Auditory Modeling and Neural Networks

Piero Cosi

Istituto di Fonetica e Dialettologia
C.N.R. Padova (Italy)
Via G. Anghinoni, 10 - 35121 Padova (Italy)
Phone +39 49 8274418
Fax: +39 49 827 4416
Email:cosi@csrf.pd.cnr.it

Abstract. The human physiology of hearing is briefly described and various computational models of peripheral auditory processing are illustrated together with their use in different speech processing applications. Finally the automatic bi-modal recognition of simple speech stimuli by an artificial neural network working on specific auditory processing parameters in cooperation with some visual articulatory cues is described.

1 Introduction

Acoustic analysis front-ends of almost all presently commercialized Automatic Speech Recognition (ASR) systems are built using speech "production-based" processing schemes.

In other words, Short-Time Fourier Transform (STFT), Cepstrum, and other related Speech Processing (SP) [1] schemes were all developed strictly considering physical phenomena that characterize the speech waveform obtained by the electrical transduction of the sound pressure wave.

Moreover LPC technique [2] and all its variants were developed directly by modeling the human speech production mechanism. In the last years, almost all these analysis schemes have been modified by incorporating, at least at a very general stage, various perceptual-related phenomena. Linear prediction on a warped frequency scale [3], STFT-derived auditory models [4], perceptually based linear predictive analysis of speech [5], [6] are few simple examples of how human auditory perceptual behavior is now taken into account while designing new signal representation algorithms. Furthermore, the most significant example of attempting to improve acoustic front-end with perceptual related knowledge, is given by the Mel-frequency cepstrum analysis of speech [7], which transforms the linear frequency domain into a logarithmic one resembling that of human auditory sensation of tone height. In fact, Mel Frequency Cepstrum Coefficients (MFCC) are almost universally used in the speech community to build acoustic front-end for ASR systems.

All these speech processing schemes make use of the "short-time" analysis framework [1]. Short segments of speech are isolated and processed as if they were short segments from a sustained sound with fixed properties. In order to better track

dynamical changes of speech properties, these short segments which are called analysis frames, overlap one another. This framework is based on the underlying assumption that, due to human articulatory characteristics, the properties of the speech signal change relatively slowly with time. Even if overlapping analysis windows are used, important fine dynamic characteristics of speech signal are discarded. Just for that reason, but without completely solving the problem of taking into account the dynamic properties of speech correctly, "velocity"-type parameters (simple differences among parameters of successive frames) and "acceleration"-type parameters (differences of differences) [8] have been recently included in acoustic front end of almost all ASR systems found on the market. The use of these temporal changes in speech spectral representation (i.e. ΔMFCC, ΔΔMFCC) has given rise to one of the greatest improvements in ASR systems. In some of the best ASR systems, the incorporation of transitional information has reduced errors by as much as 50%. [9], [10].

Moreover, in order to overcome the resolution limitation of the STFT (due to the fact that once the analysis window has been chosen, the time frequency resolution is fixed over the entire time-frequency plane, since the same window is used at all frequencies), a new technique called Continuous Wavelet Transform (CWT), characterized by the capability of implementing multiresolution analysis, has been recently introduced [11]. With this new speech processing scheme, if the analysis is viewed as a filter bank, the time resolution increases with the central frequency of the analysis filters. In other word, different analysis windows are simultaneously considered in order to simulate more precisely the frequency response of the human cochlea. As with the preceding processing schemes, this new auditory-based technique, even if it is surely more adequate than STFT analysis to represent a model of human Auditory Speech Processing (ASP), it is still based on a mathematical framework built around a transformation of the speech waveform, from which it tries directly to extrapolate a more realistic perceptual behavior.

Cochlear transformations of speech signals result in an auditory neural firing pattern significantly different from the spectral pattern obtained from the speech waveform by using one of the above mentioned techniques. In other words, speech spectral representations such as the spectrogram, a popular time-frequency-energy representation of speech, or the wavelet spectrogram (scalogram), obtained using the above described multiresolution analysis technique are quite different from the true neurogram. In recent years, basilar membrane, inner cell and nerve fiber behavior have been extensively studied by auditory physiologists and neurophysiologists and knowledge about the human auditory pathway has become more accurate. A number of studies have been accomplished and a considerable amount of data has been gathered in order to characterize the responses of nerve fibers in the eighth nerve of the mammalian auditory system using tone, tone complexes and synthetic speech stimuli [12], [13], [14], [15], [16], [17], [18], [19], [20], [21]. Phonetic features probably correspond in a rather straightforward manner to the neural discharge pattern with which speech is coded by the auditory nerve.

Various auditory models which try to physiologically reproduce the human auditory system have been developed in the past [22], and, even if they must be considered as only an approximation of physical reality, they appear to be a suitable system for identifying those aspects of the speech signal that are relevant for

automatic speech analysis and recognition. Furthermore, with these models of auditory processing, perceptual properties can be re-discovered starting not from the sound pressure wave that characterizes speech, but from a more internal representation, which is intended to represent the true information available at the eighth acoustic nerve of the human auditory system.

Advanced Auditory Modeling (AM) techniques not only follow "perception-based" criteria instead of "production-based" ones, but also overcome "short-term" analysis limitations, because they implicitly retain dynamic and nonlinear speech characteristics. For example, the dynamics of the response to non-steady-state signals, such as "forward masking" phenomena, which occur when the response to a particular sound is diminished as a consequence of a preceding, usually considerably more intense signal, are important aspects captured by efficient auditory models [23]. Various evidences can be found in the literature [24], [25], [26], [27] suggesting the validity of using AM techniques, instead of more classical ones, in building speech analysis and recognition systems. Especially when speech is greatly corrupted by noise [27], [28], [29] the effective power of AM techniques seems much more evident than that of classical digital signal processing schemes.

2 The Human Auditory System[1]

(Text and figures [30] by permission of William E. Brownell: brownell@bcm.tmc.edu)

Human social structures rely on speech communication, which requires the sensitive, rapid processing of acoustic energy that the normal inner ear provides. Encased in the hardest bone of the body, the ear contains the smallest bones, the smallest muscles, and the smallest, yet one of the most elegant organs of the body, the cochlea (part of the inner ear). The task of all hearing organs is to analyze environmental sounds and transmit the results of that analysis to the brain for their interpretation. All sensory organs have specialized sensory cells that convert an environmental signal into electrical energy [31]. The change in electrical energy is then converted to a type a digital code that is transmitted to the brain. The human auditory system performs an analysis of sound entering the ear prior to the conversion to the neural code. The inner ear first determines how much energy is contained at the different frequencies that make up a specific sound. The cochlea [32] is designed so that it is most sensitive to a specific frequency at one location and most sensitive to another frequency at another. These different locations then transmit information to the brain. This "mapping" of frequency information is just one of several strategies that the ear uses to code incoming information. The frequency analysis of environmental sounds begins in the external ear.

