{j=1}^{N} \pi_i(k) = 1$$

and, in the case of $M$ discrete features

$$\sum_{k=1}^{M} b_i(k) = 1, \; 1 \leq i \leq N$$

For standard uses of HMMs, two assumptions are valid

- The **First Order Markov** assumption, which states that

$$P(s_t|s_1^{t-1}) = P(s_t|s_{t-1})$$

  where $s_1^{t-1} = s_1, s_2, .., s_{t-1}$ is the temporal state sequence.
  So, it is assumed that the transition probability from a state to another only depends on the first preceeding state in the temporal sequence

- The **Output independence** assumption, for which

$$P(O_t|O_1^{t-1}, s_1^t) = P(O_t|s_t)$$

  where $O_1^{t-1} = O_1, O_2, .., O_{t-1}$ is the temporal observations sequence made up of the features vectors. Such assumption means that the probability distribution of a particular observation, at a certain time, depends only on the current state

Given a temporal sequence of observations $O$, a HMM is able to calculate the probability that the model has generated it. Such value can be obtained according to the following equation

$$P(O|\phi) = \sum_{i=1}^{N} \pi_i P(O_1|s_i) \prod_{t=2}^{T} P(s_t|s_{t-1}) P(O_t|s_t)$$

Where $\phi$ is the HMM model.
The *likelihood* of the sequence given the model, is so the product between the probability of the first observation to be generated by one of the states, and the combined product of the emission and the transition probabilities.

### 3.2.5 The three base problems for HMMs

The three main problems associated to the HMMs are

- The *evaluation* problem. Given a temporal sequence of observations $O$ and a model $\phi$ , how to calculate the probability $P(O|\phi)$ of the observations given the model, with a treatable complexity?
- The *decoding* problem. Given a temporal sequence of observations $O$ and a model $\phi$, how to calculate the best sequence of states associated to $O$?
- The *training* problem. Given a temporal sequence of observations $O$ and a model $\phi$, how to change $\phi$ parameters in order to maximize the probability $P(O|\phi)$?

The evaluation problem is not always used in common applications, as it calculates the probability along all the possible sequences of states. Usually, the decoding problem solution gives the reference score that is used for speech units classification in ASRs. This is a variant of the evaluation problem and will be discussed in detail in the next paragraph.
The training problem prepares the model for the recognition session. It sets the models parameters in order to give the best performances on that observation sequence. This sequence is associated to the particular unit the model represents. E.g. if $\phi$ is the model for the syllable *ma*, then the observation sequence will be a sequence of feature vectors from a piece of signal where *ma* had been uttered.

### 3.2.6 Decoding problem solution: the Viterbi algorithm

Given a HMM model $\phi$ and a temporal sequence of observations,

$$O = \{O_1, O_2, .., O_T\}$$

the decoding procedure finds the most probable sequence of states, which the model have passed to produce $O$.
More formally, the sequence $S^*$ has to be found which maximizes the probability $P(S, O|\phi)$. This is a problem very close to the optimal path search in a graph, which makes use of dynamic programming techniques.
The probability of the best sequence is indicated as $V_t(i)$ and is the score associated to the best sequence of states till time $t$, which has generated the observations and terminates in the $i$-th state.
The value of $V_T(i)$ is the score of the best sequence terminating in the $i$-th state. So choosing the best among the states gives the highest total score. A backtracking procedure is able to reconstruct the sequence from the best final state.

The complexity of the Viterbi algorithm is obviously $O(N^2T)$.

**Algorithm 1** The Viterbi algorithm.

Step 1: Initialization

$$V_1(j) = \pi_i b_i(O_1) \quad 1 \leq i \leq N$$

$$B_1(i) = 0$$

Step 2: Induction

$$V_t(j) = max_{|1 \leq i \leq N}[V_{t-1}(j)a_{ij}]b_j(O_t) \quad 2 \leq t \leq T;\ 1 \leq j \leq N$$

$$B_t(j) = argmax_{|1 \leq i \leq N}[V_{t-1}(i)a_{ij}] \quad 2 \leq t \leq T;\ 1 \leq j \leq N$$

Step 3: Termination

$$P(O|\phi) = max_{|1 \leq i \leq N}[V_T(i)]$$

$$s_t^* = argmax_{|1 \leq i \leq N}[B_T(i)]$$

Step 4: Backtracking

$$s_t^* = [B_{t+1}(s_{t+1}^*)] \quad t = T-1, T-2, .., 1$$

$$S^* = (s_1^*, s_2^*, .., s_T^*)$$

### 3.2.7 Training Problem solution: The Baum Welch algorithm

The Baum-Welch algorithm [30] for HMMs training is the standard solution for such problem, it is based on the **Expectation-Maximization** paradigm, which won't be discussed here in detail.
Given the model $\phi$ and a sequence of observations, the parameters of the models have to be adjusted in order to increase the probability of the sequence given the model.
There is no analytic solution for such problem, and it only gets an euristic one, in that an error measurement cannot be calculated. The Baum-Welch is an iterative method in which, at each step, the value of the probability is granted to be equal or greater to its value in the previous passage, refer to [30] for further details.

The current discussion won't go into the details of the technique, as it is not fundamental in this framework, differently from the Viterbi algorithm which will be compared to a novel method in the next chapters. Anyway the general techniques is depicted in algorithm 2.

**Algorithm 2** A Baum-Welch algorithm summarization.

Step 1: Initialization of the HMMs parameters.

Step 2: E-M phase

the expectation function $E[P(O|\phi)]$ of the likelihood is maximized. An associated function $Q(\phi|\hat{\phi})$ is calculated, which produces the model $\hat{\phi}$ for which it results that

$$Q(\phi|\hat{\phi}) \quad \geq \quad Q(\phi|\phi)$$

Step 3: Iteration
the setting

$$\phi = \hat{\phi}$$

is made and the step 2 is re-executed. The iteration is carried on for a predefined number of passages.

## 3.3 HTK

HTK [5] is a toolkit for ASRs building, based on Hidden Markov Models, and it has been adopted in this thesis as the baseline system as well as prosodic recognizer. The main aim of the toolkit is to build HMMs based processes. Figure depicts the general schema of the recognizer.

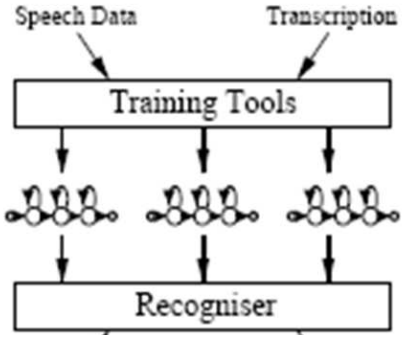



Figure 25: Schema of the HTK ASR.

As shown in figure 25, HTK is made up of two macro blocks, the first is a training tool, in which parameters of the HMMs are calculated using a language knowledge base, called *corpus*. This is a set of utterances with corresponding labels. The second process is the recognizer block, which takes utterances without labels as input, and produces their transcriptions according to the recognition process.

For the training phase, the Baum-Welch algorithm is used, while for the decoding phase, an adaptation of the Viterbi algorithm to continuous speech, called Token Passing [5], is employed.

The toolkit is able to recognize isolated as well as connected words and to produce $N$-Best lists, that are classifications of the utterances on the basis of the calculated likelihood.

HTK gives a good basis of comparison to novel models, because the whole system is able to simulate standard architectures. On the other side, the user can choose and customize the language model and the dictionary, so that also recognizers based on larger units can be simulated. The *prosodic* recognizer employed in the multi-granular model here presented, has been built on this methodology.

## 3.4 Standard Performances

A search on conference proceedings can depict a good panorama about systems performances in practical applications. Table 1 summarizes some of them.

| **Application** | **Accuracy** |
|---|---|
| Replacing Touch-Tone Menus | 99.5% |
| Call Classification and Routing | ~95% |
| Interactive Voice Responser | 90% |
| Desktop Dictation (speaker dependent) | 95% |
| Transcription of Broadcast News | 80-85% |
| Conversational Telephone Speech | 65% |
| Universal Voice Interface | ??? |

Table 1: Performances of common ASR applications [64].

Figure 26 reproduces, instead the evolution of the ASR performances in time[4]. As can be seen, performances highly depend on the task they face. Today common Interactive Voice Responsers (**IVR**) are able to manage a call with a customer, by means of DTMF or simple voice communication. Telephony constraints influence systems performances as well as the application of such instruments to a large public with different dialect inflections. Few examples can be found about call classification and routing, where the user is routed to a particular agent after he has "explained" his problem. Conversational agents, instead are far to come either for technology lacks or for social questions, because the success of an automatic system strongly depends on the grade to which people are used to that service. This affection varies from country to country, but the field is evolving fastly and IVRs are going through the direction of intelligent automatic agents.

[4]For a definition of *accuracy* refer to 7.2.

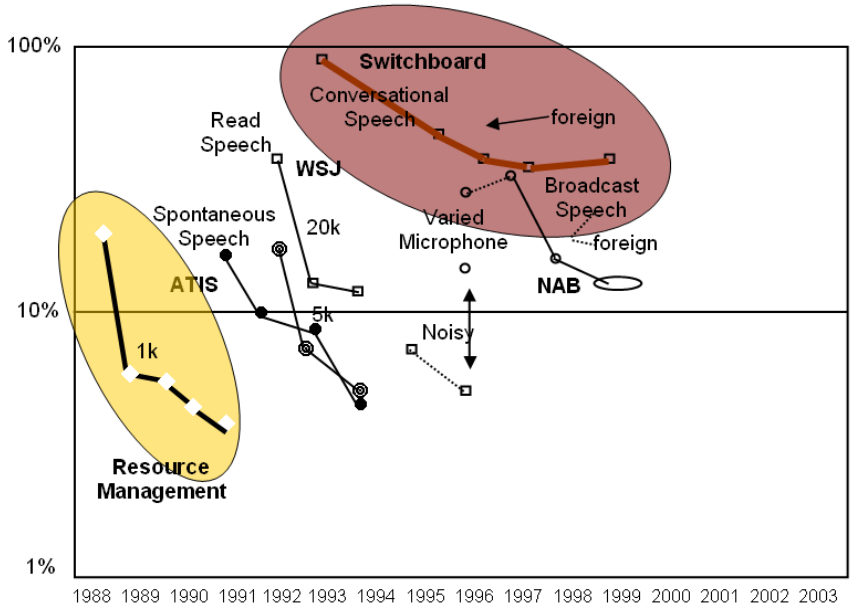



Figure 26: Performances evolution in time. Word Error Rate is represented on the y axis.

## 3.5 Discussion

In this chapter an overview of the most used techniques for ASRs building has been made. Phonetic and segmental recognizers have been described in their inner parts, as they present modules for training and recognition. The most used stochastic models, the *Hidden Markov Models* have been described, in particular the methods for calculating the likelihood of the model to a sequence of feature vectors and the training algorithm have been illustrated.
The overview, united to a schema of the performances on systems which employ this technology, is the basis for the multi-granular model presented in the next chapters. This constitutes the baseline the novel system will be compared to. Such choice has been motivated by the fact that the structure, particularly in its implementation with the HTK toolkit, is the simplest and basic one as well as being the most used in common applications.

# 4 Chapter IV: A Multigranular Segmental System

## 4.1 A fine-grain speech recognizer

In this section a novel method is presented, in which, keeping apart pragmatic, semantic and non-segmental features, the focus is strictly on the information coming from the signal in its *deep* details. The recognizer is made up of a *deep analysis* session in which acoustic models for syllables are used. This phase represents an implicit model for all the events of segmental nature, in an average period of a syllable (about 250 ms). It belongs to the class of systems which don't explicitly model all the concurring dynamics of the recognition process. It only addresses a fine processing of the signal which has to extract *hidden features* and dynamics from the acoustic observations.

The aim of this system is to create an **ASR** which is able to use information coming from two temporal analysis levels. The temporal scales analyzed are the phonetic and the syllabic ones. Wu [65] demonstrates that ASRs based upon only one of them separately, make complementary errors so that a joint recognition on the two scales of information can result in better performaces.

In this framework, a classic recognition model is modified about the acoustic model, in order to achieve a new structure. *Factorial Hidden Markov Model* (**FHMM** ref. 4.1.2) are employed, so that a new probabilistic model is built, which is able to directly catch information from the two scales. *MFCC*s (ref. 2.2.1) have been chosen for signal representation giving this approach the structure of an implicit modelling technique as described in section 1.5.

Figure 27 shows the schema of the ASR. As in standard models, the features extraction session represents the signal by means of standard MFCC features. A decoding process follows, in which FHMM syllabic acoustic models are combined with the language model to perform a recognition. The difference respect to a standard ASR is so in the acoustic models. A big importance is given to the training session, because FHMMs have the duty to automatically abstract two different dynamics from a syllabic piece of signal, one with a slow, and the other with a fast nature.

In the rest of this chapter a novel decoding procedure is presented, which is able to exploit FHMMs power at best. This has been done in order to focus on the model and performances rather than on Real-Time constraints. The loss in performances by standard ASRs systems belongs also to the decoding procedure,

which is not able to find the best solution for the alignment of the models to the speech signal, beacause of Real-Time contraints. The here presented solution is an exact alignment algorithm, which is a maximum point for ASRs that can be built with classic, Real-Time procedures, as will be furtherly demonstrated by experimental results.

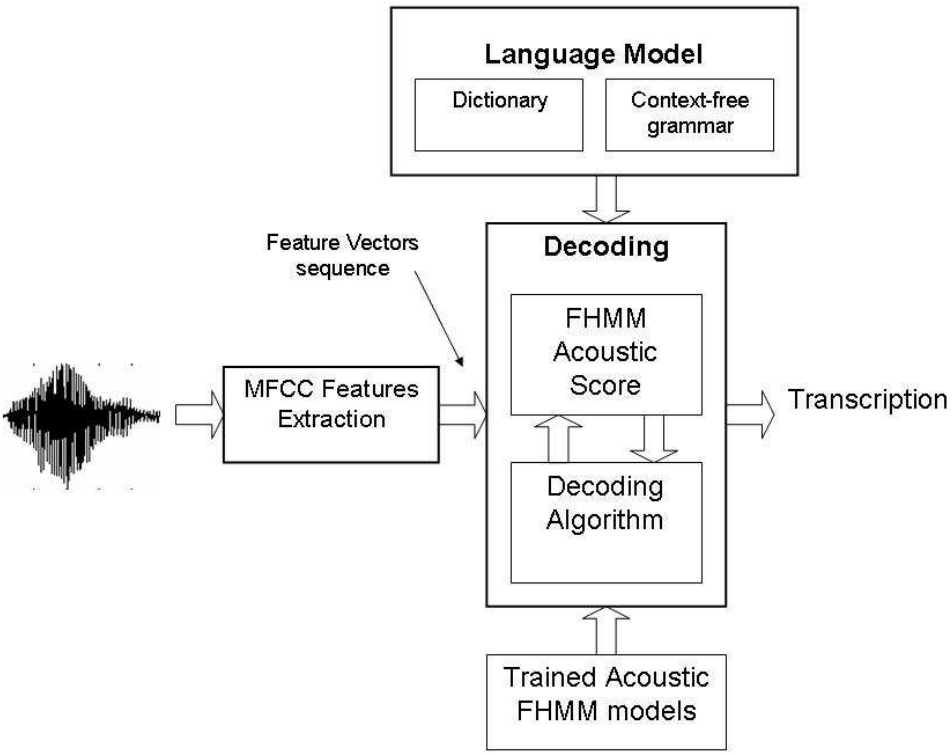



Figure 27: Deep processing recognizer schema.

### 4.1.1 The choice of the base unit

The model here introduced is meant to be *multi-granular*, in the fact that it puts together two levels of analysis, the syllabic and phonetic ones. The acoustic models employed in this framework try to extract two dynamics from a

speech signal. The first has a slow trend, and could be associated to slow evolutions of signal characteristics. The second dynamic has a fast trend and can be associated to fast information evolutions. These two phenomena could be identified as the syllabic and phonetic information lying in the signal, even if a demonstration for such correspondence is not easy. The here presented model has so the aim to extract such dynamics. The base unit of the resulting ASR will be set on the syllable, because of the necessity to extract also slow variations, which could not be detected in a brief temporal interval corresponding to a phoneme. This choice for a multi-granular system is good from the point of view of the acoustic models, which will so calculate the likelihood of the phonetic observations to a syllabic model. Instead, as it could be argued, this is not good from a language model point of view, in the fact that syllables have some counter indications:

- Too many models have to be used to cover a large vocabulary
- The pronunciation models are harder to build

The problems above are important for a theoretical investigation, but some practical considerations have to be kept into account.
Syllables include coarticulation phenomena which phonetic models are not able to catch. Giving the right examples during the training phase, can result in models to be able to recognize also altered structures. This means that the pronunciation models can be included into syllabic models, if they are trained on many cases. So the syllabic models can present an overall robustness to pronunciation variations, because these are included in the training.
Also the need for many models to cover the vocabulary can be discussed. A large number of models to be employed could result in system's slowness and difficulties in language representation. On the other side in English language, even if the complete dictionary is covered by over 30 000 syllables, only few of them are sufficient to cover the most part of it.
In figure 28 the Switchboard corpus [12] cover by english syllables is depicted. As can be noted **6000** syllables compose such corpus, but only **2000** are sufficient to cover the *95%* of it, and only **250** for the *75%*.

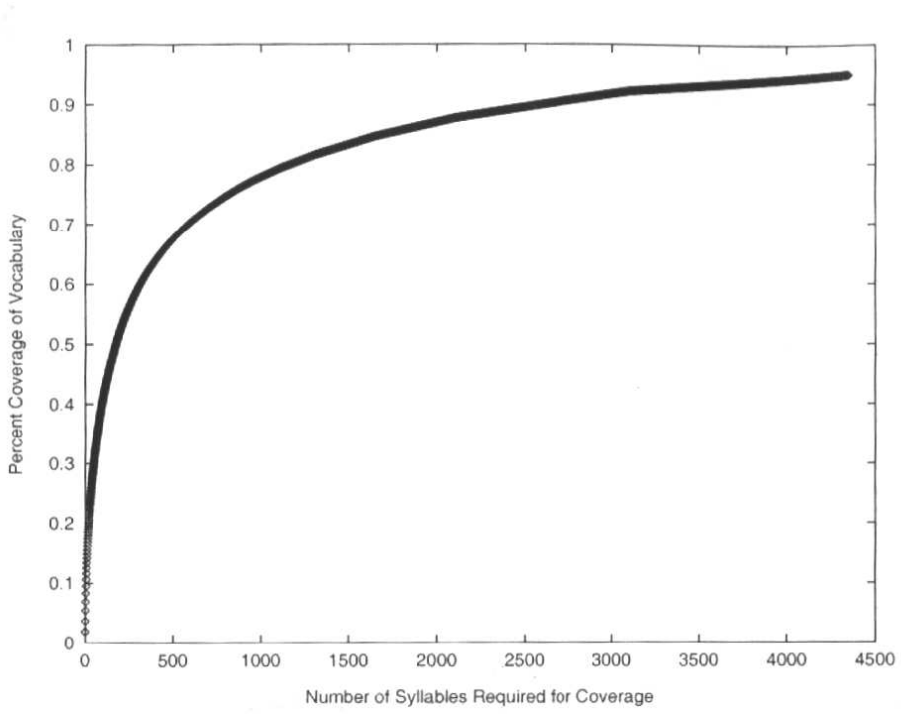



Figure 28: Syllabic statistics on the Switchboard corpus. From Wu [65].