[1] Text and figures of this chapter are extracted by the electronic publication entitled "How the Ear Works. Nature's Solutions for Listening" presented at the www address http://www.bcm.tmc.edu/oto/research/cochlea/Volta/index.html, by permission of William E. Brownell (brownell@bcm.tmc.edu), Ph.D. at the Department of Otorhinolaryngology and Communicative Sciences, Baylor College of Medicine, Houston, Texas, TX 77030.

2.2 The Middle and External Ear - the Analysis of Sound Begins

When sound passes from one media to another (as, for example, from air to water) some energy is reflected by the surface and does not pass to the new media. In order to reduce these reflections and maximize the transfer of sound energy from the air filled environment to the fluid filled inner ear, land animals evolved external ears as sound collectors and middle ears as mechanical force amplifiers. Fig. 1 shows the path that sound waves follow from the sound source where they are generated to the inner ear.

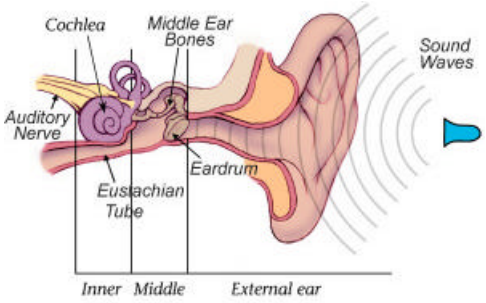

Fig. 1. Pressure waves from the speaker pass through the air to the external ear, which collects the sound and passes it to the eardrum. The middle ear consists of the eardrum, the middle ear bones, and the membrane over the oval window at the entrance to the inner ear. The cochlea of the inner ear is named with the Greek word for "snail" because of its spiral shape.

The outer portion of the external ear reflects sound towards the ear canal. Once in the ear canal, the pressure waves are aligned so they strike the eardrum at right angles. The reflection of sounds of different frequency is not the same and as a result the relative amplitude of some frequencies is greater than others. The result is that the relative amplitude of different frequencies at the eardrum differs, even if sound begins at the same intensity for all frequencies. Modification of the original sound by the external ear is a type of analysis that the brain learns to interpret. The frequency composition of familiar sounds aids your auditory system in determining where a sound is coming from. The middle ear bones conduct sound from the eardrum to the fluids of the inner ear via a small region called 'oval window'. The eardrum is bigger than the oval window. The decrease in the area of these two membranes leads to an increase in pressure. The middle ear bones act as mechanical levers and further increase the pressure of the sound at the entrance to the cochlea. All of this is necessary to maximize the sound energy that gets to the fluids of the inner ear. The pressure balance with the environment forced by the tube (called the eustachian tube) connecting the middle ear to the nose, allows the eardrum to vibrate freely.

2.3 The Inner Ear

The inner ear contains the sensory systems of balance and hearing. Its location is close to the center of the skull and it is encased in the hardest bones in the body,

which make it one of the best-protected sensory systems. The function of the auditory portion of the inner ear is that of telling the brain how much energy is contained in an environmental sound and at what frequencies that energy is located. The inner ear is divided into two fluid-filled chambers - one inside the other. Fig. 2 illustrates the basic organization of both the organs of hearing and balance.

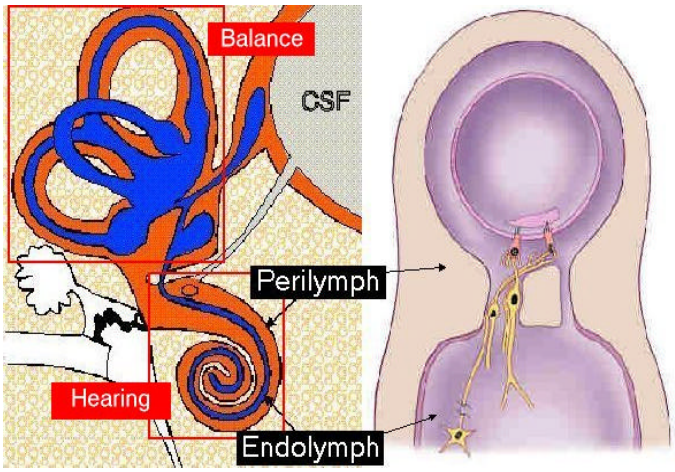

Fig. 2. Simplified diagrams showing the organization of inner ear organs of hearing and balance.

The difference in the chemical composition of these two fluids is maintained by specialized cells and provides chemical energy (like a battery) that powers the activities of the sensory cells. The wall of the membranous chamber is made up of many cells that are so tightly joined together that they prevent the two fluids from mixing. The sensory epithelium makes up only a small portion of the wall of the membranous chamber and contains sensory receptor and surrounding cells. The sensory cells are called *hair cells* because of their appearance under the microscope, as illustrated in Fig. 3, in fact they have a tuft of projections (*stereocilia*) at one end that looks like a bad haircut.

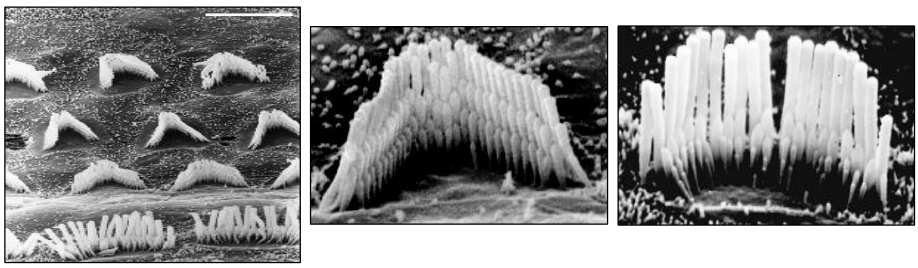

Fig. 3. From left to right, the apical surface of inner support cells, one row of inner hair cells 3 rows of outer hair cells, the stereocilia patterning of OHCs (center) and IHCs (right).

Hair cells, the sensory receptor cells of hearing, are mechanotransducers that convert the mechanical stimuli associated with hearing into neural information for

transmission to the brain. They have synapses, which are structures that permit communication between neuronal cells, located at the end of the cell opposite the stereocilia bundle. One side of the synapse is "*presynaptic*" and the other "*postsynaptic*". A chemical known as a "*neurotransmitter*" is secreted from the presynaptic cell and changes the membrane potential of the postsynaptic cell. "*Afferent*" synapses convey information into the central nervous system by exciting *"action potentials*" (pulses that travel down the fiber and carry information in a type of digital code) in the afferent nerve fibers that enter the brain. "*Efferent*" synapses modulate the membrane potential of the hair cell in response to neurotransmitter release from their presynaptic element, which is the terminal of a nerve fiber that originates deep in the brainstem. The neural signals from the brain conveyed by these efferent fibers may be viewed as having the ability to change the "*gain*" (amplification), in other words to regulate the sensitivity, of the hair cells they innervate. Neurotransmitter release at an afferent synapse is regulated by changes in the membrane potential of the hair cell in response to "*bending*" its stereocilia bundle. Each hair cell therefore codes the direction and degree of stereocilia bundle bending by either increasing or decreasing the firing rate of the postsynaptic afferent fiber in proportion to the magnitude of the bend. The inner ear sensory epithelium is among the smallest organs in the body, containing less than 20,000 sensory cells.

The hearing organ in mammals is a spiraling structure (2½ turns, nearly an inch in length, if unwounded) called the "cochlea" from the Greek word for snail. The cochlea originates from one of the balance organs, as illustrated in Fig. 4, and contains the sensory epithelium for hearing.

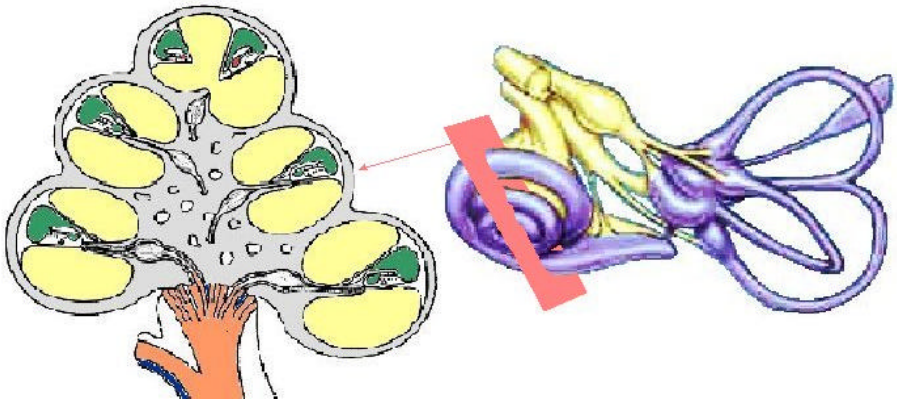

Fig. 4. A drawing of the bony chambers of the inner ear (right) and a cross section of the whole cochlea (left).

The nerve is made up of the neuronal projections that connect the hair cells with the brain and is called the eighth nerve because it is one of 12 nerves that come off the brain in the skull. The sensory epithelium of the inner ear is called the organ of Corti after the Italian scientist who first described it. Its orderly rows of outer hair cells (see Fig. 3) are unique among the organs of the body. Fig. 5 shows a short section of the organ of Corti as it spirals in the cochlea. The organ of Corti is larger and the basilar

membrane on which it sits is longer as it gets further away from the base of the cochlea. This difference in size is consistent with the fact that different frequencies of sound result in greater vibrations of the organ of Corti depending on where along the length of the cochlea you are measuring. The shorter, smaller structures near the base of the cochlea respond best to high frequencies, while the longer, larger structures near the top of cochlea respond best to low frequencies.

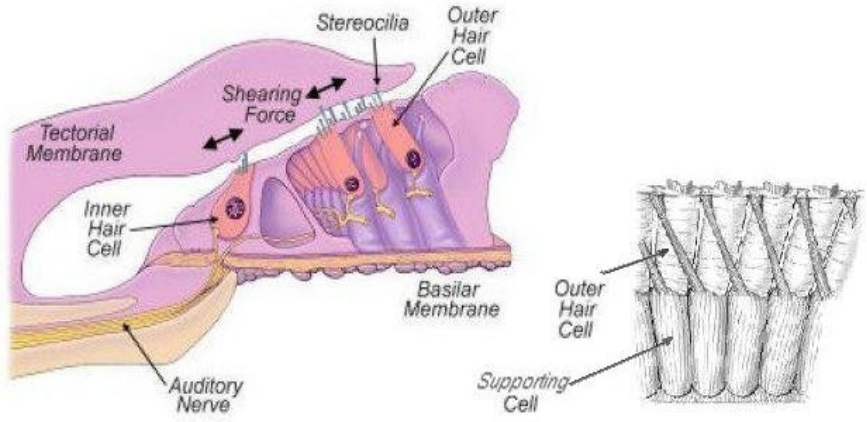

Fig. 5. A drawing of the full organ of Corti and of a particular of single row of outer hair cells.

The organ of Corti is made up of hair cells and supporting cells that sit on a flexible *basilar membrane* which is anchored to the bony shelf on the left and a ligament (not shown in the figure) on the right. A single flask shaped inner hair cell is shown on the left and three rows of cylindrically shaped outer hair cells are seen on the right. The tips of the outer hair cell stereocilia are imbedded in a gelatinous mass called the tectorial membrane, which lies on top of the organ of Corti.

When sound is transmitted to the inner ear the organ of Corti begins to vibrate up and down. Since the basilar membrane is attached to bone and ligament at its two ends, the area of maximal vibration is near the third (furthest right) row of outer hair cells. The overlying tectorial membrane is not as flexible so the stereocilia are bent as the organ of Corti moves up and down against it. The electrical potential inside the hair cells changes as the stereocilia are bent. The supporting cells of the organ of Corti send out a narrow filament that angles towards the base of the cochlea, touching the outer hair cells only at their top and bottom. Therefore, for most of the length of these cylindrically shaped cells, they are surrounded by a relatively large fluid filled space. It is known now that the spaces around the outer hair cell allow the cells to change their length during hearing.

The presence of two types of hair cells, the inner and outer hair cells, in the organ of Corti, suggested the idea that they might play different roles in hearing. In fact, it was revealed that most of the nerve fibers that carry information to the brain contact only the inner hair cells. This meant that most the information about the acoustic world reached the brain via the inner hair cells. On the contrary, neural fibers

originating from neurons deep in the brain, which send information back to the hair cells, only touch outer hair cells, thus suggesting the existing of a feedback mechanism.

Outer hair cell stereocilia are firmly embedded in the overlying tectorial membrane while inner hair cell stereocilia make only a tenuous connection. The outer hair cells are located near the center of the basilar membrane where vibrations will be greatest while the basilar membrane is anchored under the inner hair cells (see Fig. 5). These observations suggest that the movement of stereocilia and the resulting modulation of their ionic currents are likely to be greater for outer hair cells than for inner hair cells. Moreover, several studies that had examined the inner ears of deaf people shortly after they died demonstrated that outer hair cells were required for hearing.

What then is the role of outer hair cells, which are over three times more numerous than inner hair cells?

Measured (usually from dead ears) frequency selectivity and the frequency selectivity computed from the engineering analysis of mechanical vibrations of the organ of Corti did not approach the frequency selectivity of the human hearing or the frequency selectivity that could be measured from individual nerve fibers. In other words human *tuning curves* are much more selective than measured ones. More recently, improved measures from living (as opposed to dead) ears revealed that the mechanical frequency selectivity in the living ear began to approach that of human hearing. These considerations suggested that the frequency selectivity of the cochlea could be enhanced if a source of mechanical energy, called "*cochlear amplifier*", were present in the cochlea. This concept appeared validated by the discovery that sound is produced by the inner ear. In fact, "o*toacoustic"* emissions, which can be measured by placing a sensitive microphone in the ear canal, and routinely measured in the clinic to assess hearing. Within five years it was discovered that the outer hair cell could be made to elongate and shorten by electrical stimulation. The function of the outer hair cells in hearing is now perceived as that of a cochlear amplifier that refines the sensitivity and frequency selectivity of the mechanical vibrations of the cochlea. The cylindrical-shape outer hair cells are flexible, in fact they show electromotility, but they are strong enough to transmit force to the rest of the organ of Corti, thus they are pressurized with reinforced membrane along their cylindrical part to prevent them from bursting and to maintain the shape.