Figure 29 shows that the most common words used in Switchboard corpus dialogues are composed by monosyllables.
In common speech dialogues, the syllables to be employed are only *little* more in number respect to phonemes. The difference is about one order of magnitude, but the overall recognition process can manage them with treatable complexity.

| $N$ | percentage of vocabulary | percentage of corpus |
|---|---|---|
| 1 | 22.39% | 81.04% |
| 2 | 39.76% | 14.30% |
| 3 | 24.26% | 3.50% |
| 4 | 9.91% | 0.96% |
| 5 | 3.21% | 0.18% |
| 6 | 0.40% | 0.021% |
| 7 | 0.057% | 0.0013% |
| 8 | 0.0052% | 0.000037% |

Figure 29: Percentage of syllables in vocabulary and corpus words. From Wu [65].

### 4.1.2 Factorial Hidden Markov Models

A *Factorial Hidden Markov Model* (**FHMM**), firstly introduced in [25], is a HMM whose state set can be decomposed in $L$ subsets. Each subset evolves independently as a standard Markov chain and they all contribute jointly to the observable variables generation, as shown in Figure 30.

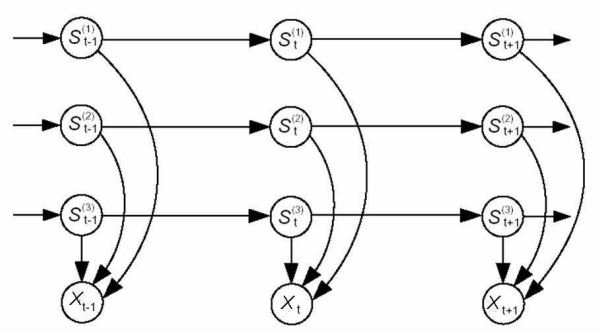



Figure 30: Factorial HMM dynamic from Jordan [25].

As stated in 3.2.4, in a standard Hidden Markov Model, a sequence of observations

$$X = X_1, X_2, ..., X_T$$

is modeled by specifying a probabilistic relation between the observations and a sequence of hidden states $S = S_1, S_2, ..., S_T$ taken by a finite set of states of dimension $K$. Moreover the model assumes that observations are independent of each other and, in many cases, that each $S_t$ is only dependent on $S_{t-1}$ (first order Markov property). HMM models are defined by the probability $P(S_{t-1}|S_t)$ of state succession, which is a $K \times K$ transition matrix and by the *emission* probabilities $P(X_t|S_t)$ which link the states to the observations. Such values can be calculated in many ways, in the case of continuous obervation vectors a gaussian mixture or a neural network can be used[6].
Factorial Hidden Markov Models expand the concept of HMM by representing a single state $S_t$ as a collection of $M$ states

$$S_t = S_t^{(1)}, S_t^{(2)}, ..., S_t^{(m)}, ..., S_t^{(M)}$$

each of which can take on $K^{(m)}$ values (for simplicity it will be assumed $K^{(m)} = K$ for all $m$). So, a FHMM consists of a state space which can be described by a $K^M \times K^M$ transition matrix. Such a system is equivalent to a HMM with $K^M$ states, and all variables are allowed to interact arbitrarily. The processing complexity is obviously exponential in $M$. Interesting phenomena come out when constraints are introduced in the state transition matrix. For what concerns the present application, each state variable $S_t^{(m)}$ is allowed to evolve according to its own dynamic, so that

$$P(S_t|S_{t-1}) = \prod_{m=1}^{M} P(S_t^{(m)}|S_{t-1}^{(m)})$$

Figure 30 depicts this structure. The transition between states can be represented as $M$ dinstinct $K \times K$ matrices.
About the emission probability of the observation $X_t$, instead, a gaussian distribution can be introduced, whose mean will depend on the $S_t^{(m)}$ states

$$\mu_t = \sum_{m=1}^{M} W^{(m)} S_t^{(m)}$$

where each $W^{(m)}$ is the contribution of $S_t^{(m)}$ to the mean. The covariance matrix length depends on the $X_t$ observation vector length.

$$P(X_t|S_t) \propto \exp(-\frac{1}{2}[(X_t - \mu_t)' C^{-1} (X_t - \mu_t)])$$

**FHMM**s have shown to be able to decompose automatically the state space into features differentiating multiple dynamics concurring in a single phenomenon. This is particularly efficient for cases in which the data are known to be generated from the interaction of multiple, loosely-coupled processes [25].
The idea here is that the multi-granular information in speech production, coming from syllabic and phonetic structure, may be thought as generated by overlapping processes with different time spanning, and a factorial model can catch these dynamics. The training can be able to associate different time-scale phenomena to different chains of the state set automatically. The layer nature of the model arises by only allowing transitions between states in the same layer. In this work only two levels of chains are used, such structure constitutes the acoustic model of the deep processing speech recognizer.

### 4.1.3 Applications of FHMMs

In ASR the factorial models has been used for the first time by [40]. In their work the authors use an acoustic model based on Factorial HMMs with two levels and three states for each level. They also use two different methods for the definition of the emission probability distribution $P(O_t|S_t)$, where $O_t$ is an observation vector and $S_t$ is a state of the factorial chain, according to section 4.1.2 notation.
The first method, which is called Linear Factorial HMM, is based on the idea of Jordan et al. [34], that the emission function is a multi-dimensional gaussian, while the second method, called Streamed Factorial HMM, use gaussian mixtures.
The Linear FHMMs give the best results, which are reproduced in table 2

| **Model** | **Word Error Rate** |
|---|---|
| Baseline HMM | 42.9% |
| Linear FHMM | 71.3% |

Table 2: Comparison between Standard HMMs and Linear Factorial HMMs using Cepstral features.

The Streamed FHMM performances, are instead reported in table 3

| Model | Features Type | Word Error Rate |
|---|---|---|
| Baseline HMM | Cepstrum + Delta Cepstrum | 42.9% |
| Baseline HMM | Cepstrum | 51.6% |
| Baseline HMM | Delta Cepstrum | 62.3% |
| Streamed FHMM | Cepstrum + Delta Cepstrum | 46.3% |

Table 3: Comparison between Standard HMMs and Streamed Factorial HMMs using Cepstral features.

Another application of the factorial model can be found in Duh [19], which employs them in Part of Speech tagging. He takes two kind of information, lexical and morpho-syntactic. The dynamics correspond to two levels of tagging and to two FHMMs layers. Respect to the classical model, Duh introduces a dependency between the states with the same index, e.g. $s_i^{(1)}$ and $s_i^{(2)}$. Furthermore he adds also dependency between adjacent states, e.g. between $s_i^{(1)}$ and $s_{i+1}^{(2)}$, and between $s_i^{(2)}$ and $s_{i+1}^{(1)}$.
The performances get an absolute **2%** increase in performances respect to a state of the art system.
Another application of Factorial HMMs can be found in [31], in ovelapped voices separation. In this case the superpositions are considered as concurrent processes to be separated.

## 4.2 Implementation Details

The studies by Jordan [25] have stated that Factorial HMMs are able to catch a different dynamic at each level, during the same process. In the case of syllable modelling, the dynamics to take into account are two: the phonetic and syllabic one. The number of levels will necessarily be two. About the number of states for each level, it has been set on the basis of experiments on classic HMMs. Standard models have given the best results with 7 states, when using MFCC features. So it has been decided the factorial model to have the same number of states for each level, which has been set to 7.
To better follow the production of the feature vectors, each chain has been defined and set as a Bakis model [1]. Such model does not allow a HMM to have backward transitions. The only possible ones are the self loops or forward

connections. Imposing a Bakis structure to a Factorial HMM means to allow only left-to-right process for each single chain in the layer.

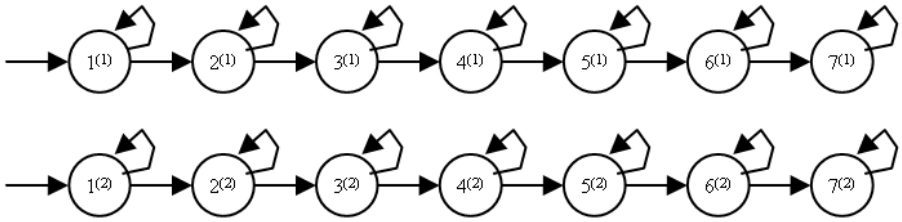



Figure 31: Factorial HMM with Bakis structure representation.

The probability transitions of the states will be initialized as

$$a_{ij+1}^{(m)} = a_{ij}^{(m)} = 0.5, \quad 1 \leq i \leq 6 \; m = 1, 2$$

$$a_{77}^{(m)} = 1 \quad m = 1, 2$$

furthermore

$$\pi(1^{(m)}) = 1 \quad m = 1, 2$$

$$\pi(i^{(m)}) = 0 \quad 1 \leq i \leq 7 \; m = 1, 2$$

Figure 31 represents the schema of the acoustic models employed in the segmental recognizer.
In the next sections the details of the training and recognition processes are explained.

### 4.2.1 FHMMs Training

Even in FHMMs the Expectation-Maximization (**EM**) algorithm is employed, but in the variant introduced by [25]. The algorithm is like the one described in section 3.2.7, and can be divided in two steps: the first *expectation* (E)

stage, fixes the current parameters and calculates the emission probabilities for the states. The second phase, *maximization* (M), uses those probabilites to maximize the likelihood of the observations to the model.
The calculation of the M step does not give problems, and has the same complexity as a standard HMMs. The problems arise in the E step, which could have an untreatable complexity, in the case of many levels. A FHMM having M layers and K states for each level is equivalent to a standard HMM with $K^M$ states, so referring to the Baum-Whelch algorithm, the complexity is $O(TK^{2M})$.
To overcome this problem, approximated methods can be used, like

- The Montecarlo method [62]
- Gibbs sampling procedure [24]
- Completely factorized variational inference method [33]
- Structured variational inference method [32]

An overview of all those methods is presented in [26].
In the here presented model, the number of levels was narrow enough to allow the use of the exact training procedure. The Factorial HMM is firstly exploded into a single standard HMM with $K^M$ states, so that the classic Baum-Welch algorithm can be used. This procedue makes the cartesian product of all the possible couples of states.
The transition probability from a couple to another is calculated as the product of the single transition probabilities, as shown in figure 32.

$$P(s_i^{(1)}s_k^{(2)}|s_j^{(1)}s_g^{(2)}) = P(s_i^{(1)}|s_j^{(1)})P(s_k^{(2)}|s_g^{(2)}) \quad 1 \leq i,j,k,g, \leq 7$$

The initial probabilities are set to 0 except for the state $s_1^{(1)}s_1^{(2)}$, while the only final state is $s_7^{(1)}s_7^{(2)}$.
The training function starts from a succession of the observations vectors containing *MFCC* parameters, which represents a syllable. Obviously the number of vectors is variable, according to the length of the syllable. Each vector refers to a 20 ms speech segment.
In the realization of the algorithm used in this thesis, the training is arrested after a maximum number of 100 iterations or if, after a while, the Maximization step does not produce sensible improvements.

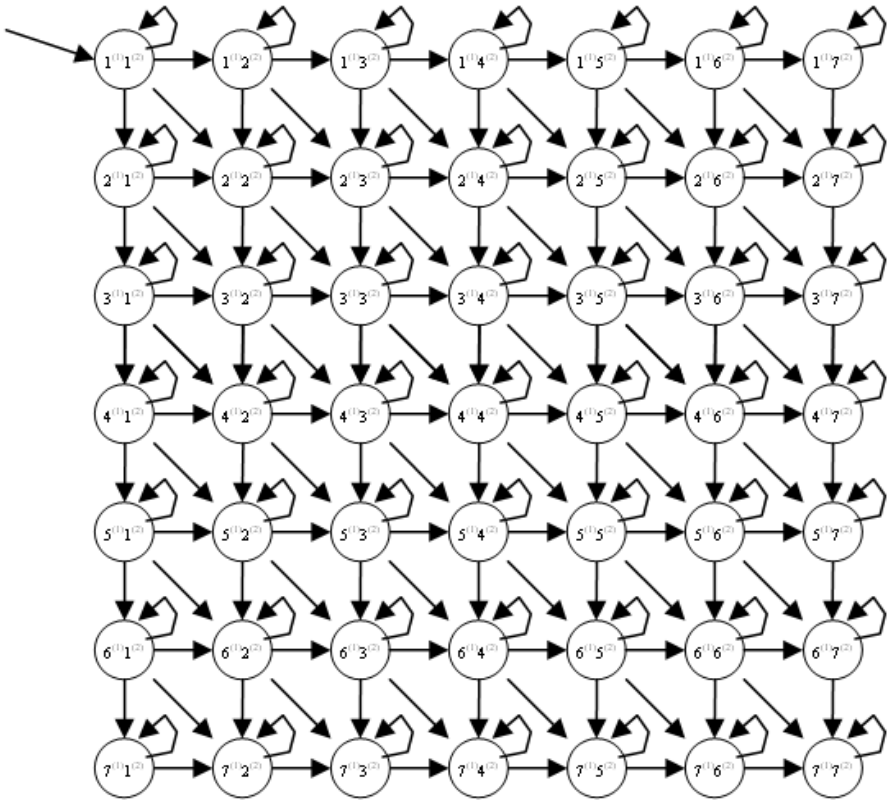


Figure 32: FHMM expansion. Each couple-state comes from the cartesian product of two states of the FHMM. The state with label $x^1y^2$ belongs to the product of the $x$ state of level 1 and the $y$ state of level 2.

### 4.2.2 Likelihood Calculation

The Viterbi algorithm (ref 3.2.6) has been used for the calculation of the likelihood of the observations to the model. This procedure uses the standard algorithm on the expanded HMM obtained from the original Factorial model. Figure 33 depicts the likelihood logarithm vs the number of frames provided to the model for the syllable *"di"*. The log-likelihood trend increases at each frame if those features refer to the syllable the model represents. If this is not the

case, then the HMM based recognition does not guarantee that the likelihood decreases. The ASR logic is that the likelihood must be highest for the right model.
Figure 34 depicts the log-likelihood for the word *"qua"* on the obervations sequence for the syllable *"di"*. Even in this case the quantity is higher and higher, but in the end it reaches a lower value respect to the *"di"* model.
The algorithm has a computational complexity of $O(S^2T)$, where $S$ is the number of states in the expanded HMM and $T$ is the number of observation frames. The treatability strictrly depends on the number of states for each layer of the Factorial HMM.

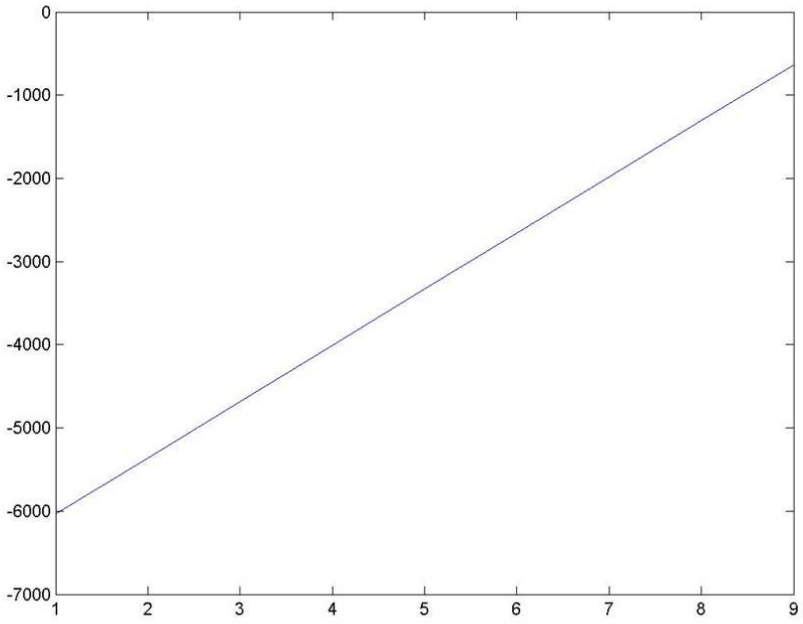



Figure 33: Log-Likelihood trend for the model of the syllable *"di"* , on frames by a *"di"* utterance.

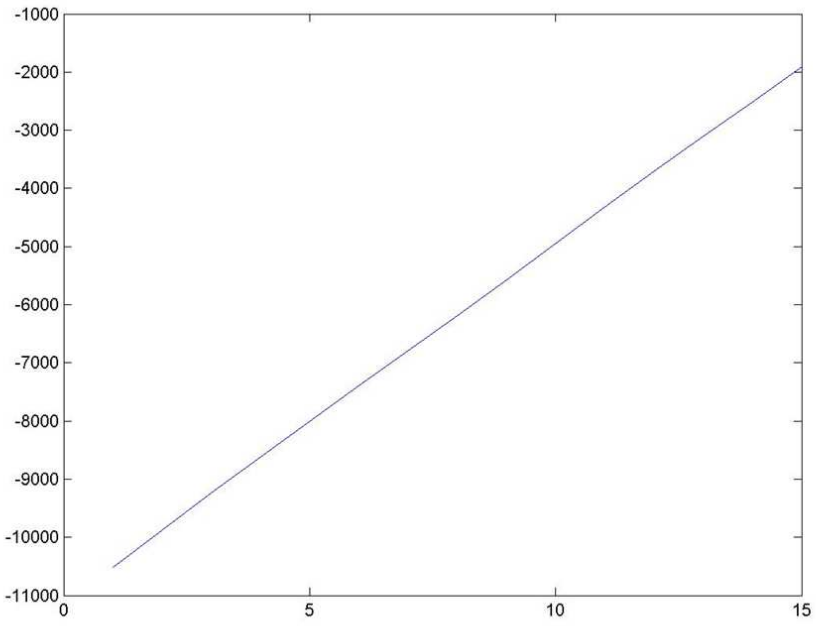



Figure 34: Log-Likelihood trend for the model of the syllable *"qua"* , on frames by a *"di"* utterance.