The end result of having outer hair cell electromotility is that we are able to discriminate between sounds that are very close in frequency. The reason for this may be easily understood by considering a simple children’s swing. A playground swing is a simple example of mechanical frequency selectivity with a passive system. The swing moves back and forth at a frequency that is determined by the length of the rope and the mass of the child. The parent can make the swing move at a different frequency by exerting considerably more effort than pushing at the natural frequency.

Active tuning is achieved when the child pumps energy into the system. When the pumping is done at the natural frequency and at the correct time in the cycle the swing goes higher and tuning is improved. The improved tuning is best appreciated by attempting to pump at a frequency that is different than the natural frequency. In contrast with the parent pushing, no amount of pumping will make the swing move at

the new frequency, in fact the magnitude of the swinging rapidly decreases. This ability to narrow the range of frequencies at which the swing will oscillate is equivalent to what the outer hair cell does in hearing. In fact, one way to imagine the cochlea is to envision a row of 3000 swings that have progressively longer ropes and heavier children as you move to the right. Each swing has a preferred frequency that is lower than the swing to the left. The parents behind each child are equivalent to sound vibrations and if the children do not pump the entire set of swings is equivalent to a dead cochlea with relatively poor frequency selectivity. If the children pump the set of swings is now equivalent to a living or active cochlea with greatly improved frequency selectivity.

In summary, the role of the outer hair cells in hearing is both sensory and mechanical. When the organ of Corti begins to vibrate in response to the incoming sound, each hair cell will sense the vibration through the bending of its stereocilia. The bending results in a change in the outer hair cell's internal electrical potential, which drives electromotility. If the resulting mechanical force is at the natural frequency of that portion of the cochlea then the magnitude of the vibration will increase. If the electomotile force is at a different frequency, the vibrations will decrease. The system now has greater sensitivity and frequency selectivity than when the outer hair cells are missing or damaged. The refined mechanical vibrations of the organ of Corti are transmitted to the inner hair cells, which excite the 8th nerve fibers at their base and tell the central nervous system that there is energy at a specific frequency in the sound entering the ear.

This frequency map then projects to the brain, which performs the almost unbelievable task of reconstructing the original 3 dimensional acoustic world. The analysis of speech appears to take place in parts of the brain that are highly developed only in man. The amazing machinery that accomplishes the reconstruction of the acoustic world relies on the delicate structures of the inner ear that deconstruct the original sounds.

3 Auditory Speech Processing (ASP): Computational Models and Applications

The auditory system of humans consists, as underlined in the previous Section, of various parts that interact converting the sound pressure waves entering the outer ear into neural stimulus entering the central nervous system. By understanding how these parts act, it is today possible to describe how signals are elaborated by the auditory system, but it is also possible to analyze signals using mathematical models that reproduce the auditory mechanism. In this way we have the possibility to understand which kind of representations our higher levels in the brain use to isolate signals from noise, or to separate signals which have different pitches.

Among all the possible auditory based speech processing techniques, ranging from short-term analyses modified by psychoacoustic or auditory findings [3], [4], [5], [6], [7], to more physiologically based analyses [22], two different models belonging to

the second category will be described in details and few applications will be introduced. in this section.

3.1 The Lyon's Auditory Model: the Correlogram

As all the other physiologically based models, the Lyon's auditory model [33], [34], describes with particular attention the behavior of the cochlea, the most important part of the inner ear, that act substantially as a non-linear filter bank.

The outline of the Lyon's cochlear model, following the work of M. Slaney and R.F. Lyon [35], [36], is described in Fig. 6, where the main characteristic blocks of the model are illustrated.

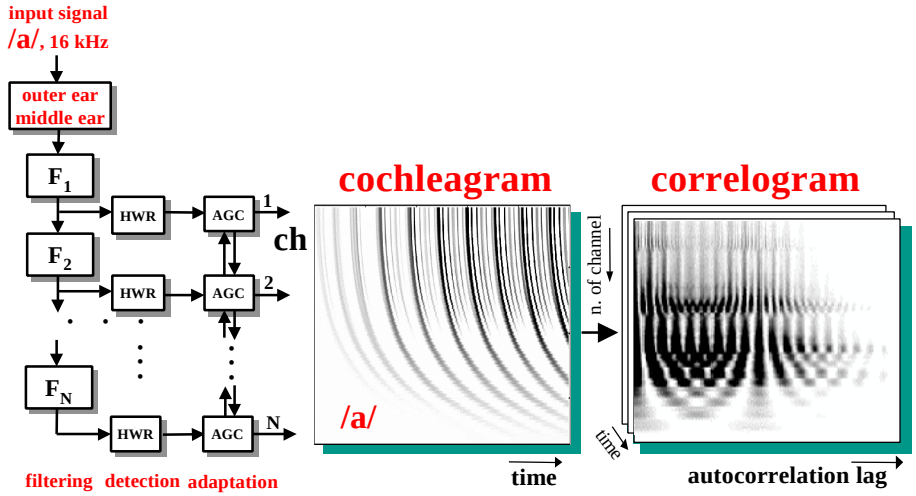

Fig. 6. Outline of the Lyon's cochlear model. The cocheagram and correlogram refer to the Italian vowel /a/.

The first filtering stage models, by a broadly tuned cascade of low-pass filters, the propagation of energy as waves on the basilar membrane (BM). The bigger the number of these filters the more accurate is the model. Usually, before this stage there is another block that simulates the effects of the outer and middle ear (pre-emphasis). In the original implementation 86 filters were considered. This number depends on the sampling rate of the signals and on other parameters of the model such as the overlapping factor of the band of the filters, or the quality factor of the resonant part of the filters.

The second detection stage by the application of a non-linear Half Wave Rectification converts BM velocity into a representation of Inner Hair Cells (IHC) receptor potential or auditory nerve (AN) firing rate. This stage has the function to drop the negative portions of the waveform, modeling the directional behavior of the inner hair cells, thus cutting the energy of the signal by approximately two.

Finally the third compression stage continuously adapts, by an automatic gain control (AGC) mechanism, the operating point of the system in response to its level of

activity. In particular, four automatic gain control stages are cascaded in such a way that the value of the gain of each stage depends on a time constant, on the value of the preceding output sample and on the values of the preceding output samples of the adjacent channels. In this way it is possible to reproduce the well-known psychoacustic masking effect. The different time constants simulate the different adaptive times of our auditory system: the first AGC stage has the biggest time constant so that it reacts to the input signal more slowly, while the following stages have decreasing time constants. This stage compresses widely varying sound input levels into a limited dynamic range of BM motion, IHC receptor potential, and AN firing rate. In summary, the Lyon's cochlear model, instead of trying to replicate the detailed time structured firing events of each IHCs or AN fibers, aims to globally model the neural firing rate as a function of cochlear place or best frequency (BF) or characteristic frequency (CF) versus time. In other words the outputs of the model, the "*cochleagram*", approximately represent the neural firing rates produced by the solicitation of various parts of the cochlea due to the sound pressure waves entering the outer ear. Finally, each of the outputs of the Lyon's cochlear model is short-time windowed, by an L-length window, and autocorrelated thus obtaining as many STAs as the number of channels. The "*correlogram*" consist of the ensemble of these STAs.

For each channel *i*, of the Lyon's cochleagram, the following STA is computed:

$$r_{x_i}(m,\tau)=\frac{1}{L}\sum_{n=0}^{L-|\tau|-1} x_{w_i}(n,m)\,x_{w_i}(n-|\tau|,m) \qquad (1)$$

where: i=1,…,N channel index (N=86 in the present implementation)
L window length in samples
τ autocorrelation lag
and $x_{w_i}(n,m) = x_i(n)\cdot w(m-n)$

Given that the autocorrelation itself is a function of a third variable, the autocorrelation lag, the resulting correlogram is a three-dimensional function of frequency or number of channel of the cochlear filter bank, time, and autocorrelation lag or delay.

3.2 Pitch Extraction and Speech Enhancement

The correlogram is not a static representation but a dynamic one. Given that the periodicity information remains unchanged along all the transformations occurring in the Lyon's cochlear channels, if a sound is periodic, the autocorrelation functions of a big number channels shows a clear peak at the position t_p corresponding to the period of the input signal. As stated in the original implementation of the algorithm [37], if a sum is executed with the values of the correlogram *cor(t,i)* along all the channels, normalized by the correlation of the window used to reduce the classical frame border effect, a function called "*summary correlogram*" *SC(t)* is obtained:

$$SC(\tau) = \sum_{i=1}^{N} \frac{cor(\tau, i)}{cor_w(\tau)} \quad \textbf{(2)}$$

where the peak corresponding to the period of the input signal, related to the pitch frequency by the relation $f_p = 1/t_p$ results enhanced. An outline of the whole pitch extraction procedure together with an example of its application to a simple vowel stimulus is illustrated in Fig. 7. When the signal is highly degraded by noise the above procedure has to be modified in order to become more robust. In the case of a stationary noise, a sort of spectral-subtraction technique [38] in the correlogram domain named "*correlogram subtraction*" is proposed [39]. In fact, a mean correlogram of the noise, is subtracted from the correlogram of the signal plus the noise. The main hypothesis for this procedure to be applicable is that there must be some parts of the signal where only the noise is present; in this way its statistics can be computed. This is obviously a quite common situation in the case of the speech signal where silence is frequently inter-mixed within speech. The above procedure is shown to result quite effective despite the fact that the non-linearity transformations implied in the auditory front-end should not justify the super-imposition effect which is instead explicitly accepted considering the subtraction operation.

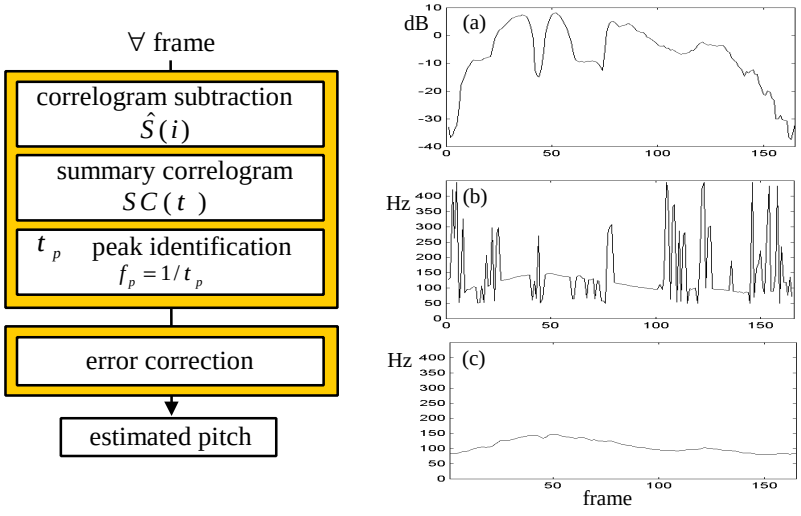

Fig. 7. Outline of the pitch extraction algorithm [39] on the left and an example of its application on the right. Referring to the the Italian noisy voiced word /lavan'daja/ ('washerwoman'): (a) the frame-by-frame signal to noise ratio corresponding to a global 0dB SNR; (b) the pitch estimate from the original Summary Correlogram [37] is given; (c) the final estimated pitch computed with the proposed algorithm.

The same ‘correlogram subtraction’ technique is then applied to enhance noisy speech signal [39]. In fact, before applying the inversion strategy procedure described in [40], [41], a clean correlogram is reconstructed by the above technique and a much more clean speech is synthesized, as illustrated in the following example shown in Fig. 8.

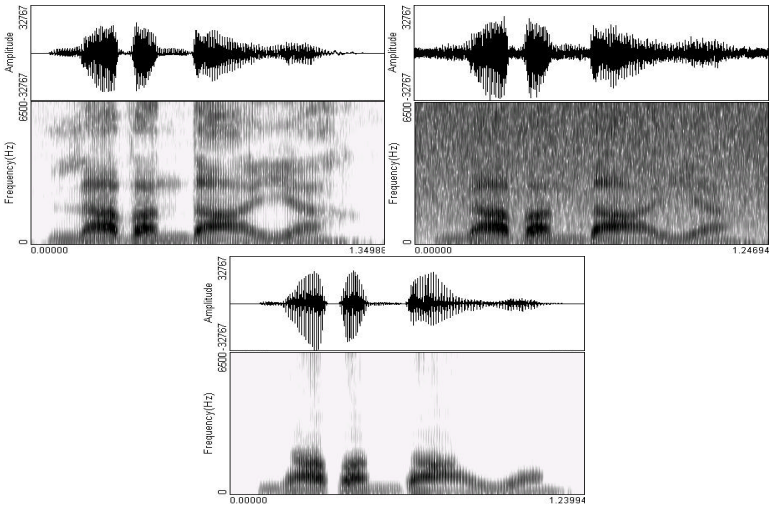

Fig. 8. Application of the '*correlogram subtraction*' enhancement technique to the Italian noisy (0dB SNR) word /lavanda'ja/ ('washerwoman'): (a) clean; (b) noisy; (c) enhanced.

Very promising results have been so far obtained, both for pitch extraction and for the enhancement procedure for noisy speech, thus leading to hypothesize the future use of these auditory-based techniques in new front-ends for robust speech applications.

3.2 The Seneff *Joint Synchrony/Mean-Rate*' Auditory Speech Processing

The computational scheme proposed by S. Seneff [23], [42] to model the human auditory system is called *Joint Synchrony/Mean-Rate* model and similarly to the Lyon's model tries to capture the essential features extracted by the cochlea in response to sound pressure waves. The overall system, described in the block diagram in Fig. 9, includes three blocks. The first two of them deal with peripheral transformations occurring in the early stages of the hearing process while the third one attempts to extract information relevant to perception such as *formants* and to enhance sharpness of onset and offset of different speech segments.