### 4.2.3 The Silence Model

The silence detection is an important feature for an ASR, because pieces of silence should not be given to acoustic models. Such method avoids the possibility of an initial piece of silence, or inter-word pauses, to be recognized as a words.
The silence model has been considered a further unit, similar to a syllable, but different in the meaning, for which a Factorial HMM has been created. It presents a Bakis structure too. In other realizations [5], the silence model presents a simple three states structure, where also a connection between the first and the last state is allowed. This structure, has given bad perfromances in the present model, especially in initial long silence deletion. So the employed model uses a 2 layers Factorial HMM with 7 states for each layer.

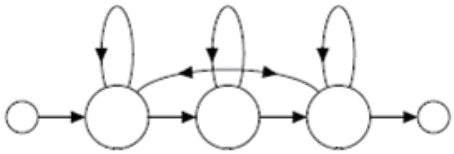


Figure 35: Representation of a Standard HMM for silence model [5].

### 4.2.4 The Language Model

The aim of the language model is to associate a probability to a sequence of words or, more generally of units. If $W = w_1, w_2, .., w_n$ is a sequence of units, the language model calculates the probability of such sequence by the following equation

$$\begin{aligned} P(W) &= P(w_1, w_2, .., w_n) = \\ & P(w_1)P(w_2|w_1)P(w_3|w_1w_2)\,..\,P(w_n|w_1, w_2, .., w_{n-1}) = \\ & \prod_{i=1}^{n} P(w_i|w_1, w_2, .., w_{i-1}) \end{aligned}$$

For complexity needs, the backward dependency is often limited to only $N$ preceeding units. The resulting language model is said to be based on $N$-grams. The choice here has been a *bi*-gram language model, which stores the probabilities $P(w_i|w_{i-1})$ of the concatenation between the syllable $w_i$ and its immediate preceeding. The language model score will be so calculated as

$$\begin{aligned} P(W) &= P(w_1, w_2, .., w_n) = P(w_1)P(w_2|w_1)P(w_3|w_2)\,..\,P(w_n|w_{n-1}) = \\ & \prod_{i=1}^{n} P(w_i|w_{i-1}) \end{aligned}$$

The set of syllables used in the current experiment has been taken from the numbers from *0* to *999,999,* so that the language model has been built upon those words. In that case the introduction of a probabilistic grammars was not necessary. The above formula for probabilities calculations are still correct even if they are calculated on exact estimations. Chapter VII explains the details of the implementation.

### 4.2.5 The Decoding Algorithm

In order to exploit the model at best, an efficient algorithm for syllable decoding had to be developed. Standard algorithms usually act in real time using dynamic programming methods and some approximations (as in the case of the beam search algorithm) with the aim to reduce execution time. These procedures can introduce many errors as the recognition is strongly dependent on the left-to-right time processing. A more efficient procedure could try any possible alignment between words and signal, as it could retreat some decisions made during the left-to-right processing, and could try to shift the models backwards or forwards to achieve the best alignment and words separation. This procedure has been developed using dynamic programming. It does not act in real time because of the request to be independent from the signal runtime generation. This is undoubtedly an high complexity procedure, however it can drive acoustic models performance at best. The main aim of the algorithm is to navigate the structure formed by the union of the language and acoustic model in order to maximize the probability $P(W|X)$ for a sequence of units (syllables in this case) $W = w_1 w_2..w_n$ given the observation sequence $X = X_1 X_2..X_T$.
$P(W|X)$ could be calculated as follows

$$P(W|X) = P(w_1 w_2..w_m|X_1^t)P(w_n|w_m)^\gamma P(w_n|X_{t+1}^T)$$

where $w_n$ is the last unit if $W$ is not empty and $w_m$ is the preceeding syllable in the sequence. $P(w_n|w_m)$ is the language model probability between $w_m$and $w_n$, $\gamma$ is the language model weight, and $t$ is the optimal time boundary between the units. Lets demonstrate that if $P(W|X)$ is the optimal solution for the units alignment problem, then $P(w_1 w_2..w_m|X_1^t)$ is the optimal solution for the problem of units alignment in the time interval [1,$t$], where $t$ is the best first boundary for $w_n$. This is trivial in the fact that if there was another sequence $w_1^{'} w_2^{'}..w_m^{'}$ for which $P(w_1^{'} w_2^{'}..w_m^{'}|X_1^t) > P(w_1 w_2..w_m|X_1^t)$ then it would be

$$P(w_1^{'} w_2^{'}..w_m^{'}|X_1^t)P(w_n|w_m^{'})^\gamma P(w_n|X_{t+1}^T) > P(w_1...w_m w_n|X_1^T)$$

against the hypothesis of $P(W|X)$ to be the optimal solution for the problem. This discussion leads us to introduce the following recurrence relation for the solution $f(m,t)$ to the subproblem of units alignment in the time interval [1,$t$]

$$f(m,t) = max \begin{cases} P(X_1^t|m)\pi(m) \\ max_{|1 \leq t* < t, n \in Syl}\{f(n,t*)P(m|n)P(X_{t*+1}^t|m)\} \end{cases}$$

where *Syl* is the set of all the units involved, P($m|n$) is the probability of $n$

and $m$ unit concatenation, $P(X_1^t|m)$ is the likelihood of the model $m$ to the observations $X_1X_2..X_t$, and $\pi(m)$ is the probability for $m$ to be a starting unit for a sequence. Notice the dependency from $f(n,t^*)$, which is the best solution to the subproblem of units alignment till time instant $t^*$. The optimal solution will be retrieved as follows

$$P(W|X) = max_{|m \in Syl}\{f(m,T)E(m)\}$$

Where $E(m)$ is the probability for the model $m$ to be a plausible ending unit. Starting from this solution, a backtracking procedure produces the best alignment. The algorithm is also based on the calculation of the matrix $V$, which contains the likelihoods of a model $m$ to all the intervals of observations

$$V = \begin{pmatrix} P\left(X_1^1|m\right) & P\left(X_1^2|m\right) & \dots & P\left(X_1^{T-1}|m\right) & P\left(X_1^T|m\right) \\ -\infty & P\left(X_2^2|m\right) & \dots & P\left(X_2^{T-1}|m\right) & P\left(X_2^T|m\right) \\ \dots & \dots & \dots & \dots & \dots \\ -\infty & -\infty & \dots & -\infty & P\left(X_T^T|m\right) \end{pmatrix}$$

The algorithm complexity is $O(T^2N^2C(V))$, where $C(V)$ is the complexity of the likelihood calculations for a single model. If $S$ is the number of states in the acoustic model, then $C(V) = O(S^2T)$. This value can be reduced by considering that using controlled (even if connected) speech, a single syllable can rarely have a maximum duration greater than a fixed values (e.g. *500* ms). At this point, the matrix $V$ will get a band aspect which allows optimization about complexity issues.

If observations are taken every 10 ms, then we can calculate the likelihoods only on 50 observations intervals leaving to zero longer span probabilities. The complexity of this algorithm seems to be quite high for practical applications, especially if the utterance is too long, but it leads to an optimal alignment. Tests have stated that, on an AMD 2800+ processor, the response is about 30 seconds, for a 3 seconds utterance, after the recording stops. Efforts should be fronted in the next future to improve this performance.

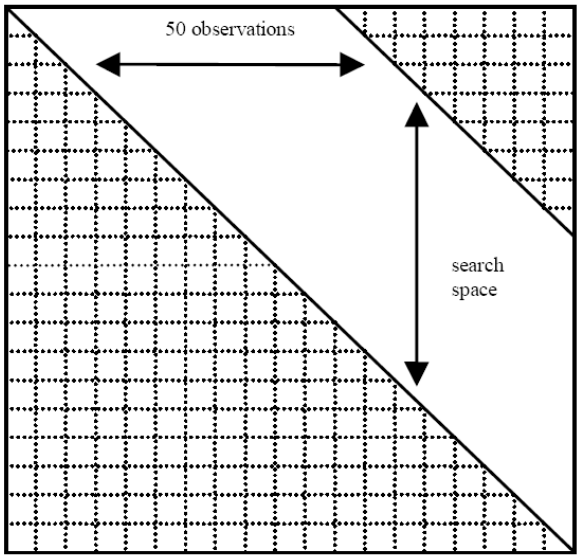



Figure 36: Band matrix for complexity reduction based on assumptions about syllable length.

## 4.3 Discussion

A novel approach to segmental multi-granular recognition has been introduced[8]. An ASR employing Factorial HMMs and a new decoding algorithm has been described. Factorial HMMs have the power to extract information coming from ovelapping dynamics and use them in likelihood calculation. In the here presented model, FHMMs constitute the acoustic models, whose sequence is managed by a language model. The perfomances of such system will be presented in chapter VII, where the property of dynamics separation will be confirmed by experiments on numbers recognition. The model outperforms standard segmental ASRs because the use of Factorial HMMs with syllabic acoustic model is able to catch multiple information lying in a succession of fine features addressing to phonetic characteristics. This is the first layer of the whole multi-granular model here presented. The further information that will be added, belongs to signal analysis, but with super-segmental nature. The next chapter clarifies such approach.

# 5 Chapter V: A Non-Segmental Speech Recognizer

## 5.1 Non-Segmental Recognition

Starting from the assumption we have a speech signal containing at least a single complete word (without fragmentation), we can try to find a rough description which can help the system in guessing a set of possible identities for the word, without going deeply into the analysis of the segment. If we had been in presence of a unit like a phoneme we could have found a set of features with a certain amount of discriminant power. In some experiments words have been modeled as sequences of phonetic features but no effort has been made towards the discovery of new features addressing a whole word unit. In the model here presented, a set of *tracts* is introduced which could entirely characterize a quasi-syllabic unit, which is harder to describe than a phoneme, but is simpler than a word. A choice has been made among non-segmental features, and eventually *prosodic profile*, *vowel configuration* and *syllabic borders* have been chosen to define a feature set which can have a rough discriminant power. The idea is to build up a recognizer using a little amount of the computational effort required by a deep analysis involving segmental recognition, which does not aim to find a precise matching for a word, but instead to reduce the number of possible candidates. The recognizer proposed, makes use of a stochastic model in order to associate a set of features, extracted from the segment, to a set of syllables in the vocabulary. In order to choose the proper set of features that could describe an entire syllable, some considerations have to be made. Referring to Iwano et al. [22], an interesting feature is the derivative of the pitch curve. In [22] syllabic HMMs act on both phonetic and prosodic features on a connected digits dictionary. The experimental results show an absolute improvement of about *4.5%* with a signal-to-noise ratio of *20dB*. Delta pitch is used in combination with phonetic features to achieve the result, while the stochastic model is what they call *multi-streaming*[5], which reduces to the Factorial HMMs described in section 4.1.2, when the covariance matrix is diagonal. Following the discussion in section 1.6 , this is an *implicit* approach, because the issue of extracting a model for the phenomenon is deployed to the stochastic learning session. The sole delta pitch feature is obviously not enough to discriminate among a set of words, so some other information has to be introduced.

[5]The concept of *multi-streaming* they introduce in this framework is what is meant here by *multi-granular*.

The energy profile of a word can indicate some areas corresponding to vowels or sonorant consonants. The derivative of such profile could do the same work of the pitch trend to suggest how the word has been pronounced. Such an information can be another feature for a word model, e.g. "*uno*" and "*due*" can be separated istantaneously by only means of pitch and energy derivatives. As can be guessed, such set of features is too rough for performing a good recognition. Word more complex than monosyllables can be easily confused, e.g. "*kwattro*" and "*tsinkwe*" can have very similar profile in some situations. A furtherly bit more specific information has to be given. The *vowel configuration* is a good candidate for such a task. The term refers to the sequence of vowels included in the word, their distribution among the syllables involved and the distance between them. As said above, vowels can be a strong aid in words discrimination and can be suitable for a prosody based level of a multi-granular model. The process of vowels identification in a speech segment has not to be carried on by a complex analysis as in the case of phoneme based recognition. The recognition process follows a pitch analysis and the vowel identification is made only on areas where the pitch has a slow variation and the signal has a strong energy. So the computational complexity is mantained less than that of the deep analysis level.

Furthermore some other information can be added in order to make the system more performant. Remind that the aim of this phase is not to build a very performant recognizer, but only to aid the deep analysis level in the recognition process. The process of dictionary pruning has to be able to reduce the number of candidate words even of some orders of magnitude.

Other useful information coming from prosodic analysis is the set of *voiced parts*, *stressed segments* and *syllables profile*, which are recognizable phenomena and can be discriminant features.

Such a set has to be translated to some word succession taken from the dictionary, but a stochastic model has to be used because of the high speech variability even in prosodic aspects. A Hidden Markov Model can be suitable for such a purpose. The output will be constituted by an $N$-Best hypotheses list for the candidate words.

This kind of recognizer is able to perform a rough recognition of some segments using a kind of information which comes from prosody and is so complementary to that of the deep syllabic or phonetic analysis. Vowels, and other information about phonetic tracts enrich high level information and can make the system more perfomant. This is the first part of the multi-granular model presented,

which has to be integrated with the second kind of process, that is more complex in computation but also more discriminant. The merging session between the upper and the lower level will be discussed in later section, where the role of the rough recognizer here presented will be properly clarified.
A second approach has been attempted, following the trend by other experiments. Prosody can be used in post-processing phase, when a speech recognizer has produced a list of the first $N$ best phrases, according to the score coming from the decoding process. Prosody can be a good method to rescore such list and get the correct solution. Vergyri et al. [54], as explaind in 1.6 have worked in this way. The authors have inserted prosodic analysis at multiple levels, but alway basing on a fine-grain recognizer. According to them, prosody is a feature that can be introduced as a feature either at the acoustic lever or at the language level. Also the work by Hirose [52] is an example of *bottom-up* approach, in that prosodic information is used after the phonetic recognizer computation. In the here presented model, two methods for prosodic information use in ASRs will be proposed. The first is an attempt to build a stand-alone recognizer, which is only based on prosodic features. The second is a procedure, to be attached to a fine-grain recognizer, which makes a decision on a syllable segment of speech to mark its prosodic "*coherence*". This will be inserted in a rescoring module for a segmental ASR.

## 5.2 A Top-Down Prosodic Recognizer

### 5.2.1 Description

The prosodic recognizer, has the aim to output a list of phrases as candidates for the utterance transcription. The ASR produced in this phase cannot be able to perform as well as a phonetic or syllabic one, because

- Prosodic information is not suitable to ASR classical structure
- Commonly used feature coding, does not completely catch the phenomenon
- Prosodic features do not contain full discriminative information

The ASR here presented has been created to build an initial stage, in which the whole space of solutions can be reduced by means of a rough prosodic

recognition. The system starts from prosodic and non-segmental features in order to catch all the possible information to help an underlying recognizer. Prosody is meant to be a substitution of the fine-grain recognizer only in the case in which the problem is trivial. Referring to de Saussure [18], the idea is that the power of the fine analysis has to be used only in non-trivial cases. The model here proposed has been developed by means of standard HMMs, because it has to be fast and has not the requirement of producing an exact recognition. Two problems have been faced, the first is the choice for the acoustic models. Here the basic unit of speech is supposed to be *prosody*, an entity without a precise formal definition and which is distributed along the whole utterance. Building a recognizer for such phenomenon using a classic structure, means that a forcing has to be made.
The second issue is about the representation of prosody. In section 2.2.2 a discussion has been made about common methods to formally describe prosodic characteristics. The description does not completely define the phenomenon, in the fact that other information is present in the signal, even if some parts of it are still unknown. In the present realization, all the methods described in section 2.2.2 have been used together in order to catch as much as possible information. Redundancy has been useful in increasing system's performances.

### 5.2.2 The choice for the Base Unit

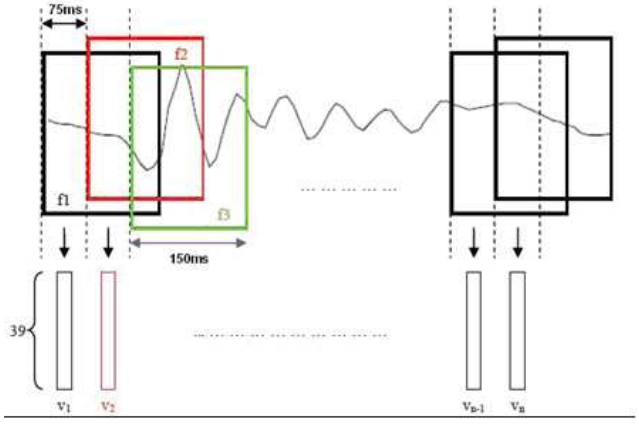
The choice for the Base Unit of speech has alway been searched among segmental or quasi-segmental entities. In the present case such choice is not easy, because prosodic information is distributed along the whole signal. Units representation is crucial to an ASR, and usually refers to a succession of feature vectors, each referring to a speech segment of a certain length.
The choice made in this thesis has been of two types: the prosodic phenomena will be extracted from a piece of signal much larger than a phoneme. Speech segments of *150* ms have been chosen, with *75* ms superpositions. This is because the most part of the methods for features extraction refer up to such length, as in the case of the Modulation Spectrogram (ref. 2.2.2), which catches the slow variations of the spectrum.
Prosody has been modeled as a succession of long range features, in which the slow modulations, united to pitch and energy trends, to accent and phrase component, and to vowels indications are mixed together. The forcing here is in the

fact that, even if prosody is not a phenomenon identifiable in a single speech segment, the succession of long range features could catch it into an ASR structure. The vantage here is that a standard ASR using wide analysis windows, has a much lower complexity, because less vectors are needed to represent an utterance.

On the other side, an acoustic model has to be used in order to associate such successions of features to a speech unit. Theoretically the choice would have required a direct correspondence between the vectors and an utterance, in that prosody cannot be divided into classes. Obviously this cannot be possible for computational and language representation reasons. Having an acoustic model for each single word in the dictionary, would have implied to build up a huge number of models which could have made the fine-grain processing less complex. On the other side, features describe too long units to be catched by a phonetic model.