The speech signal, band-limited and sampled at 16 kHz, is first pre-filtered through a set of four complex zero pairs to eliminate the very high and very low frequency components. Then it passes through the first block, a 40-channel critical-band linear filter bank whose single channels were designed, similarly to Lyon's model, in order to fit physiological data.

The second block is called the hair cell synapse model. It is nonlinear and is intended to capture prominent features of the transformation from basilar membrane vibration, represented by the outputs of the filter bank, to probabilistic response properties of auditory nerve fibers. The outputs of this stage represent the probability of firing as a function of time for a set of similar fibers acting as a group.

The third and last block is a double-unit block with two parallel outputs. The *Generalized Synchrony Detector* (*GSD*), which implements the known "*phase-locking*" property of nerve fibers, represents the first unit and is designed with the aim of enhancing spectral peaks due to vocal tract resonances. The second unit, called *Envelope Detector* (*ED*) computes the envelope of the signals at the output of the previous stage of the model and seems more important for capturing the very rapidly changing dynamic nature of speech. The outputs of this unit should be more important in characterizing transient sounds.

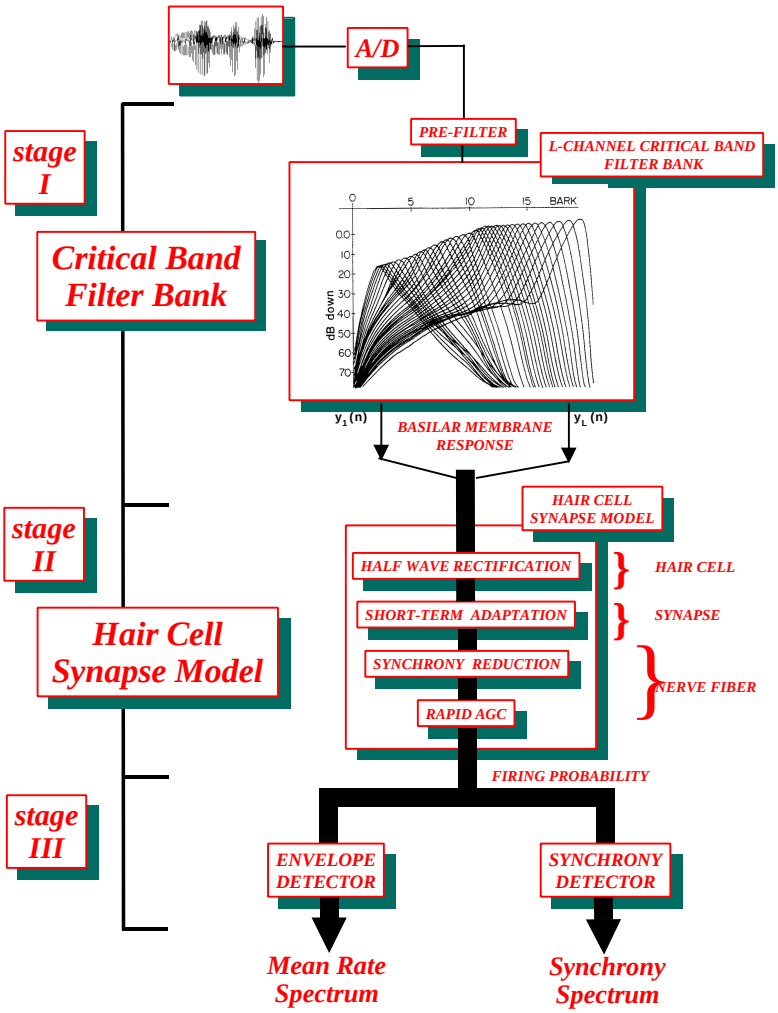

Fig. 9. Block diagram of the Seneff '*Joint Synchrony/Mean-Rate Auditory Speech Processing*' (for a complete description of the model refer to [23]).

Referring to the Italian word /pa'tata/ (potato), shown in Fig. 10, uttered in isolation by an Italian male speaker, the output of the Seneff model is quite effective

in producing GSD spectra with a limited number of well defined spectral lines[2] and also in tracking the dynamic modifications of speech. Transitions from one phonetic segment to the next are clearly delineated by onsets and offsets in the output representation better represented by the ED auditory spectrogram, and this is probably due to the forward masking mechanism which is directly included in the model.

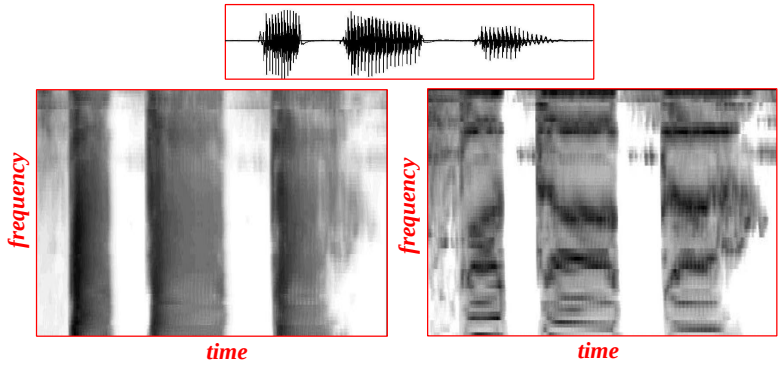

Fig. 10. Output of the Seneff auditory speech processing applied to the Italian words /pa'tata/ (potato). Envelope and Synchrony parameters are shown on the left and right respectively.

3.3 Segmentation

Considering Seneff words in [23]:

> "…..The output of this model is delivered to two parallel channels, each of which produces spectral representations appropriate for distinct subtasks of a speech recognition system. One path yields an overall energy measure for each channel that can be identified with the average rate on neural discharge. The outputs of this path appear to be useful for locating acoustic events and assigning segments to broad phonetic categories. In the other path, the extent of dominance of periodicities at each channel's center frequency is captured by a synchrony measure, which yields a spectral representation with enhanced spectral contrast, relative to the mean-rate spectrogram. The outputs of this stage show distinct formant peaks during sonorant regions, with smooth transitions over time, as well as preserving spectral prominences in the high-frequency region for fricatives and stops….."

it is quite evident that this processing scheme constitutes a good starting point to build an effective segmentation system.

In fact, following the work of Glass [43] on the so-called Multi-Level Segmentation[3] (MLS) theory [44] a PC-based segmentation system called SLAM

---

[2] This represents a good use of speech knowledge according to which formants are voiced sound parameters with low variance.

[3] Within the framework of MLS theory [44], speech is considered as a temporal sequence of quasi-stationary acoustic segments, and the points within such segments are more similar to each other than to the points in adjacent segments. Following this viewpoint, the segmentation problem can be simply reduced to a local clustering problem where the decision to be taken regards the similarity of any particular frame with the signal immediately preceding or following it. Using only relative measures of acoustic similarity, this technique should be quite independent of the speaker, vocabulary, and background noise.

(Semi Automatic Segmentation Module)[4] was designed and implemented [45]. SLAM makes use of the MLS hierarchical technique, that, incorporating some kind of temporal constraints, is quite useful to appropriately rank the significance of acoustic events. By a recursive technique, involving the computation of Euclidean-based similarity measure for each target frame, some initial adjacent "seed regions", which constitute the basis for the MLS "hierarchical structuring" segmentation procedure, are created. These regions, using the same similarity measure, are themselves merged together and, by keeping track of the distance at which two regions merge into one, a multi-level structure is built up that describes the hypothesized segmentation landmarks, usually called *dendogram* [44], is built up. The effectiveness of the combination of the MLS strategy with the Seneff auditory processing versus other more classical frame-based analysis techniques such as FFT and LPC could be verified in Fig. 11 where two segmentation examples referring to two Italian syllables /'ba/ and /'ka/ are illustrated

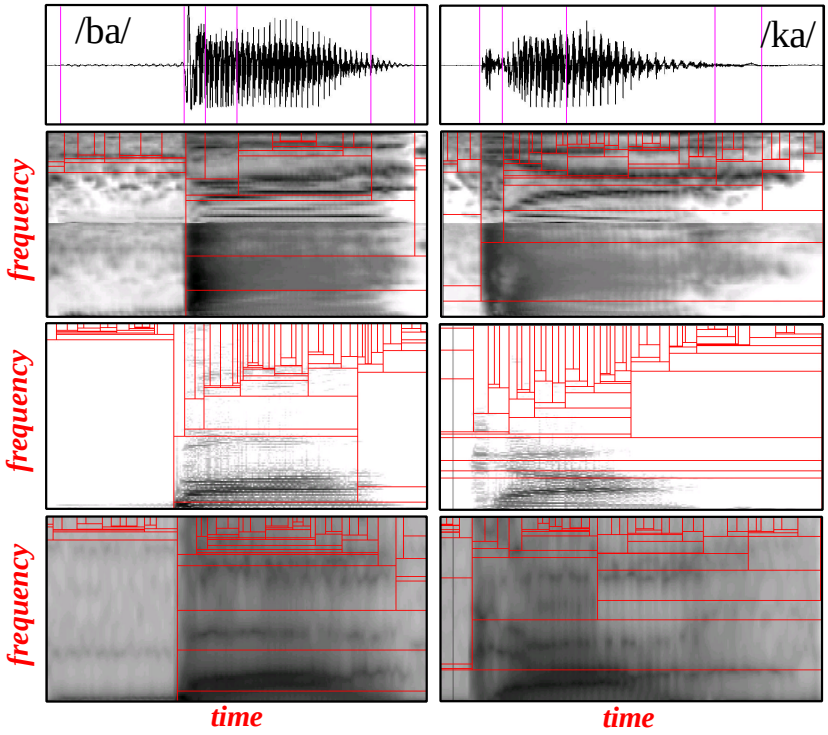

Fig. 11. Example of segmentation of the Italian syllables /'ba/ and /'ka/ with different auditory and spectral representations. For each syllable, the Seneff AM, the narrow-band FFT, and the LPC-derived spectrograms are illustrated below the waveform from top to bottom.

Looking at the dendograms with the hypothesized segmentation landmarks superimposed to every spectrogram plots in the figure, it is quite evident that with the

[4] SLAM version 1.0 works on Windows 3.1, 3.11, 95 and NT and is available at the following ftp address: www.csrf.pd.cnr.it (/Webdisk/Pub/Cosi/Slam).

Seneff ASP the final correct segmentation is much more easily identified. Using this procedure, along the lines followed for segmenting and labeling American English speech material [24], [45], various Italian speech data have been semi-automatically segmented and labeled with SLAM obtaining an accuracy similar to that obtained by manual labeling by expert phoneticians [46].

4 Artificial Neural Networks (ANNs) and Auditory Speech Processing (ASP) for Automatic Speech Recognition

As underlined in the previous Sections, various studies suggest the effectiveness of auditory-based front-ends in different speech processing applications, especially in adverse conditions.

Moreover, experimental results show that Artificial Neural Networks (ANNs) represent an effective alternative to classical pattern recognition methods in various applications [47]. In fact, several neural network models have been recently investigated by researchers for dealing with signal processing and particularly with automatic speech recognition [48]. The combination of ASP with ANN techniques give rise to new tools resulted quite effective to tackle the problem of continuous speech recognition, especially at the '*very low*' phonetic level.

4.1 Multi-Layered Neural Networks (MLNNs)

The Multi-Layered Neural Networks (MLNNs) trained with Back-Propagation (BP) are probably the most used as static networks [47].

MLNNs are networks with an input layer of nodes, one or more hidden layers and an output layer whose nodes represent a coded version of the input. As illustrated in Fig. 12, nodes are connected by links and weights are associated to links. All the links bringing a signal into a node contribute to the calculation of the excitation of that node. The excitation is the sum of the product of the weights of each link and the value of the output coming from the node the link carries its signal from. The output of a node is a function of the node excitation.

By choosing the link weights a large variety of classifiers can be designed having specific properties. Link weights can be obtained by a learning process. Learning can be supervised or unsupervised. When learning is supervised, the network input is fed by sets of patterns. Each set corresponds to a class of patterns that have to be coded with the same values appearing at the output nodes. The output nodes are clamped with the desired values and algorithms exist for computing the values of the link weights in such a way that the network codes the sets of input patterns as desired. These learning algorithms have a relevant generalization capability. Recently, a large number of scientists are investigating and applying learning systems based on MLNNs. Definitions of MLNNs, motivations and algorithms for their use can be found in [48]. Theoretical results have shown that MLNNs can perform a variety of complex functions [49]. Applications have also shown that MLNNs have interesting generalization performances capable of capturing information related to pattern

structures as well as characterization of parameter variation [50], [51]. Algorithms exist for MLNNs with proven mathematical properties that allow learning to be competitive and to focus on the properties that make different patterns belonging to different classes. Furthermore, in MLNNs the knowledge about a set of competing classes (in our case speech units or phonemes) is distributed in the weights associated with the links among nodes. If we interpret each output of the classifier as representing a phonetic property, then an output value can be seen as the degree of evidence with which that property has been observed in the data.

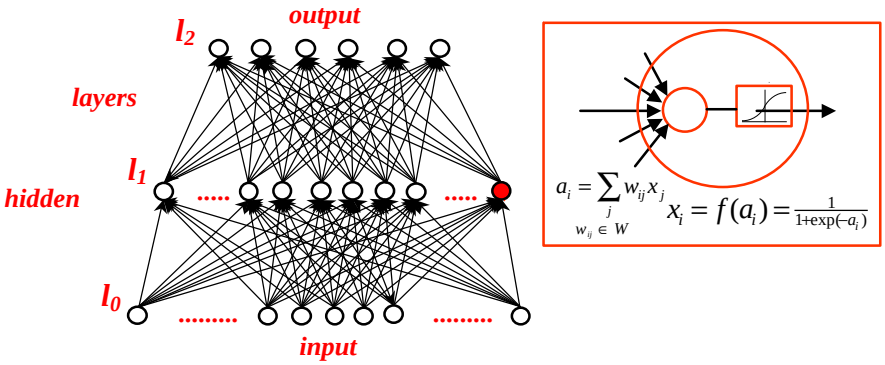

Fig. 12. Outline of a simple Artificial Neural Network (ANN) architecture. A single node is magnified on the right and its computing function is indicated.