Syllables have been chosen as acoustic models for the ASR. In this case, the classification is separated from the representation, because the model has to recognize syllables from non-syllabic observations. Some experiments [11] have demonstrated that features like the Modulation Spectrogram are able to build up a syllabic recognizer, which does not get the performances of a phonetic recognizer, but is able to catch some discriminant information. The next section explains the details of the resulting ASR.

### 5.2.3 The ASR




The features employed in this thesis, are the ones described in 2.2.2. As the

Figure 37: Representation of the features extraction process in the Prosodic Recognizer.

figure above depicts, the analysis is taken on *150* ms windows overlapped by *75* ms. For each segment, a vector of features is extracted, containing the element resumed in table 4.
The meaning of the features has been explained in section 2.2.2.

| Features | # | Features | # |
|---|---|---|---|
| Energy+Δ+ΔΔ | 3 | Modulation Spectrogram | 18 |
| Diff. Energy+Δ+ΔΔ | 3 | Accents Markers | 1 |
| Pitch+Δ+ΔΔ | 3 | Vowels | 1 |
| Voiced/Unvoiced | 1 | | |
| Fujisaki Rectangles | 3 | | |

Table 4: Features employed in the Prosodic Recognizer. The right column reports the corresponding number of parameters, for each feature.

As can be argued from the table, a single 150 ms segment of speech is represented by 33 prosodic features.
The first phase of the ASR building has interested the configuration of the acoustic models.

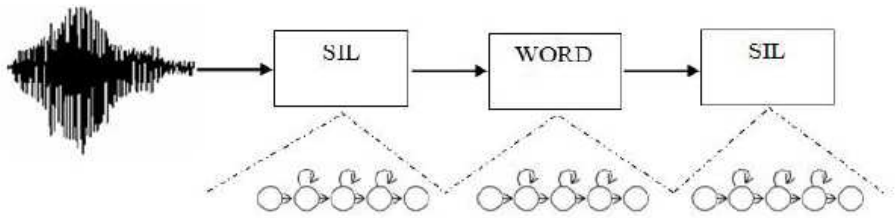



Figure 38: Representation of the complete model for a word.

The emission probability function has been investigated, in order to get the more performant acoustic models configuration. Experiments have been made varying the number of states and gaussians for emission probability simulation.

Table 5 reports the results on the recognition of numbers ranging from *0* to *999,999*.

| # States | # Mixtures | # Features | Sentence Correctness |
|---|---|---|---|
| 4 | 3 | 32 | 28.64% |
| 4 | 3 | 36 | 30.05% |
| 4 | 3 | 36 | 37.52% |
| 5 | 4 | 36 | 33.86% |
| 6 | 3 | 36 | 61.74% |
| 6 | 0 | 36 | 30.52% |
| 6 | 10 | 36 | 30.40% |

Table 5: Reporting of the recognizer scores, at the variation of the number of gaussian mixtures, features and states for the acoustic models.

The number of features has been changed too, in order to understand if the carried information, was completely redundant. The results show that the best configuration uses all the features, 6 states for each acoustic model and 3 gaussians mixture for the emission probability.

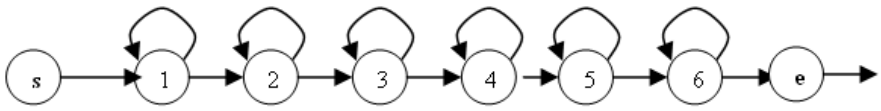



Figure 39: Acoustic model employed in the final Prosodic ASR.

The model has been implemented using the HTK toolkit [5]. The general structure follows the standard ASR architecture, modificated in order to meet the requirements of prosodic recognition.
The output of the recognizer is a list of $N$-Best recognized phrases, because the ASR is born with the idea to be an overlying recognizer for a successive deep processing phase. For each phrase, the confidence level is calculated and the list is truncated when a very low score is registred. In the experiments made for this thesis, the produced $N$-Best list either contained the correct word at least in the first 24 phrases or it didn't include it at all, so $N$ has been set to 24.

Relating to the decoding and classification algorithms, the overall complexity is reduced by the fact that the number of frames is about one order of magnitude less than a standard phonetic recognizer.

## 5.3 A Bottom-Up Prosodic Recognizer

The second approach for prosodic information integration, can be defined *bottom-up* because it acts after another recognizer has transcribed an utterance. The basic idea is that a segmental ASR can indicate the succession of syllables which the utterance is made of, with their relative borders. Those syllables could be even insertions or errors by the recognizer, but a prosodic analysis could revise such segments and decide that some of them are not *coherent* with the adjacent syllables.
In section 1.6 approaches using prosody for $N$-Best list rescoring have been presented. The present approach does not address to the whole utterance, but the syllables it is made of. The idea is so to understand how prosody can be useful at more detailed scale. The system here presented is a syllable rather than utterance rescoring process.
The system acts on the basis of the following rules

- A syllabe is analyzed only if an *anomaly* is detected
- Such syllable is *altered* or *deleted* on the basis of a static rule

The concept of *anomaly* for a syllable has been introduced in order to distinguish between correct and altered syllables. Referring to the words of a vocabulary, some reference values can be calculated for the prosodic characteristics of the syllables composing them.
A statistical evaluation on a rich corpus can be done and, for each syllable the following values can be calculated

- The minimum value of the *energy* the syllable assumes on the entire corpus $E_{min}$
- The minimum number of *voiced samples* the syllable contains on the entire corpus $Uv_{min}$

- The minimum duration in samples $Dur_{min}$
- A flag indicating if a prominent vowel is always present in all the instances of the syllable $Vow$

All the parameters are calculated on a normalized signal.

| | $E_{min}$ | $Uv_{min}$ | $Dur_{min}$ | $Vow$ |
|---|---|---|---|---|
| di | 0.0278 | 0 | 321279 | 0 |
| die | 0.153 | 2071 | 1325000 | 0 |
| do | 0.657 | 2052 | 1300000 | 1 |
| due | 0.084 | 0 | 1000000 | 0 |
| dze | 0.004 | 48 | 1625000 | 0 |
| sei | 0.110 | 0 | 1525000 | 0 |
| ... | ... | ... | ... | ... |

Table 6: An example of some values for the prosodic rescoring.

*A syllable taken from the result of a recognition on the test set, is said to* "present an anomaly" *if its value of energy or duration or the number of voiced samples is less than the minimum calculated on the training corpus.*

When a syllabic anomaly is detected, a set of rules decides the transformation of such syllable. The rules have been extracted from the notice that systematic errors were committed by the segmental recognizer employed.

- First Rule: *Assimilation.* If the anomaly is on a syllable with structure CV1, and the next syllable is a vowel, V2, which does not present an anomaly, then from the two a unique syllable is obtained which is CV2. Example : a possibile segmental recognition output could be "*o-ttan-to-u-no*" with "*to*" being anomaly while "*u*" not. Then the result will be "*o-ttan-tu-no*"[6].
- Second Rule: *Deletion.* If a syllable presents an anomaly, and the *assimilation* rule does not apply, then the right and left adjacent syllables are taken into account. The tri-syllable is analyzed checking if it could belong to a word of the language, if not, the syllable is deleted.

[6]The syllable division here follows acoustic rather than linguistic rules.

Example : suppose the tri-syllable "*sei-tre-sil*"[7] has been recognized, and the word "*tre*" presents an anomaly. Then if the language is only made up of numbers ranging from *0* to *999,999*, then the syllable "*tre*" is deleted.

- Third Rule: *Exception*. If an anomaly is presented at the first syllable after the initial silence, then the tri-sillable to analyze will start from the anomaly.
  Example : suppose an anomaly is registred on "*due*" from the tri-syllable "*sil-due-mi-lle*". Then the tri-syllable to be analyzed will be "*due-mi-lle*", where the *deletion* rule will apply.

The benefits due to the use of this rescoring procedure will be evident in chapter VII, where the results of the multigranular model with the bottom-up approach will be showed. The rules above refer to systematic events associated to the particular segmental recognizer employed, but the adopted paradigm can be applied to all the segmental recognizers which are able to produce a sequence of syllabic markers. Notice that even in this framework the syllable is adopted as the minimal functional unit, in which the prosodic phenomena can be used.

## 5.4 Discussion

So far two different systems aiming to integrate prosodic information into a speech recognizer have been presented. The first adopts an approach which can be defined as "*top-down*", in that the prosodic information results as a reduction of the search space for a successive segmental recognizer. The second approach aims to set the coherence of the recognized syllables by a segmental ASR, and so uses a "*bottom-up*" approach. In both cases, the syllable is assumed to be the minimal form of information which can incapsulate prosodic features. In the case of the top-down approach, an ASR is built, based on syllabic acoustic models but using prosodic features taken from large analysis windows. This is a novel approach in that the feature sequence is not a direct representation of the syllable, but can be useful for discriminating among the phrases. In the case of the bottom-up system, syllables are decided to be deleted, or altered as they mantain a coherence with all the other corresponding instances in the corpus. Also in this case syllabic decisions are taken depending on prosodic cues.

[7] *sil* is the syllabic model for the silence.

The next chapter faces the problem of integrating this two models with the multi-granular segmental ASR, described previously. Many methods can be used, depending on word structure and nature hypotheses.

# 6 Chapter VI: A Multi-Granular Speech Recognizer

## 6.1 Architecture

The present section illustrates several methods for multi-granular integration between the systems showed in sections IV and V.
There are two proposed merging phases, each corresponding to a certain paradigm of integration. The first can be classified as *bottom-up* in that the merging happens after the segmental recognition session. In section 5.3 a prosodic rescoring has been illustrated, which is able to state the prosodic coherence of a syllable. Such method, discussed in further sections, tries to improve system's performances. It starts from the basic assumption that segmental recognition has to be the real core of an ASR, and its results can be only incremented, but nothing can substitute it in signal recognition. All the information coming from the signal has to be processed at segmental level, even if multi-granular analysis is allowed at that level. In the here presented case, the system uses a mixture of phonetic and syllabic information by means of Factorial HMMs (ref. 4.1.2) and only after, the prosodic rescoring is used.
All this theory about multi-granularity is based on considerations about the signal, without exploring other kind of knowledge sources. This choice has been made because only signal processing is investigated by current methods, which is linked to more informatic, rather than social, environment. Multi-granularity theories have explored also human behaviour, by means of many experiments as seen in section 1.5, and many of them give a great importance to the pure signal analysis, which can be processed in parallel on different scales.

The second proposed approach, follows a *top-down* trend, which is linked to the general framework of *multi-pass* strategies in ASR building.

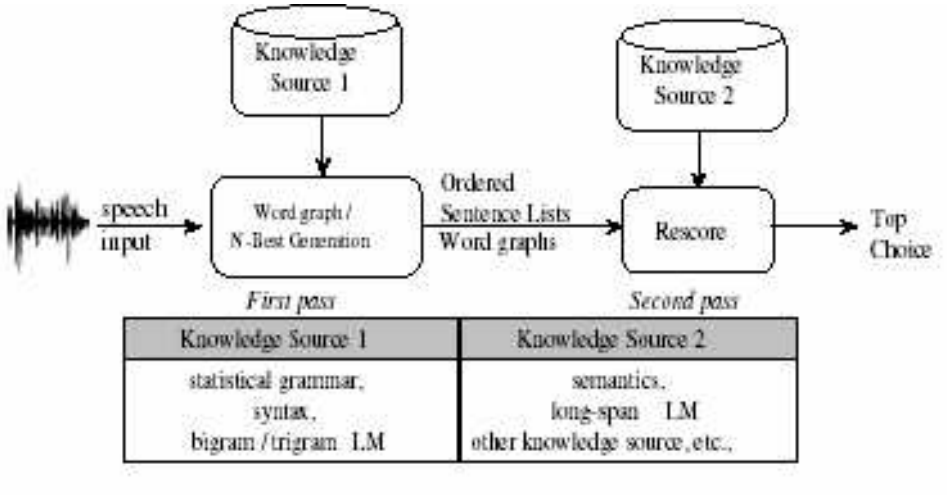



Figure 40: Representation of the Multi-Pass method for ASRs [30]. In the $N$-best search framework, the most discriminant and unexpensive knowledge sources (KS 1) are used first to generate the $N$-best. The remaining knowledge sources (KS 2, usually expensive to apply) are used in the rescoring phase to pick up the optimal solution.

Ideally, a search algorithm set on a single level of analysis, for example focusing the phonemes, should consider all possible hypotheses based on a unified probabilistic framework that integrates information coming from acoustic, language, and lexical pronunciation models, which can be integrated in an HMM state search. It is desirable to use the most detailed models, such as context-dependent models, interword context-dependent models, and high-order n-grams, in the search as early as possible. When the explored search space becomes unmanageable, due to the increasing size of vocabulary, search might be infeasible to implement. As the development of more powerful techniques grows up, the complexity of models tends to increase dramatically. For example, language understanding models can require long-distance relationships. In addition, many of these techniques are not operating in the standard left-to-right manner.

The Multi-pass search is a possibile alternative to such situation. Several knowledge sources are applied at different stages (ref. figure 40), in the proper order to constrain the search progressively. In the initial pass, computationally af-

fordable knowledge sources are used to reduce the number of hypotheses. In subsequent passes, progressively reduced sets of hypotheses are examined, and more powerful and expensive KSs are then used until the optimal solution is found. In top-down approaches, the first stage of the process is also the less discriminant one, while in the bottom-up approaches it is the most discriminant but also complex one.
In the top-down case, the early passes of multipass search can be considered fast matches that eliminate the unlikely hypotheses. Multi-pass search is, in general, not admissible because the optimal word sequence could be wrongly pruned prematurely, due to the fact that not all sources are used in the earlier passes. However, for complicated tasks, the benefits of computation complexity reduction usually outweigh the non-admissibility. In practice, a multi-pass search strategy using progressive knowledge sources, could generate better results than a search algorithm forced to use less powerful models due to computation and memory constraints.
The most straightforward strategy is the so-called $N$-best search paradigm. The idea is to use affordable sources to first produce a list of the $N$ most probable word sequences in a reasonable time. Then these hypotheses are rescored using more detailed models to obtain the most likely word sequence. The idea of the $N$-best list can be furtherly extended to create a more compact representation namely word lattice or graph, which will not be discussed in this section.
The approach using the standalone recognizer, introduced in section 4.1, will be a multi-passing strategy using a first prosodic stage for $N$-best production, followed by a deep segmental recognizer.

## 6.2 Multi-Granular Integrations

### 6.2.1 Long and Short Utterances

Before introducing the proposed integrations for ASRs in a multi-granular model, a set of considerations have to be made about the acoustic features on which the prosodic analysis takes place. A deep gap exists between long and shord words. In the first case the prosodic profile can be very complex and the system can be confused, while in the second case, short words can be discriminated because their prosodic profiles can be very different as well as simple.
From the point of view of an ASR, instead, the differences can be resumed by the following points

- Features are poorly discriminant
- Syllabic acoustic models are not directly associated to the information extracted
- Sometimes there is not difference between a long word and a short one pronunciated slowly
- There are neither pauses or Tone Units[8] which can separate the words

All the points above imply that the recognition of long words by only the prosodic information can be an hard task.
A different treatment has been used to long word recognition in multi-granular systems. The main problem is about the inner nature of such linguistic phenomena. Three hypotheses can be made about

- A long word is a composition of more short words
- A long word is an entire word to be recognized at whole
- A long word is an unexpected phenomenon which can only be treated at fine detail

The first assumption means that a long word is like a fastly uttered phrase, and it should be treated as a succession of words. Algorithms for words separation can be used in order to cut the utterance, and detect the component subwords. Much of the effort is deployed to this last phase.
The second assumption refers to the fact that a long word has to be treated like any other word in the language. So an attempt has to be made in order to extract a list of hypotheses in which also the long word is contained.
The third assumption states that neither the first or the second approach are sufficient to define a long word, and only a successive analysis can be made in this case.
According to the above assuptions, ASRs have been constructed. The results will also be interesting for understanding which is the real nature of a "*long*"

[8] Features describing the intonative contour of a phrase. Often they mark an utterance in which a concept has been exposed.

word. The word *long*, has be defined in this framework, as a word presenting more than 6 syllables.
The next section illustrates the models in detail.

### 6.2.2 Systems Integrations

A set of ASRs has been contructed in order to account for prosodic recognition with different approaches. The here presented models are multi-granular in the fact that they employ multiple levels of analysis and address to different information sources with different temporal scales. The first described will be the *bottom-up* model, which uses the prosodic processing after the segmental recognition. Later, the *top-down* approaches will be shown, in which a pre-processing is made following the multi-passing paradigm. The post processing phase is introduced anyway, for model robustness. The several systems refer to different hypotheses about the nature of "long" words, with "long" indicating words with more than 6 syllables (counted with the automatic procedure in [67]).

#### *Bottom-Up* Multi-Granular ASR (*BU*)

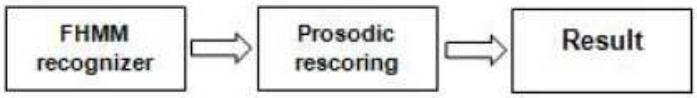



Figure 41: Multi-Granular ASR using a *bottom-up* approach.

This system is made up of a phonetic-syllabic recognizer based on Factorial HMMs (4.1.2), followed by a prosodic post-processor (5.3) for results control. The basic assumption here is that a word must always be analyzed in its fine details. The prosodic information is useful only in results verification and re-arrangement. The integration method is *bottom-up* and the basic steps can be resumed as follows

- A phonetic-syllabic recognition is acted on the entire signal
- A successive processing rearranges the syllabic succession by the previous phase, on the basis of prosodic analysis

**Multi-Granular ASR with *Entire* word recognition (*TDE*)**

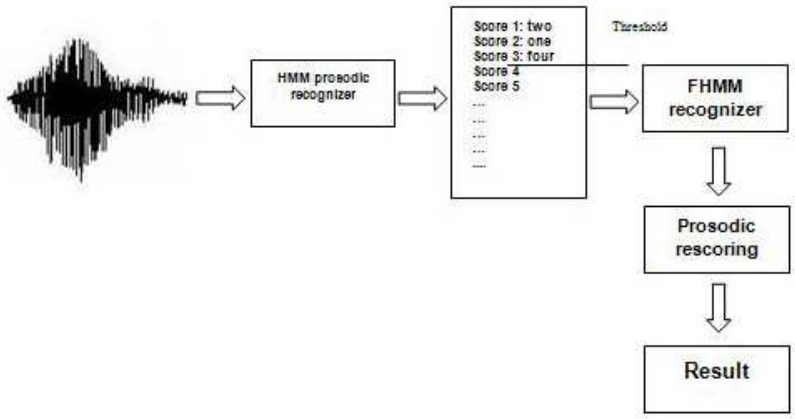



Figure 42: Multi-Granular ASR with single prosodic analysis.

The system is made up of a multi-pass architecture (ref. 6.1) which uses an initial prosodic recognizer, followed by a phonetic-syllabic recognizer. In the end a rescoring procedure modificates the syllabic successions by the previous step. The approach stresses the *top-down* method, but also *bottom-up* processing is used. The assumption on which the model is based is that the fine-analysis can be helped by a prosodic pruning. The integration is at the acoustic models and at the language model levels. The not involved syllables are not processed, and the syllabic connections not appearing in the $N$-Best list are annihiled. A long word is here treated as an entire word which cannot be decomposed in sub-words.

The process can be resumed as follows

- A prosodic recognition is performed
- From the previous stage a $N$-Best list is obtained
- The phonetic-syllabic recognizer acts by setting to zero the acoustic models scores and the connection to the syllables not included in the $N$-Best list
- A successive processing rearranges the syllabic succession by the previous phase, on the basis of prosodic analysis

The above model has not been useful in long word recognition. In such cases the word was not included *at all* in the $N$-Best list, and so the model was equivalent to perform the deep analysis directly, when in presence of long words.
The process can be so riformulated as follows

- Analysis of a signal and classification in *long* or *short*. The choice is based on Fujisaki impulses and on the number of automatically extracted syllables
- In the case of a long signal, a fine-recognition is performed
- A bottom-up processing is made on the result
- In the case of a short signal a prosodic recognition is performed
- From the previous stage a $N$-Best list is obtained
- The phonetic-syllabic recognizer acts by setting to zero, the acoustic models scores of the syllables not included in the $N$-Best list
- A successive processing rearranges the syllabic succession by the previous phase, on the basis of prosodic analysis

**Multi-Granular ASR with *Multiple* sub-words recognition (*TDM*)**

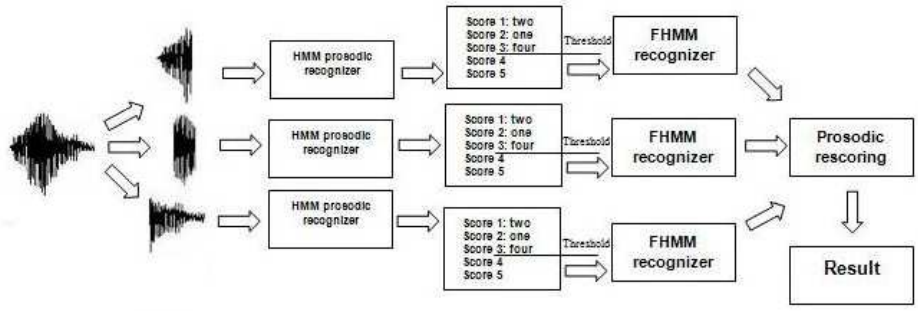



Figure 43: Multi-Granular ASR, decomposing a long signal in short signals succession, with a factorial recognizer for each extracted piece.

Even in this case the system is made up of a multi-pass architecture which uses an initial prosodic recognizer, followed by a phonetic-syllabic recognizer. In the end a rescoring procedure modificates the syllabic successions by the previous step. The approach stresses the *top-down* model, but also *bottom-up* processing is used. The assumption on which the model is based is that the fine-analysis can be helped by a prosodic pruning. The difference respect to the previous recognizer is that now a *long* word is assumed to be a composition of more sub-words by the language, which have to be separately recognized. The merging phase is *after* the recognition of the single sub-words is teminated. The bottom-up rearrangment acts on the output by the merging phase.
This is the procedure summary

- Analysis of a signal and classification in *long* or short. The choice is base in Fujisaki impulses and on the number of automatically extracted syllables
- In the case of a short signal, the procedure in ***TDE*** is performed
- In the case of a long signal, the following procedure applies
- The signal is segmented according to Fujisaki impulses and automatic syllabic separation

- Each segment is treated as a separate word
- A bottom-up processing is made on the result
- A prosodic recognition is performed
- From the previous stage a $N$-Best list is obtained
- *For each word-segment,* the phonetic-syllabic recognizer acts by setting to zero, the acoustic models scores of the syllables not included in the $N$-Best list
- In the end, the results are concatenated in order to get a unique result

- A successive processing rearranges the syllabic succession by the previous phase, on the basis of prosodic analysis

**Multi-Granular ASR with sub-words Prosodic recognition (*TDP*)**

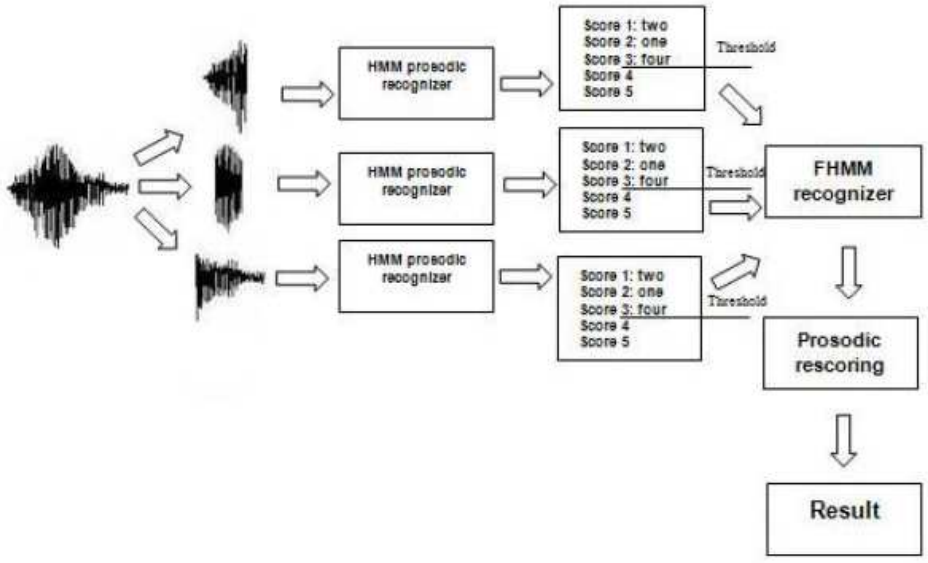



Figure 44: Multi-Granular ASR, decomposing a long signal in short signals succession, with a unique factorial recognizer for all the pieces.

As in the previous case, the system is made up of a multi-pass architecture which uses an initial prosodic recognizer, followed by a phonetic-syllabic recognizer. In the end a rescoring procedure modificates the syllabic successions by the previous step. The approach stresses the *top-down* method, but also *bottom-up* processing is used. The assumption on which the model is based is that the fine-analysis can be helped by a prosodic pruning. The difference here is that a *long* word is identified, *only at prosodic level*, as a composition of sub-words from the language.
The following is a summary of the process

- Analysis of a signal and classification in *long* or short. The choice is base in Fujisaki impulses and on the number of automatically extracted syllables
- In the case of a short signal, the procedure in ***TDE*** is performed

- In the case of a long signal, the following procedure applies
- The signal is segmented according to Fujisaki impulses and automatic syllabic separation
- Each segment is treated as a separate word
- A bottom-up processing is made on the result
- A prosodic recognition is performed
- From the previous stage a $N$-Best list is obtained
- A merging phase is performed, where from the many lists, a single one is obtained
- *On the merged list,* the phonetic-syllabic recognizer acts by setting to zero, the acoustic models scores of the syllables not included in the $N$-Best list
- In the end, the results are concatenated in order to get a unique result
- A successive processing rearranges the syllabic succession by the previous phase, on the basis of prosodic analysis

The merging phase here is obtained by the cross product between all the occurrences in the lists, respecting the temporal order of the segments they belong to.

## 6.3 Discussion

So far we have showed two multi-granular kinds of systems. The one based on a multi-pass strategy and another using a rescoring module for multi-scale integration. Each of them try to integrate information coming out from a prosodic scale of analysis into a more detailed process, based on a mixture of base units. The multi-granular analysis acts at different stages. It can be implicitly found into the segmental recognizer, or it can be noticed in the systems interaction. A particular attention has to be kept to "*long*" words. They can be treated as combinations of shord words or they can be thought as entire units. The soltion to this problem is not trivial and strongly influences systems performances. The first hypothesis leads to find an algorithm for words decomposition in subwords,

which could extract also units that are not contemplated by the language. The second hypothesis, instead, has to deal with the fact that the only prosodic information cannot distinguish among long words, because their internal complexity increases. Prosodic features discriminant power is limited only to *"short"* words.

Each proposed multi-granular ASR start from a different set of initial assumptions about the nature of words and speech processing.

The next chapter exposes the results by all the systems, introducing also the chosen experimental framework.

# 7 Chapter VII: Results

## 7.1 The Corpus

In this section the employed corpus will be described. As stated previously, this is a collection of audio files, each with an associated transcription at multiple levels. Transcriptions about phonemes as well as words or syllable are present, and often also prosodic events annotations can be found. The corpus used for this thesis has been taken from SPEECON [13]. 18 different languages are present, which cover the most used european languages. Also dialect inflections can be found in the collection.
The extracted piece is a set of spoken numbers in Italian language, pronounciated by only male speakers. Each signal has a sampling frequency of 16000 Hz. There are 1906 recordings, pronounciated by about 400 different speakers, who have recorded about 5 sentences for each one. Among these, 4007 files have been taken corresponding to numbers ranging from 0 to 999,999. In each recording a single word in the range is present.

| parola | # occ | parola | # occ | parola | # occ |
|---|---|---|---|---|---|
| zero | 121 | sedici | 22 | sessanta | 62 |
| uno | 126 | diciassette | 62 | sessantuno | 13 |
| due | 233 | diciotto | 55 | sessantotto | 12 |
| tre | 242 | diciannove | 52 | settanta | 49 |
| quattro | 246 | venti | 55 | settantuno | 5 |
| cinque | 215 | ventuno | 6 | settantotto | 6 |
| sei | 220 | ventotto | 0 | ottanta | 44 |
| sette | 222 | trenta | 53 | ottantuno | 4 |
| otto | 190 | trentuno | 6 | ottantotto | 3 |
| nove | 230 | trentotto | 5 | novanta | 49 |
| dieci | 67 | quaranta | 40 | novantuno | 4 |
| undici | 57 | quarantuno | 11 | novantotto | 4 |
| dodici | 57 | quarantotto | 11 | cento | 581 |
| tredici | 17 | cinquanta | 74 | mille | 52 |
| quattordici | 52 | cinquantuno | 5 | mila | 310 |
| quindici | 50 | cinquantotto | 7 | | |

Table 7: Dictionary words with relative occurrencies count.

A grammar and a dictionary have been extracted from the corpus, where only 47 words are necessary to build up all the words in the language. The occurency of the dictionary words are reported in table 7.
The here presented experiment is centred on the syllables other than phonemes and prosody. A syllabic transcription was necessary for the learning session of the models, unfortunately this was not present in the corpus. A syllabic annotation session has been manually made on the entire corpus to set up the experimental environment by means of the Wavesurfer tool [61].
The corpus has been divided as follows

- $\frac{1}{3}$ of the corpus as been used for testing and development, the rest has been used for training
- A speaker present in training set is not present in the test set and viceversa

Some considerations must be made about the motivations for the choice of numbers as the experimental environment. First of all the choice for numbers will be justified along with the subdivision of the dictionary in 47 words, along with the "*perceptive*" subdivision in syllables.

### 7.1.1 The Language

In experiments about speech recognition, it is necessary to select a dictionary, which is neither too big, because of development time problems, or too little, because it would not represent all the problems can be found with large dictionaries.

The domain of numbers is sufficiently varied and has a wide number of features can be found in natural language applications. Such domain is quite little but not trivial. The chosen range is *0-999,999*, where few syllables can build up all the vocabulary and even long words. Numbers have a well defined grammar, for which no statistical analysis is necessary, but ambiguities and superpositions are present because the same set of syllables is shared among several words. The same thing would not happen in the case of digits (numbers from *0* to *9*) where few ambiguities and superpositions are present. The possibility to have a static and defined grammar avoids the use of approximation in the extimations

of concatenation probabilities. So the focus can be set on the acoustic models and the decoding procedure, as the experiment here presented needs.

```
$zero=zero;

$uni1 =due|tre|quattro|cinque|sei|sette|nove;

$uni2 =uno|otto;

$uni3=due|tre|quattro|cinque|sei|sette|otto|nove;

$uni=$uni1|$uni2;

$dec=dieci|undici|dodici|tredici|quattordici|quindici|sedici|diciassette|
     |diciotto|diciannove|venti|venti-$uni1|ventuno|trenta|trenta-$uni1|
     |trentuno|trentotto|quaranta|quaranta-$uni1|quarantuno|quarantotto|
     |cinquanta|cinquanta-$uni1|cinquantuno|cinquantotto|sessanta|
     |sessanta-$uni1|sessantuno|sessantotto|settanta|settanta-$uni1|
     |settantuno|settantotto|ottanta|ottanta-$uni1|ottantuno|ottantotto|
     |novanta|novanta-$uni1|novantuno|novantotto;

$cen=cento|cento-$dec|cento-$uni|$uni3-cento|$uni3-cento-$dec|
     |$uni3-cento-$uni;

$mi1=mille|mille-$cen|mille-$dec|mille-$uni;

$mi2=$cen-mila|$cen-mila-$cen|$cen-mila-$dec|$cen-mila-$uni|$dec-mila|
     |$dec-mila-$cen|$dec-mila-$dec|$dec-mila-$uni|$uni3-mila|
     |$uni3-mila-$cen|$uni3-mila-$dec|$uni3-mila-$uni;


   ( SENT-START ($zero|$uni|$dec|$cen|$mi1|$mi2) SENT-END )
```

Figure 45: ABNF Grammar for numbers from *0* to *999,999*

### 7.1.2 The Acoustic Models

Syllables representation for words in the vocabulary had to be chosen, along with words representing the utterances. An acoustic-perceptive choice has been made. Starting from the signal, energy islands have been isolated and annotated by hand. Table 8 reports the extracted syllables along with their number of occurencies. As can be noticed, the subdivision does not agreee with the linguistic decompositions. This is because automatic systems have to be used in ASRs, some of them addressing to automatic syllabic decompositions. The annotation criteria has so been alligned to the rules used by automatic syllabators, like that in [67]. Energy islands presenting an **onset**, **nucleus** and **coda**, have so been detected and the signal transcribed.

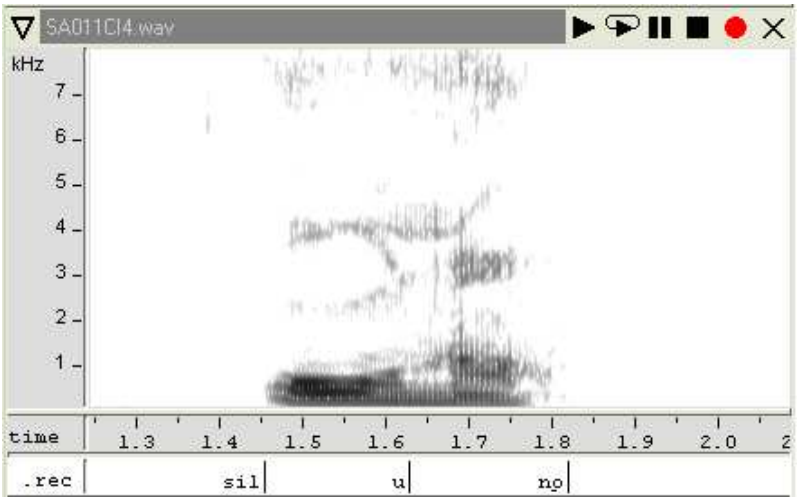



Figure 46: Example of syllabic annotation.

| sillaba | # occ | sillaba | # occ | sillaba | # occ |
|---|---|---|---|---|---|
| cen | 581 | o | 241 | tre | 259 |
| ci | 322 | qua | 360 | tren | 64 |
| cia | 114 | quan | 86 | ttan | 111 |
| cin | 301 | que | 215 | tte | 284 |
| cio | 55 | quin | 50 | tto | 293 |
| di | 424 | ran | 62 | ttor | 52 |
| die | 67 | ro | 121 | ttro | 246 |
| do | 57 | se | 391 | tu | 54 |
| due | 233 | sei | 220 | u | 126 |
| la | 310 | ssan | 87 | un | 57 |
| lle | 52 | sse | 62 | van | 57 |
| mi | 362 | ta | 371 | ve | 282 |
| nno | 52 | ti | 55 | ven | 61 |
| no | 467 | to | 629 | ze | 121 |

Table 8: Syllables extracted from the corpus.

Words in the dictionary have been also chosen according to this criteria, because a word like *diciotto* has a spectral realization which cannot be decomposed in units (e.g. *dici-otto*), as in that case also other words (e.g. *dici*) should be introduced in the dictionary, which are not frequent and are also highly dependent on pronunciation.

## 7.2 Results

The following sections show the performances of the systems previously described. For each ASR results are reported at various levels.

- At **Syllable** level. That is in syllables classification
- At **Dictionary Words** level. That is in dictionary words recognition
- At **Sentence** level. That is in the recognition of the uttered number

Notice that the experiments are on single number recognition in the range 0-999999. So, for syllables and dictionary words, also a calculation about false insertions, deletions or substitutions will be reported, in the value for the *accuracy*. In the case of sentences, only the report of the number of correct words will be shown, as there are no cancellations, substitutions or deletions.
For results interpretation some definitions have to be introduced, which are commonly used reference values in literature.

$$Correctness = \frac{H}{N} X100$$

$$Accuracy = \frac{H - I}{N} X100$$

$$Word\,Error\,Rate = \frac{I + S + D}{N} X100 = 1 - Accuracy$$

Where

- $H$: number of correctly recognized units
- $I$: number of over-inserted units
- $S$: number of substituted units
- $D$: number of deleted units
- $N = H+S+D$: total number of units present in the manual transcription

The first value accounts only for the fact that the unit (syllable, word or sentence) is in the transcription, the second value controls if the recognition is effectively "*right*", by checking also if a substitution, deletion or insertion is associated to the unit.
The next sections will report the results, which will be discussed in the next chapter.

### 7.2.1 Baseline ASR Performances

The baseline system is described in section 3.2. Results are reported at the variation of the speech unit considered for acoustic models.
The employed algorithm for decoding is the Viterbi (ref. 3.2.6).

- Using **phonetic** acoustic models, the results on the transcription of the entire utterance and of dictionary words are

| | **On Utterance** | **On Dictionary Words** | | |
|---|---|---|---|---|
| **Number of states** | **Corr** | **Corr** | **Acc** | **WER** |
| 5 | 58.15% | 79.65% | 69.87% | 30.13% |
| 6 | 60.69% | 81.65% | 75.32% | 24.68% |
| 7 | 62.14% | **80.21%** | **75.56%** | **24.45%** |

Table 9: Performances of the baseline system with phonetic units, at the variation of the number of states.

The best results are achieved using 7 states acoustic models. The experiments have detected a fast reduction in performances when the number of states is more than 7, because units larger than phonemes are modelles in that case.

- Using **syllabic** acoustic models, the results on the transcription of the entire utterance and of dictionary words are

| | On Utterance | On Dictionary Words | | |
|---|---|---|---|---|
| **Number of states** | **Corr** | **Corr** | **Acc** | **WER** |
| 7 | 30.07% | 67.47% | 37.18% | 62.82% |
| 9 | 46.38% | 70.11% | 57.69% | 42.30% |
| 10 | 41.85% | 66.03% | 51.12% | 48.88% |
| 11 | 52.90% | **73.00%** | **63.46%** | **36.53%** |
| 12 | 50.54% | 68.51% | 60.02% | 39.98% |
| 13 | 50.72% | 68.43% | 61.30% | 38.70% |

Table 10: Performances of the baseline system with syllabic units, at the variation of the number of states.

In this case, the best results are achieved with 11 states.

- Using an acoustic model for each **word in the dictionary**, the results on the transcription of the entire utterance and of dictionary words are

| | On Utterance | On Dictionary Words | | |
|---|---|---|---|---|
| **Number of states** | **Corr** | **Corr** | **Acc** | **WER** |
| 18 | 29.17% | **55.85%** | **37.74%** | **62.26%** |

Table 11: Performances of the baseline system with entire dictionary words taken as units, at the variation of the number of states.

This experiment has produced very low results when the number of states is different from 18. This last model has not been used in the rest of the experiments because using an acoustic model for each word leads to a complex and not generalizable ASR.

### 7.2.2 Factorial ASR Performances

The Factorial system is described in section 4.1. Results are reported in comparison to the baseline system at the variation of the focus unit.
The employed algorithm in such model is the new algoritm introduced in section 4.2.5.

- In syllables classification, the performances between the Factorial and Standard HMMs are reported in the following table

| **Syllabic Model** | **Accuracy** | **Correctness** |
|---|---|---|
| FHMM 2 lev. with 7 states each | **84.81%** | **94.30%** |
| Standard HMM with 7 states and syllabic models | 81.33% | 89.11% |
| Standard HMM with 5 states and phonetic models | 85.74% | 93.91% |

Table 12: Results on syllables classification.

Factorial model seem to use better the multiple information inside a syllabic length segment.

- In utterance transcription, the performances of the Factorial model is compared to the baseline system, referred to as *HTK*, and to a system employing standard HMMs acoustic models for syllables, but the new presented algorithm for word decoding.

| Model | Correctness |
|---|---|
| ASR with Factorial HMMs and proposed Decoding Algorithm | **68.84%** |
| ASR with Standard HMMs and proposed Decoding Algorithm | 65,19% |
| HTK with 7 states phonetic models | 62.14% |
| HTK with 11 states syllabic models | 52.90% |

Table 13: Results on utterances transcription.

Even in this case, the Factorial model performs better. The benefit comes from the dual layer nature and the decoding procedure. Better performances are calculated also for classical models employing the exact decoding algorithm introduced in section 4.2.5.

### 7.2.3 Mean permanence in state for each layer

In order to put in evidence the behaviour of the Factorial HMMs, a measure has been introduced, which is able to demonstrate that the Factorial model is effectively able to extract a slower and a faster dynamic from the signal.
This is achieved by means of the ***mean permanence in state***, which is calculated as

$$t(\phi, m) = \frac{1}{N-2} \sum_{i=2}^{N-1} \frac{1}{1 - a_{ii}^{(m)}}, \quad m = 1, 2$$

where $\phi$ is the acoustic model, $m$ is the layer and $a_{ii}^{(m)}$ is the probability transition from the state $i$ to itself. The quantity refers to the mean number of times the system makes self loops for each state in the layers.
The calculation can be made on all the models to have an overall idea of the effective presence of two dynamics.

$$\bar{t}(m) = \frac{\sum_{\forall \phi} t(\phi, m)}{Numb.\phi} \quad m = 1, 2$$

The results for this quantity is reported in table 14.

| Slow Level | Fast Lever |
|---|---|
| **4.50** | **7.84** |

Table 14: Mean permanence in state for the two Factorial levels.

It is evident that two dynamics are present for the different layers. One evolves fastly, the other slowly. This is in agreement to the hypothesis for Factorial HMMs to be able to model multiple dynamics of the signal. According to the model, this should refer to syllabic and phonetic phenomena, even if the real nature of such trends cannot be extracted fromt the results. The identification can be only supposed but not demonstrated.

### 7.2.4 Multi-Granular ASR Performances

The multigranular systems showed, are that described in section 6. In particular

- **BU**: Multigranular ASR with Factorial HMM models, followed by prosodic rescoring
- **TDE**: Multigranular ASR with Factorial HMM models preceed by prosodic recognition, in the case of short signals (less than 6 syllables). On long signals the **BU** process is directly applied.
- **TDM**: Multigranular ASR with Factorial HMM models preceed by prosodic recognition, in the case of short signal (less than 6 syllables). Long signals are instead divided in short signals which are recognized separately.
- **TDP**: Multigranular ASR with Factorial HMM models preceed by prosodic recognition, in the case of short signal (less than 6 syllables). Long signals are instead divided in short signals which are prosodically recognized separately. The factorial recognition is acted on the merged result coming from the previous stage.

| Model | Accuracy | Correctness |
|---|---|---|
| ***BU*** Multi-Granular System | **90.15%** | **93.69%** |
| ***TDE*** Multi-Granular System | 89.11% | 92.80% |
| ***TDM*** Multi-Granular System | 64.42% | 72.02% |
| ***TDP*** Multi-Granular System | 1.33% | 73.74% |
| FHMM 2 lev. with 7 states each | 84.81% | 94.30% |
| Standard HMM with 7 states and syllabic models | 81.33% | 89.11% |
| Standard HMM with 5 states and phonetic models | 85.74% | 93.91% |

Table 15: Results on syllables classification.

- Table 15 shows the results on syllables recognition. The **Bottom-up** approach gets the best results, nearly followed by the **TDE** model.

| Model | Accuracy | Correctness |
|---|---|---|
| ***BU*** Multi-Granular System | **84.49%** | **89.89%** |
| ***TDE*** Multi-Granular System | 82.77% | 88.84% |
| ***TDM*** Multi-Granular System | - | 56.14% |
| ***TDP*** Multi-Granular System | - | 57.11% |

Table 16: Results on Dictionary Words recognition.

- Table 16 shows the results on dictionary words recognition. While the BU is very close to the TDE, the other two systems get no results on accuracy, in that always insertion, cancellations or substitutions are produced. This means that the automatic subdivision of a long signal in many short signals does not perform well.

| Model | Correctness |
|---|---|
| ***BU*** Multi-Granular System | **79.17%** |
| ***TDE*** Multi-Granular System | **78.06%** |
| ***TDM*** Multi-Granular System | 68.04% |
| ***TDP*** Multi-Granular System | 65.98% |
| FHMM 2 lev. with 7 states each | 68.84% |
| Standard HMM with 12 states, syllabic models and proposed decoding alg. | 65,19% |
| Standard HMM with 7 states, phonetic models and proposed decoding alg. | 64.09% |
| HTK with 7 states phonetic models | 62.14% |
| HTK with 11 states syllabic models | 52.90% |

Table 17: Results on entire utterance transcription.

- Table 17 shows the results on entire word recognition. Notice that accuracy is not reported because the recognition is on a single word in the range *0-999,999*. Even in this case the BU system gets the best performances. Scores are reported also for a system using standard HMMs using the exact decoding algorithm introduced in section 4.2.5.

From the above results, it is clear that the multi-granular system really introduces useful factors, which are able to improve performances of **17.03%** respect to stardard architectures, and of **26.27%** respect to standard syllabic systems! The next chapter will also talk about the difficulties in applying the model to real cases, because the constraint here is that the real-time requirements have been ignored.

Notice that even the exact decoding algorithm introduces sensible improvements, as the performances respect to the standard Viterbi are increased by **12.29%** in the case of syllabic models.

## 7.3 Discussion

The performances of the implemented systems have been shown on the task of numbers recognition ranging from *0* to *999,999*. The choice for the corpus and the language is due to the need of a workbench which is able to let the experiment focus on the acoustic models and decoding phase, rather than on the language model statistical estimation or large vocabulary recognition. The numbers have been chosen because they present many of the problems can be found with big vocabularies.

The performances of the systems have been calculated starting from the baseline ASR built with HTK, and then with the multi-granular segmental ASR, based on Factorial HMMs. In the end, the multi-granular system has been compared to the others and the results show a very high performance improvement. Other experiments can be found in the comparison table, with systems using standard acoustic models in combination with a novel, exact, algorithm for utterance decoding. Such algorithm is able to exploit the acoustic models at best and gives the best alignment of the models to the signal. The benefits are evident in that the performances strongly change.

The next chapter will argue the consequences of the obtained results, each of which is associated to some assumptions about the nature of words and of the recognition process. The discussion will focus on the meaning of the model as it can also guess something about human recognition, as previewed by psicoacoustic theories.

# 8 Chapter VIII: Discussion

## 8.1 Summary

The multi-granular system presented has shown to get very high performances respect to a baseline system.
The following figures shows a comparison between a standard system (e.g. HTK, ref 3.2) and the proposed multi-granular model.

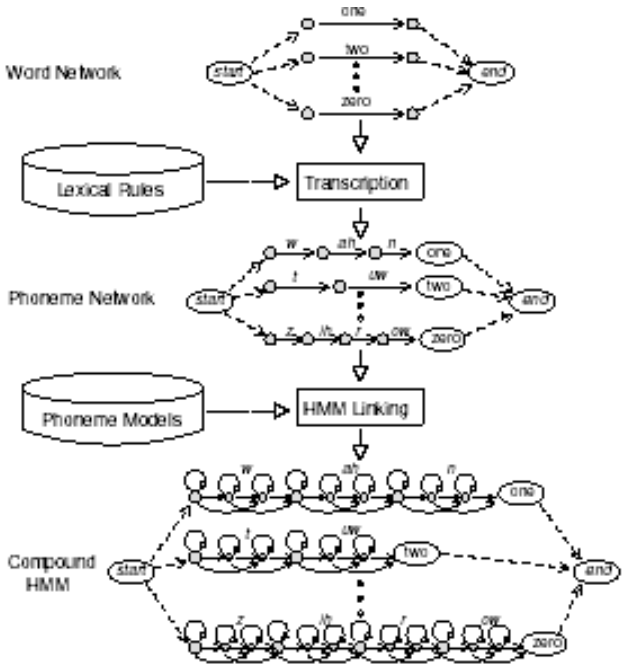



Figure 47: HTK ASR schema [5].

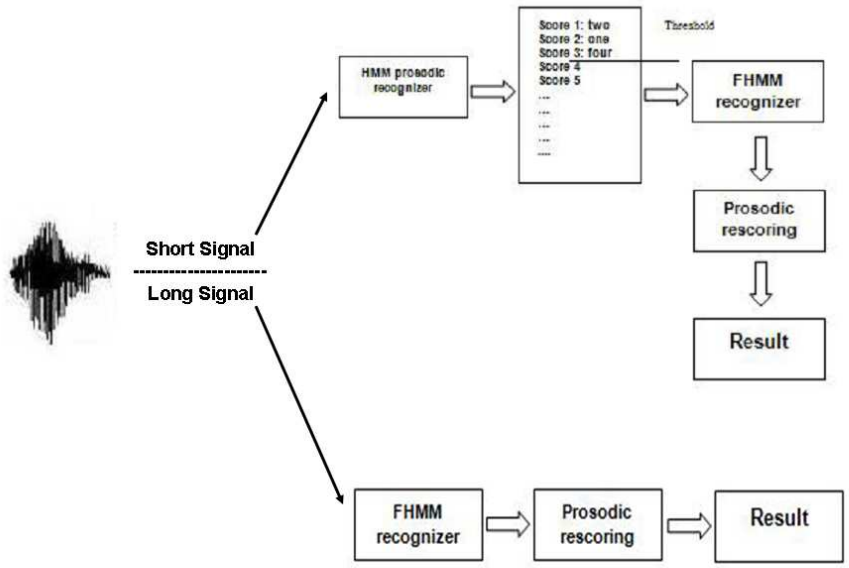



Figure 48: Multi-Granular ASR schema.

As can be noticed, a distinction is made between long and short signals, which is taken on the basis of the number of syllables. A long signal is one having more than 6 syllables, on which a prosodic recognizer is completely unable to guess the result. It does not appear even in the list of the first 200 best recognized phrases. This means that a separation in the functioning has to be made in order to exploit system's functionalities at best. On the other side the results in section 7.2.4 show that the best results are achieved when the only bottom-up system is employed. So it seems that the prosodic recognizer used in the multi-pass step is not so useful. Going deeply into details, the difference is in the fact that the complexity of the factorial model strongly depends on the number of active models and the likelihood calculation.

| Signal Name | TDE M-G ASR | Bottom-Up ASR | Reduction Perc. |
|---|---|---|---|
| SA005CI | 40.24 | 65.29 | 38.37% |
| SA015CI | 43.59 | 72.17 | 39.61% |
| SA305CI | 48.12 | 71.01 | 32.23% |
| SA401CI | 41.44 | 61.77 | 32.93% |
| SA524CI | 57.37 | 82.11 | 30.13% |
| SA541CI | 45.99 | 75.32 | 38,94% |
| SA287CI | 52.55 | 71.72 | 26,72% |
| SA230CI | 42.63 | 69.58 | 38.73% |
| SA209CI | 47.08 | 80.02 | 41.17% |
| SA178CI | 52.70 | 85.23 | 38.16% |

Table 18: Comparison (in **seconds**) between Multi-Granular Top-Down and Bottom-Up recognizers.

| Model | Correctness |
|---|---|
| ***BU*** Multi-Granular System | **79.17%** |
| ***TDE*** Multi-Granular System | **78.06%** |
| ***TDM*** Multi-Granular System | 68.04% |
| ***TDP*** Multi-Granular System | 65.98% |
| FHMM 2 lev. with 7 states each | 68.84% |
| Standard HMM with 12 states, syllabic models and proposed decoding alg. | 65,19% |
| Standard HMM with 7 states, phonetic models and proposed decoding alg. | 64.09% |
| HTK with 7 states phonetic models | 62.14% |
| HTK with 11 states syllabic models | 52.90% |

Table 19: Remind of the results on entire utterance transcription.

Table 18 shows a comparison between the computational time for the recognition of some files, in the case of the **BU** and the **TDE** model. A mean difference of 35,7% is calculated on the entire corpus, so that the TDE model has the advantage to reduce the processing time.

A particular stress has to be given to the novel decoding algorithm introduced in section 4.2.5. The advantage is in the fact that even standard models can achieve better results, even if the computational time increases.
In summary the following considerations can be argued from the results

- The Bottom-Up approach reaches the best score of **79.17%** in correctness
- The Multi-Pass Strategy makes only sense as a catalyst of the entire process
- The exact decoding algorithm is able to increase the recognizer performances, even in the case of standard HMMs acoustic models
- The automatic segmentation of the signal in sub-words is not effective

The multi-granular approach seems to make sense if the prosodic information is employed after the segmental recognition. The top-down approach, instead, can be used, on short files, where it is useful as a catalist of the whole processing. The automatic segmentation is not able to divide the signal in elements corresponding to dictionary words. This is due to the fact that the signal is severely altered by the fast inner coarticulation of a word. The next section analyzes such problems.

## 8.2 Issues

In this section, the problems of the here presented models will be highlighted. The main evident issue regards the distinction between "long" and "short" words, where the first term refers to more than 6 syllables words. Why such dinstiction has to be made?
This problem refers to the real nature of long words, which must be distinguished from long sentences. This difference is evident in the recognition of single words ranging from 0 to 999,999. From the point of view of a prosody based recognizer, there is no way to recognize a long word, not even in the first 200 best utterances. From a spectral point of view this long signal can be confused with a shord word pronunciated slowly. This is because features are quite rough, and models cannot discriminate by only means of them. Another problem can be the fact that the used architecture is a standard one, referring to syllabic acoustic models

which are not in a direct correspondence to the observations. The acoustic models could be not suitable, but even the decoding strategy could lack in something.
In summary, the problems for the long words are

- The poorly discriminant features
- The acoustic models not corresponding to the observations
- A decoding algorithm not suitable to the problem complexity

Attempts to solve such problems have been made, by varying the type of acoustic models employed, but no way has been found. Maybe the real nature of the problem, which is the main reason of the failure of the TDM and TDP recognizers, is that a long word is not really made up of many sub-words corresponding to dictionary words, but it is made up of altered sub-words. Such differences can be due to the deletion or substitution of some phonemes.
In this framework the bottom-up approach, discussed in section 5.3, can be introduced, which is an attempt to model the prosodic events in another way. A set of static rules is employed on syllables, in order to make the intervention of prosody more generalizable. A rescoring process on the whole utterance, would have not been generalizable to other vocabularies.
The problem of long words decomposition can have two explanations

- A procedure for decomposing a sentence in sub-words from the dictionary *cannot* exist, but the obtained segments can only be associated to altered sub-words, to be managed in the dictionary
- The procedure *can* exist but the features to use for the decomposition are not yet clear

In the first case a possible solution is to alter the grammar in order to contemplate also altered words, but this could compromise the whole recognizer performances.
On the other side, it could be guessed if it makes sense to build up a speech recognizer based on prosody. Prosody could be used only in bottom-up approaches

because of its super-segmental nature, in the sense that it only adds information to the segmental level, but cannot substitute it. This is another open point left by the present thesis.
The problem can be generalized to a more wide question: Does it make sense to build up a multi-pass structure for an ASR, using prosody as first step? From the discussion above it seems that the procedure is not so useful, but it can be confusing sometimes. As can be seen from table 19, the bottom-up process is more performant, while the top-down approach is useful only as catalyst. So it seems the multi-pass strategy is not so useful in this framework.

## 8.3 Future Work

In this section, future work about multi-granular ASRs will be discussed. Commercial ASRs lack in the fact that applications to wide public is not robust. Speech variability is too complex to be catched with a simple phonetic recognizer. A factorial or a multi-granular model could be more robust, but it can present complexity issues. To overcome this problem, approximate methods, as described in section 4.2.1 could be used. The calculation of factorial models likelihood is the most complex phase in the presented approach, but an increment in speed can be achieved by getting the results with approximate algorithms. The important thing is to preserve the relative differences between the likelihoods during the decoding phase: the correct recognition must have the highest likelihood.
Further work will so focus on the real-time processing, which has been neglected in this thesis.
Also the decoding strategy employs much more time than the standard Viterbi algorithm, even if the benefts are evident. A workaroud in this case can be to use beam seach strategies [30], to reduce the solution search space.
In summary, the experiment is really encouraging besides the high computational complexity. Notice that this highly depends on the number of frames and states. On the other side if the word is short such problems do not arise. This means that with little vocabularies containing short words, this kind of recognizer can achieve very high results and be also efficient.
Interactive Voice Responsers (**IVR**) are among the systems which highly use such kind of recognition. Automatic telephony agents address to a wide public and so usually employ simple speech recognition on little static grammars. A

factorial model can be embedded in one of these systems and could be perfect for many applications. An example can be the recognition of isolated or connected digits, which are the most employed in IVR applications. A recent proposal has been that of introducing multi-granular recognizer in the Avaya Devconnect Program [49] for experimentation.
A last note is about applications on large vocabularies. With the present situation, it is not thinkable to apply the multi-granular recognizer to the so called Natural Language applications, in which a person can speak freely to an automatic system. The problem is in the high dependency on the number of frames which the signal is made of. In real cases it can be guessed if a dialogue is really made up of long signals. The prosodic analysis of the **TDM** and **TDP** models, has stressed the possibility to subdivide the signal in many signals. In spoken dialogues words are made up of two syllables on average (in Italian language). A segmentation by means of tone units and Fujisaki accent components, could so divide the whole signal in many sub-signals with a treatable complexity from a multi-granular ASR point of view. The result would not achieve bad results as in the case of long words segmentation, in that sentences are very different from long words. Prosodic variations are evident, especially when a complete meaning has been expressed in a sentence [67]. A dialogue can be seen as a set of many short sentences, many constituted of two syllables. So a multi-granular strategy could in principle be applied.
The discussion about multigranular models can be resumed by the following points

- Multi-Granular models are more robust to speech variations
- They can be used in little vocabularies, with short words, achieving very high performances
- IVR framework can be suitable for such application
- Application to large vocabularies need to introduce approximated methods for likelihood calculation and words decoding
- Long Dialogues can be seen like sequences of shorter phrases, where multi-granular models can be applied

All this statements are left as future work on such models.
The striking thing here is that an automatic speech recognizer has been built, starting from ideas belonging to the origin of the language studies, supported in time by psycoacoustic experiments and by some mathematical models. The results demonstrate that those ideas are able to equip an artificial system of powerful instruments to overperform a standard approach. In our opinion, this can be really the base for robust and future generation automatic speech recognizers.

# 9 Chapter IX: A Practical Application

## 9.1 Introduction

The previous chapter has highlighted the good and bad aspects of a model for a multi-granular ASR, based on syllabic acoustic models and prosodic features. The resulting system is depicted in figure 49.

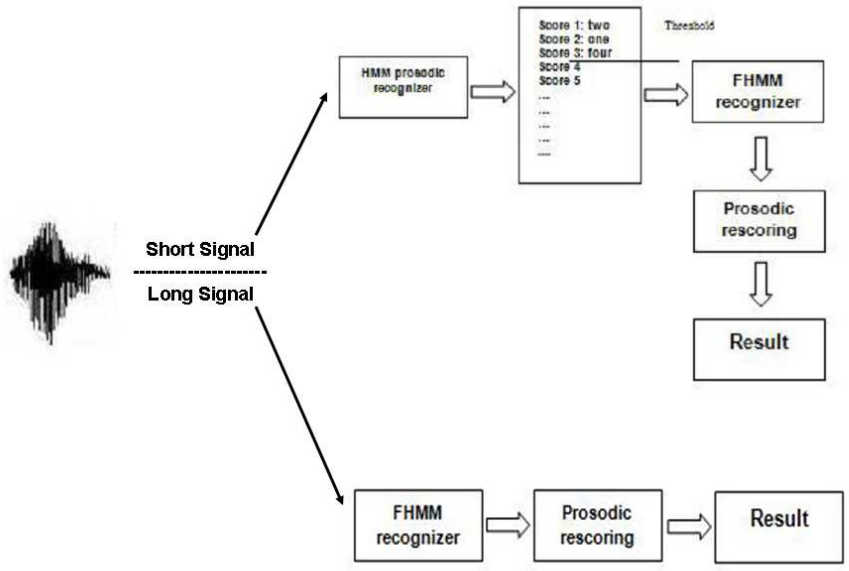



Figure 49: Overall ASR schema.

The next step, before any further work aiming to make the system perform better or faster, is the following question:

*Is it suitable to face real application problems, and in which context?*

First of all notice that the preliminar prosodic analysis, in the novel system, is able to accelerate the whole recognition process, even if such technique is effective only on "*short*" words recognition. The algorithms here presented have

a complexity strictly dependent on the number of frames to analyze and the number of models involved into the task. It would be impossible to apply this ASR to a spontaneous speech dialogue, without any further work in the direction of complexity reduction.
On the other side, the power of multi-granular processing, makes the ASR suitable to application addressing *large public*, where speaker independence and robustness requirements are necessary.
In summary the enviroment for a straightforward application of the here presented model must have the following characteristics:

- Words to be recognized have to be *short*, for the response to be faster.
- Few words in the dictionary are needed, because the decoding strategy complexity is strictly dependent on that.
- The environment needs to face large speaker variety and has robustness requirings.

The landscape above could seem very restrictive, but that is not completely true. Many applications in real word, do not need all the power of large vocabulary ASRs, while they really need something suitable for the task they are going to face. Powerful speech recognizers in commerce, such Nuance [38] or Loquendo [59] products, are hard to adapt to small tasks where the recognizer has a little set of words among which to choose, but the variety of speakers is really large, as well as a very noisy enviroment is generally present. Think for example to an application for an energy supply company, which has to develop an automatic telephonic responser for its services. In that case, people variety is very high in pronunciation and use to automatic services, moreover noisy environment will be surely present, expecially when the call comes from people in street traffic.
A task specific recognizer for digits was introduced by Avaya [48] in 2004, with the Interactive Responser platform. The machine accounted for automatic telephony answering, for call flows filtering and management. The applications running on the platform were able to integrate also ASR or Text-To-Speech facilities during the interaction with the caller. For digits recognition, a task specific ASR, called "*whole word*" recognizer, was introduced. This system presented a single HMM for each digit, and classified on the basis of the highest likelihood. Usually on such tasks, whole word HMMs perform better than systems

addressing to sub-word units, but this is not the case for telephony applications. The "*whole word*" failed when in presence of a large number of customers and people preferred the interaction by means of DTMFs, moreover problems rose when in presence of "*barge-in*", that is when a user was allowed to interrupt the machine speaking. A demostration was given by the Interactive Response System of **A.E.M.** (a company for electricity and gas furniture in Milan, Italy), which could accept up to 120 contemporary calls. Each menu of the response system had a *barge-in-able* prompt, and interaction by means of DTMF or voice was possible. Statistics showed that over the *70%* of the callers preferred the DTMF interaction because the whole word recognizer was really not performant from their point of view, and people didn't like the time loss due to recognition errors. The issue was only on speech recognition, because the agreement on the overall service was over the *90%*. Simple structure recognizers are not suitable to such applications, so that a more powerful tool has to be used. The *whole word* recognizer is not embedded in the Avaya IR systems anymore, and its use is discouraged for customer applications.

From the discussion above, it could be argued that the introduction of a multi-granular ASR could be suitable on such tasks, which represent the largest part of today automatic telephonic responsers using speech recognition facilities.

## 9.2 ASR for an IVR application

The most suitable ASR schema for an IVR application for isolated digits recognition could be that in figure 50.

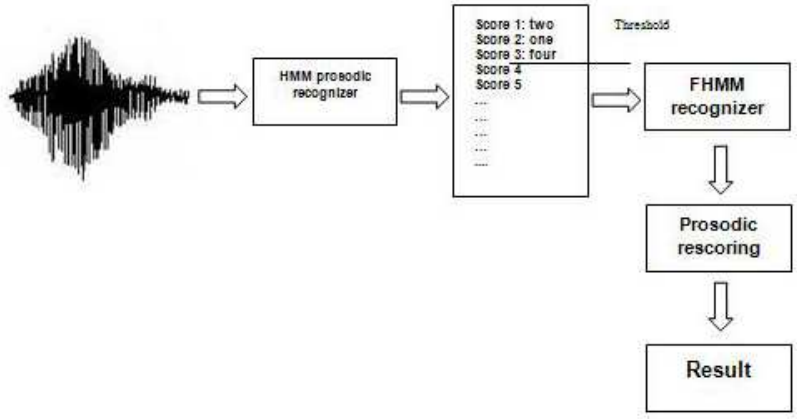



Figure 50: Multi-Granular ASR for isolated digits recognition.

A single digit is short enough to meet the above requirements, and the multi-granular structure can account for the robustness requirements. The role of the preliminary prosodic analysis is to make the computation fast enough for practical applications.
Problems can arise in the recognition of sequences of digits. In that case, a long sequence can be viewed as a long word, where the constituting elements are more easily separable. So that the schema of an ASR could be the one in figure 51.

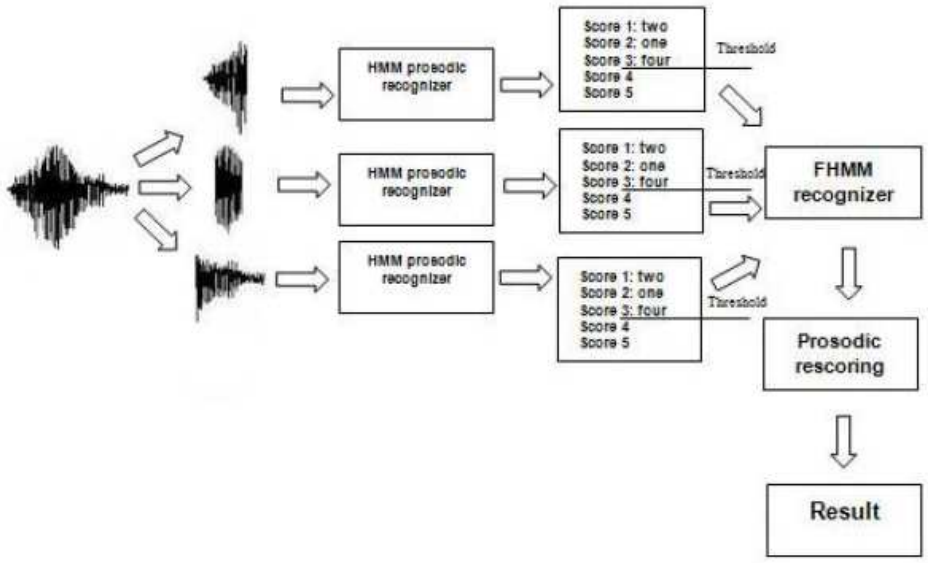



Figure 51: Multi-Granular ASR for connected digits recognition.

Obviously the last schema should be verified and results are necessary to state its efficacy on such task.

## 9.3 The Environment

A possible enviroment for the employment of a multigranular ASR is that of IVR applications. In the most cases a user is asked to choose among a finite set of possibilities, by voice or touchtones. A typical IVR system is depicted in figure 52.
The application lies on a web server while a VoiceXML interpreter manages the call details. A telephonic interface has the duty to answer the call and use the ASR and Text-To-Speech features.
When the user is asked to interact, the telephonic interface gets the audio input and processes it, in order to set the start and the end of speech. The recognition process acts while the caller is speaking, or sometimes in "batch" mode, after the user has finished to talk.

A multigranular ASR can act in this situation when the choice is on a menu or there is a little finite set of words, associated to some actions.

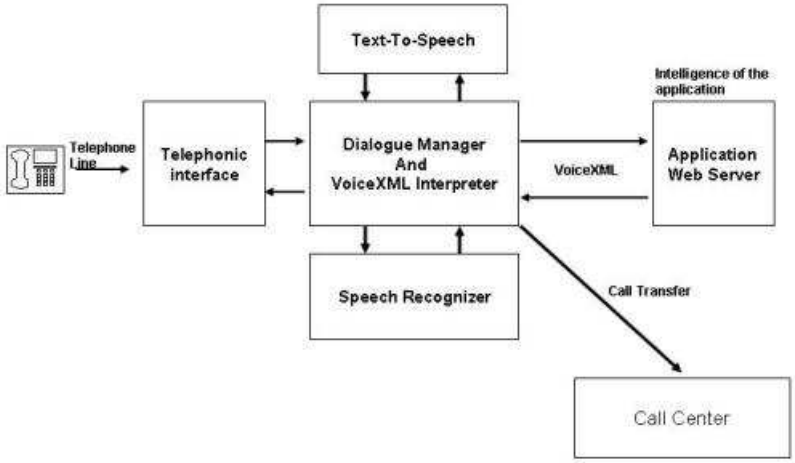



Figure 52: General IVR Architecture.

The most IVR systems are oriented to such applications, where the set of possibilities is very narrow, but a very powerful ASR is needed.
Even if the interactive responsers are moving towards the direction of artificial agents, this aim is very far to come. This is due to the lack in performances of modern commercial speaker independent ASRs. Nuance [38] has abandoned the idea to build up a telephony speaker independent recognizer with large vocabulary, so that spontaneous dialogues are even discouraged by the builders themselves. This world is so suitable to the introduction of task specific ASRs, like multi-granular ones.
Dialogues design literature is wide and all the advices are oriented to build up applications using fragmentation of information, so that the user has to answer to simple questions and has to use a simple set of words. This is because telephony applications are oriented to people not used to this technology, which have many difficulties to interact with a machine. Such users need the most powerful speech recognition processing techniques, and the use of general purpose systems is not revealing successful. Ignoring the needs of practical applications is not useful, because success and people education on such systems start from simple

but functioning cases. Common people interest and trust, on these solutions, are the most important parameters for companies investments.

## 9.4 Proposals

The above reasons have been the principal impulse for some companies to set up a connection network between IVR solutions developers. The aim is to create more robust applications and multichannel services, which can integrate many different worlds and technologies. The trend is towards the invention of interactive agents which can be really useful and very fast. The main problem of standard applications is that the users choices are very controlled and no much freedom is left to the caller about the expression of his needs. A sequence of menus can be really tyring and the user is not always able to find the point he is searching for. The possibility of free speech requires much research, but the technology nowadays is ready for the development of robust applications in that way. The problem is that information and techniques have to be shared.
Examples of developers networks are given by companies like Avaya, Genesys [56] and HP . The first has introduced an international program (Avaya DevConnect [49]), which aims to bring as many natural language processing skills as possible, as well as telephony integrations for speech recognition and synthesis. As the author of the tesis is a member of this program, the here presented project has been submitted to the attention of the other members, in order to make the multi-granular ASR the successor of the old *whole word* recognizer, and there are high probabilities the new approach will be experimented as the next embedded recognizer for simple words or digits on the Avaya IR platforms. Other products oriented to natural language processing have been successfully accepted as official IVR solutions, about probabilistic grammars developments and about Computer-Telephony integrations [60].
The proposed solution is so the insertion of a multigranular ASR on IVR platforms in order to test the performances on a large public.
In summary, the result of this thesis has not only been the exploration of the multigranular theory, in order to get evidence of was stated in psychoacoustics and linguistics, but it has become also a useful tool which will be experimented on real cases, with a large quantity and variety of speakers, in order to test the usefulness, other than the robustness of the model. The perception of the good performances of a system is intersected with its robustness, but the two aspects

do not coincide.

## Further Reading

It is not possible to quote all the literature about multi-granularity, so, before the reference section, adviced lectures for a better understanding of the contents of this thesis will be listed.

In acoustics and psychoacoustics, there are important works referring to human speech recognition as a multiple sources concurrency:

*Prosodic organization of speech based on syllables: the C|D model,*

O. Fujimura, In Proceedings of the XIIIth International Congress of Phonetic Sciences, V. 3, p. 10-17, 1995.

*Speech Perception: New directions in research and theory,*

L. C. Nygaard, D. B. Pisoni, In Joanne L. Miller and Peter D. Eimas, (Eds.) Speech, Language and Communication, Vol. 11 of Handbook of Perception and Cognition, Ch. 3 , pp. 63-96. Academic Press, 1995.

*The Temporal Unfolding of Local Acoustic Information and Sentence Context,*

S. Borsky, L. P. Shapiro, B. Tuller, Journal of Psycholinguistic Research, 29, 155-168.

*The vowel-sequence illusion: Intrasubject stability and intersubject agreement of syllabic forms,*

R. M. Warren, E. W. Healy, M. H. Chalikia, Journal of the Acoustical Society of America, 100(4): 2452-2461, October 1996.

*The origins of speech intelligibility in the real world,*

S. Greenberbg, In Proceedings of the ESCA Workshop on Robust Speech Recognition for Unknown Channels, p. 23-32. ESCA, 1997.

*Perceptual processing of speech and other perceptual: some similarities and differences,*

R. M. Warren, In Steven Greenberg and William Ainsworth, Eds. Listening to Speech: An Auditory Perspective. Oxford University Press,1998.

*Syllable timing computation in the $C|D$ model,*

O. Fujimura, In ICSP, pp. 519-522, 1994.

*Phonetics and Phonology,*

J. Clark, C. Yallop, Chapter 5, p. 124-127,287. Basil Blackwell, 1990.

Multigranular models have been experimented in other ways, aside the one here showed. The followings are other works, not mentioned in Chapter I, which can make the reader know other paradigms in ASR building:

*Architetture parallele basate su modelli nascosti di Markov per il riconoscimento di numeri,*

F. Persico, Thesis in Informatics at Università degli Studi di Napoli Federico II aa. 2004/05.

*Multiresolutional hierarchical decision support systems,*

A.M. Meystel, Systems, Man and Cybernetics, Part C: Applications and Reviews, 2003.

*Speech Recognition by Composition of Weighted Finite Automata,*

F. C. N. Pereira, M. D. Riley, Available as cmp-lg/9603001 from http://xxx.lanl.gov/cmp-lg.

*Merging information in speech recognition: Feedback is never necessary,*

D. Norris, Medical Research Council Cognition and Brain Sciences Unit

The literature about ASRs using multiple knowledge sources is wide. Many people have used several kinds of information to improve recognizers performances, not referring to a multi-granular inner structure of speech. The followings are examples of those experiments:

*Use of word level side information to improve speech recognition,*
D. Vergyri, In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2000.

*Using Natural language knowledge sources In Speech Recognition,*
R. Moore, In Keith Ponting, editor, Speech Pattern Processing ,1999.

*Using multiple time scales in a multi-stream speech recognition system,*
S. Dupont, H. Bourlard, C. Ris, In Eurospeech 1997, pp. 3-6.

*Syllabe segmentation of continuos speech with artificial neural networks*,
W. Reichl and G. Ruske, in Proceedings of Eurospeech 93, 3rd European Conference on Speech Communication and Technology, Berlin, pp. 1771-1774, 1993.

*Syllabe detection and segmentation using temporal flowneural networks*,
L. Shastri , S. Chang , S. Greenberg, Proceedings of the Fourteenth International Congress of Phonetic Sciences, San Francisco. 1999.

*Syllabe segmentation of continuos speech with artificial neural networks*,
W. Reichl and G. Ruske, in Proceedings of Eurospeech 93, 3rd European Conference on Speech Communication and Technology, Berlin, pp. 1771-1774, 1993.

*Syllabe detection and segmentation using temporal flowneural networks*,
L. Shastri , S. Chang , S. Greenberg, Proceedings of the Fourteenth International Congress of Phonetic Sciences, San Francisco. 1999.

*Using Dialog-Level Knowledge Sources to Improve Speech Recognition,*
A. G. Hauptmann, S. R. Young, W. H. Ward, National Conference on Artificial Intelligence,p. 729-733,1988.

*A comparison of Statecharts step semantics,*
A. Maggiolo-Schettini, A. Peron, S. Tini, Theoretical Computer Science, 2 October 2001.

*Connected word recognition using whole word templates,*
J. S. Bridle and M. D. Brown, Proc. Inst. Acoust., pp. 25-28, 1979.

*High Level Knowledge Sources in usable speech recognition systems,*
S. R. Young, A. G. Hauptmann, W. Ward, E. Smith, P. Werne, Communications of the ACM archive V. 32, Issue 2 1989

Many works on prosody and non-segmental features have shown their usefulness in speech recognition. The followings are examples of this approach in speech recognition and understanding:

*A prosodical guided speech understanding strategy,*
W. A. Lea, M. F. Medress, T. E. Skinner, IEEE Transactions on Acoustic, Speech and Signal Processing, 38(1), p. 35-45, 1990.

*Automatically predicting dialogue structure using prosodic features,*
H. W. Hastie, M. Poesio, S. Isard ,Speech Communication V. 36 , 2002.

*Phonetic characterisation and lexical access in non-segmental speech recognition,*
M. Huckval, Proc. 13th Int. Congress.Phonetic Sciences, 1995.

*Using high level dialogue information for dialogue act recognition usign prosodic features,*
H. Wright, M. Poesio, S. Isard, Int. Journal of Network Management, V. 9 (2), p. 118-125, 1999.

*Word Fragment Identification Using Acoustic-Prosodic Features in Conversational Speech,*
Y. Liu, Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human
Language Technology.

*Speech Processing Based on Syllable Identification by using Phonological Patterns,*

A. Tanaka, S. Kamiya, Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86.

*Syllable-level desynchronisation of phonetic features for speech recognition,*

K. Kirchhoff, Speech Communication archive, V. 49 (2), 2007.

*Using Prosodic Features of Speech and Audio Localization in Graphical User Interfaces,*

A. Olwal, S. Feiner, Columbia University Department of Computer Science, Technical Report CUCS-016-03, 2003.

*Prosodic structure and spoken word recognition*,

F. Grosjean and J. P. Gee., Cognition Special Issue, pp. 135-155. MIT Press, 1987.

*Feature-based Pronunciation Modeling for Speech Recognition,*

K. Livescu, J. Glass, In Proc. HLT/NAACL, 2004.

*Recognizing reverberant speech with RASTA-PLP*,

B. Kingsbury and N. Morgan, ICASSP, vol 2, pp. 1259-1262, Munich, Germany, 1997, IEEE.

*Tweaking the lexicon: Organization of vowel sequences into words,*

R. M. Warren, J. A. Bashford, D. A. Gardner, Perception and Psichophysics. 47(5): 423-432, 1990.

*Continuous Speech Recognition of Japanese Using Prosodic Word Boundaries Detected by Mora Transition Modeling of*

*Fundamental Frequency Contours,*

K. Hirose, N. Minematsu, Y. Hashimoto, K. Iwano.

*Robust speech recognition using the Modulation Spectrogram,*

B. E. D. Kingsbury, N. Morgan, S. Greenberg. Speech Comunication. 1998.

*Perceptually inspired signal processing strategies for robust speech recognition in reverberant environments,*
B. E. D. Kingsbury, PhD thesis, UC Berkeley, 1998.

*APA an object oriented system for automatic prosodic analysis,*
M. Petrillo, PhD Thesis at Università di Napoli Federico II,1999.

*On the Robust Automatic Segmentation of Spontaneous Speech,*
B. Petek, O. Andersen, P. Dalsgaard, Proc. of ICSLP '96, V. 2, 1996.

*Un sistema automatico per la localizzazioni delle zone formantiche nella identificazione del parlante,*
G. Coro, M. Falcone, Proc. of AISV 2005.

The following are examples of signal processing techiques for enhancing ASR performances:

*Efficient coding leads to novel features for speech recognition,*
W. Smit, E. Barnard, 5th Annual Symposium of the Pattern Recognition Association of South Africa, p. 25-26, 2004.

*Articulatory Motivated Acoustic Features for Speech Recognition,*
D. Kocharov A. Zolnay, R. Schluter, H. Nev, Proc. of Interspeech 2005.

*Autoencoders, minimum description length, and Helmholtz free energy,*
G.E. Hilton, R.S. Zernel In J. D. Cowan, G. Tesauro, & J. Alspector (Eds.), Advances in neural information processing systems 6, pp. 3-10. San Francisco, CA: Morgan Kaufmann, 1994.

*Signal processing for robust speech recognition,*
R. M. Stern, F. H. Liu, P. J. Moreno, A. Acero,Kluwer Academic Publ., p. 351-378,1996.

In this thesis, a new kind of acoustic model has been introduced for syllables recognition. The followings are other experiments in new representations of speech units:

*Comparison of Parametric Representations for Monosyllable Word Recognition in Continuously Spoken Sentences,*

S. Davis, P. Mermelstein, IEEE Trans. on Acoustics, Speech and Signal Processing, 28(4), pp. 357-366, 1980.

*Acoustic model clustering based on syllable structure*,

I. Shafran and M. Ostendorf, Computer Speech and Language 17 (2003), p. 311-328.

*A survey of Discriminative and Connectionist methods for Speech Processing*,

D. Aberdeen, 2002.

*Integrating thumbnail features for speech recognition using conditional exponential models,*

H. Y., A. Waibel, Proc. of Acoustics, Speech, and Signal Processing, V. 1, I- 893-6, 2004.

Factorial HMMs carry an entire world rotating around them. It could have been impossible to resume all the techniques in this thesis. The followings are lectures on HMMs and Factorial HMMs, which make clear the whole statistical analysis performed by these machine learning systems, as well as their properties:

*Factorial learning and the EM algorithm,*

Z. Ghahramani, In G. Tesauro, D.S. Touretzky, T.K. Leen (Eds.), Advances in neural information processing systems 7, MA: MIT Press, 1995, p. 617-624.

*A tutorial on Hidden Markov Model and selected Applications in speech recognition,*

L. Rabiner, Proceeding of the IEEE, vol 77, n. 2, 1989, pp. 257-266.

*Maximum likelihood from incomplete data via the EM algorithm,*

A. Dempster, N. Laird, D. Rubin , Journal of the Royal Statistical Society Series B, 39, p. 1-38, 1977.

*The Monte Carlo method,*

N. Metropolis, S. Ulam, J. Am. Statistical Association, vol. 44, pp. 335341, 1949.

*Software for Factorial Hidden Markov Models developments,*
Z. Ghahramani, www.gatsby.ucl.ac.uk/~zoubin/software.html

*Token Passing: a conceptual model for connected speech recognition systems,*

S.J. Young, N.H. Russell, J.H.S. Thornton, CUED Technical report F_INFENG/TR38, Cambridge University, 1989.

*An Introduction to Variational Methods for Graphical Model*,

M. I. Jordan, Z. Ghahramani, Proceedings of the NATO Advanced Study Institute on Learning in graphical models table of contents

Erice, Italy, p. 105 - 161.

*Factorial Hidden Markov Models for Speech Recognition: Preliminary Experiments*,

B. Logan, P. J. Moreno, Cambridge Research Laboratories Technical Report 97/7, 1997.

*A minimum description length framework for unsupervised learning,*

R.S. Zemel, Ph D. Thesis Department of Computer Science, University of Toronto, Toronto Canada, 1993.

*A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains,*

L. Baum, T. Petrie, G. Soules, N. Weiss, The Annals of Mathematical Statistics, 41, 1970, pp. 164171.

# References


[1] R. Bakis. Continuous speech recognition via centisecond acoustic states. *91st Meeting of the Acoustical Society of America*, 1976.

[2] H. B. Savin T. G. Bever. The non perceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior*, 9:295–302, 1970.

[3] K. Livescu J. Glass J. Bilmes. Hidden feature modeling for speech recognition using dynamic bayesian networks. *Proceedings of the Eurospeech*, 2003.

[4] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[5] Htk Book. *Manual of the HTK toolkit*. Cambridge University Engineering Department http://htk.eng.cam.ac.uk/, 2002.

[6] N Morgan H Bourlard. An introduction to hybrid hmm connectionist continuous speech recognition. *IEEE Signal Processing Magazine*, 1995.

[7] J. S. Bridle. Stochastic models and template matching: some important relationships between two apparently different techniques for automatic speech recognition. *Proc. Inst. Acoust. Autumn meeting*, 1984.

[8] F. Caropreso. Speech recognition by means of factorial hmms phd. thesis at university federico ii of naples. 2006.

[9] S. Chang. A syllable articulatory-feature and stress-accent model of speech recognition. *Ph.D. Thesis University of California Berkeley*, 2002.

[10] A. Slater J. Coleman. Non-segmental analysis and synthesis based on a speech database. *Proc. ICSLP 96*, 4:2379–2382, 1996.

[11] G. Coro. Il modulation spectrogram nel riconoscimento automatico del parlato. *Proceedings of AISV 2004*, 2004.

[12] Switchboard corpus. Recorded telephone conversations. *Sponsored by DARPA*, 1992.

[13] The SPEECON corpus. http://www.speechdat.org/speecon/project.htm.

[14] P. Cosi. Cslu toolkit il riconoscimento automatico del linguaggio naturale alla portata di tutti. *Atti delle IX Giornate di Studio del G.F.S*, pages 147–159.

[15] D. Cristal. *Prosodic System and Intonation in English.* Cambrigde University Press, 1969.

[16] D. Cristal. *Enciclopedia Cambridge delle Scienze del Linguaggio.* Zanichelli, 1998.

[17] G. Coro F. Cutugno. Il modulation spectrogram nel riconoscimento automatico del parlato. *Proceedings of AISV*, 2004.

[18] F. de Saussure. *Cours de linguistique generale.* Editions Payot, 1922.

[19] K. Duh. Jointly labeling multiple sequences: A factorial hmm approach. 2005.

[20] G. Coro M. Falcone. Un sistema automatico per la localizzazione delle zone formantiche nella identificazione del parlante. *Proceedings of AISV 2005*, 2005.

[21] O. Fujimura. Syllable as a unit of speech recognition. *IEEE Transactions on Acoustic Speech and Signal Processing*, 1:82–87, 1975.

[22] K. Iwano T. Seki S. Furui. Noise robust speech recognition using prosodic information. *DSP for In-Vehicle and Mobile Systems*, pages 139–152.

[23] Garzanti. *Dizionario della lingua italiana.* Garzanti Editore Garzanti Linguistica, 2000.

[24] S. Geman D. Geman. Stochastic relaxation gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[25] M. Jordan Z. Ghahramani. Factorial hidden markov models. *Machine Learning*, 29:245–273, 1997.

[26] Z. Ghahramani. Factorial learning and the em algorithm. *In G. Tesauro D.S. Touretzky T.K. Leen (Eds.) Advances in neural information processing systems*, 7:617–624, 1995.

[27] R. Silipo S. Greenberg. Automatic transcription of prosodic stress for spontaneous english discourse. *Proceedings of ICPhS-99*, 1999.

[28] S. Greenberg. Understanding speech understanding towards a unified theory of speech perception. *Workshop on Auditory Basis of Speech Perception ESCA*.

[29] H. Fujisaki K. Hirose. Modelling the dynamic characteristics of voice fundamental frequency with applications to analysis and synthesis of intonation. *In Preprints of the Working Group on Intonation 13th Intl. Congress of Linguists*, pages 57–70.

[30] X. Huang A. Acero H. Hon. *Spoken Language Processing.* Prentice Hall, 2001.

[31] A.N. Deoras M. Johnson. Factorial hmm approach to simultaneous recognition of isolated digits spoken by multiple talkers on one audio channel. *University of Illinois at Urbana-Champaign*, 2005.

[32] L. Saul M.I. Jordan. Exploiting tractable substructures in intractable networks. *In D. S. Touretzky M. C. Mozer M. E. Hasselmo (Eds.) Advances in neural information.*

[33] L. Saul T. Jaakkola M.I. Jordan. Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4:61–76, 1996.

[34] Z. Ghahramani M. Jordan. Factorial hidden markov model. *Machine Learning*, 29, 1997.

[35] D. Kahn. *Syllable based Generalizations in English Phonology. Outstanding Disertations in Linguistics.* Garland Publihing, 1980.

[36] S. Greenberg B. E. D. Kingsbury. The modulation spectrogram: In pursuit of an invariant representation of speech. *Proceedings of ICASSP*, 3:1647–1650, 1997.

[37] P. Ladefoged. *A Course in Phonetics.* Hartcourt Brace Jovanovich, 1993.

[38] Nuance The leading supplier of speech recognition Imaging PDF and OCR Solutions. www.nuance.com.

[39] D. W. Massaro. Perceptual images processing time and perceptual units in auditory perception. *Psychological Review*, 2:124–145, 1972.

[40] B. Logan P. Moreno. Factorial hmms for acoustic modeling. *Cambridge Research Laboratories Digital Equipment Corporation*, 1998.

[41] B. Kingsbury N. Morgan. Recognizing reverberant speech with rasta-plp. *Proceedings of ICASSP*, 2:1259–1262, 1997.

[42] D. Kocharov A. Zolnay R. Schlüter H. Ney. Articulatory motivated acoustic features for speech recognition. *Proceedings of the European Conference on Speech Communication and Technology Interspeech*, pages 1101–1104, 2005.

[43] A. V. Oppenheim. *Digital Signal Processing*. Prentice Hall, 1975.

[44] N.M Veilleux M. Ostendorf. Probabilistic parse scoring with prosodic information. *Acoustics Speech and Signal Processing ICASSP*, 2:51–54, 1993.

[45] A. Ganapathiraju J. Hamaker J. Picone. Syllable-based large vocabulary continuous speech recognition. *Transaction on Speech and Audio Processing*, 9, 2001.

[46] L. C. Nygaard D. B. Pisoni. Speech perception: New directions in research and theory. *Speech Language and Communication*, 11:63–96, 1995.

[47] D. Poeppel. The analysis of speech in different temporal integration windows: Cerebral lateralization as asymmetric sampling in time. *Speech Communication*, 41:245–255, 2003.

[48] Avaya Communications Enabled Business Processes. www.avaya.com.

[49] Avaya Developer Connection Program. deconnect.avaya.com.

[50] L.D. Erman F. Hayes-Roth V.R. Lesser D.R Reddy. The hearsay-ii speech understanding system: Integrating knowledge to resolve uncertainty. *Blackboard Systems*, pages 31–86, 1988.

[51] S2S. Sound to sense sense to sound http://www.s2s2.org/.

[52] K. Hirose K. Iwano A. Sakurai. Use of prosodic features in speech recognition. *Proceedings of 1996 IEEE Invited Workshop on Pattern Recognition for Multimedia Techniques*, 10:99–108, 1996.

[53] C. Wang S. Seneff. Lexical stress modeling for improved speech recognition of spontaneous telephone speech in the jupiter domain. *Proceedings of the 7th European Conference on Speech Communication and Technology*.

[54] D. Vergyri A. Stolcke V. Gadde L. Ferrer E. Shriberg. Prosodic knowledge sources for automatic speech recognition. *Proceedings of ICASSP*, pages 208–211, 2003.

[55] S. Hawkins R. Smith. Polysp: a polysystemic phonetically-rich approach to speech understanding. *Rivista di Linguistica*, pages 99–189, 2001.

[56] Genesys The World's Number 1 Contact Center Software. www.genesyslab.com.

[57] E. Shriberg A. Stolcke. Prosody modeling for automatic speech recognition and understanding. *Proceedings of Workshop on Mathematical Foundations of Natural Language Modeling*, 2002.

[58] S. King T. Stephenson S. Isard P. Taylor A. Strachan. Speech recognition via phonetically featured syllables. *Proceedings of ICSLP*, 2:124–145, 1972.

[59] Loquendo Vocal Technology and Services. www.loquendo.com.

[60] Abla Beyond the Voice. www.abla.it.

[61] The Wavesurfer Tool. http://www.speech.kth.se/wavesurfer/.

[62] N. Metropolis S. Ulam. The monte carlo method. *J. Am. Statistical Association*, 44, 1949.

[63] A. P. Varga. Speech recognition. *Proc. IEEE*, 1979.

[64] IMA 2000 workshop. http://www.ima.umn.edu/reactive/spring/tm.html. 2000.

[65] S. L. Wu. Incorporating information from syllable-length time scales into automatic speech recognition. *PhD thesis, ICSI*, 1998.

[66] G. Yule. *Introduzione alla linguistica*. Edizioni Il Mulino, 1997.

[67] L. D'Anna M. Petrillo F. Cutugno E. Zovato. Apa: towards an automatic tool for prosodic analysis. *In Acts of Speech Prosody*, pages 231–234, 2002.