4.2 Speaker Independent (SI) Vowel Recognition

The combination of the Seneff ASP and MLNNs give rise to an effective generalization among speakers in coding vowels.

In fact, the results of the application of Seneff ASP in vowel classification experiments were compared with those obtained with a classical front-end built with an FFT-based filter-bank. A subset of the American-English vowels built up with the 10 vowels /i, ɪ, ɛ, æ, ʌ, ə, ɑ, ɔ, ʊ, u/ were extracted from the words BEEP, PIT, BED, BAT, BUT, FUR, FAR, SAW, PUT, BOOT. With a MLNN-based system, a 96% correct classification performance was obtained with the ASP front-end while with a classical FFT-based front-end an 87% correct classification performance was achieved [52]. Similar results have been obtained for the complete set of Italian vowels /i, e, ɛ, a, ɔ, o, u/, which have been extracted from the words PIPA, PEPE, PEPPA, PAPA, POPE, POPPA, PUPA [53].

4.3 Dynamic Multi-Layered Neural Networks (DMLNNs)

A dynamical behavior has been differently added to static MLNNs thanks to various techniques. Among them:

(a) the transformation of recurrent networks in feedforward ones [49];

(b) the introduction of feedback connections [54];
(b) the addition of buffered context at the input [55];
(c) adding buffered context at the input and at the hidden layers [56].

Especially the last approach has given good results in speech recognition experiments [57]. All these new models are inherently limited in their representation of the past to a fixed period of time. On the contrary, the dynamic architectures called Dynamic Multi-Layered Networks (DMLNs) are characterized by the execution of the supervision without considering a static input [58]. In fact, instead of waiting for a fixed point in time a learning algorithm known as Extended Back Propagation for Sequences, can be utilized. Within this framework, the learning environment is defined by a sequence of frames representing the natural time evolution of speech signals and the output supervision is done during the evolution of the activations. The dynamic model usually considered is discrete instead of continuous and its transitions occur when a new frame is applied at the input. A very simple and usually adopted class of DMLNs is that in which dynamic neurons, with feedback connections to themselves, have only incoming connections from the input layer.

4.4 I-set and E-set Speaker Indipendent Recognition

By the use of ASP and DMLNNs various successful applications have been designed. In particular, in Fig. 13, the two architectures utilized for the classification of the Italian alphabet I-set and E-set[5] [60] are simultaneously illustrated.

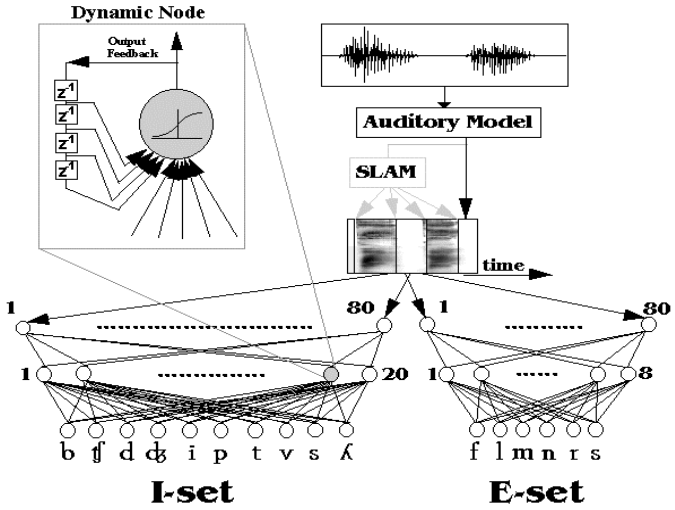

Fig. 13. Outline of the architectures used for the I-set and the E-set classification

[5] Italian I-set: /bi/, /tSi/, /di/, /dZi/, /i/, /pi/, /ti/, /vi/ plus other two "out of alphabet" stimuli /Li/, /si/. Italian E-set: /'Effe/, /'Elle/, /'Emme/, /'Enne/, /'Erre/, /'Esse/ (see SAMPA Phonetic Alphabet [59]).

As illustrated in the figure, Input speech signal, sampled at 16kHz in a quiet office room, it is analyzed with the Seneff ASP, and successively segmented using the SLAM segmentation program in order to locate onset and offset of target stimuli. Finally, two DMLNNs with dynamic neurons with local feedback connections to themselves, having only incoming connections from the input layer, are considered as the recognition framework. Frame rate was set to 2ms for the I-set experiment and to 8ms for the E-set experiment while the delay value of dynamic neuron was set to 4 frames for both experiments. Learning supervision time was forced only at the last frame of the target stimuli. Speech data-base is made up of 7 male talkers repeating five times, in random order, each of the selected non-sense stimuli, and circularly one speaker is tested while using the remaining 6 speakers for learning. The achieved Speaker Independent (SI) mean recognition error-rate is around 35% and 12% for the I-set and the E-set respectively.

4.4 Bi-Modal Recognition

Considering the fact that humans make use of various sources of information, especially visual, in order to recognize and understand speech with high accuracy, the idea of building new automatic speech recognizers able to use other sources of information than the acoustic signal such as those given by our visual channel is becoming more and more attractive within the scientific community [61].

Various audio-visual automatic speech recognition (ASR) systems able of enhancing recognition performance, mostly in noisy conditions, have been developed in the past. Among them, the system described in Fig 14 refers to the specific architecture utilized in an Italian speaker-dependent [62] and speaker-independent [63] plosive-set classification experiment.

The system being described makes use of a system for automatic jaw and lips movement 3D analysis called ELITE. The speech signal is acquired in synchrony with the articulatory data extracted by ELITE and is pre-filtered and sampled at 16 kHz. The Seneff ASP is applied to the acoustic input channel. The final vector made of 40 acoustic and 14 articulatory parameters, tracking the movements of few reflecting-paper markers positioned on the lip of the subjects, is sent, at a 500 Hz frame rate, to a DMLNN trained by EBPS to discriminate among input stimuli. Learning supervision times are forced only at the middle and last frames of the target stimuli. Speech data-base referring to disyllabic symmetric /'VCV/ nonsense words, where C=/p,t,k,b,d,g/ and V=/a,i,u/, is made up of 10 male talkers repeating five times, in random order, each of the selected non-sense stimuli. As for the I-set and E-set experiments described in the previous section, circularly one speaker is tested while using the remaining 9 speakers for learning. The achieved speaker independent mean recognition error-rate is around 29% (23% if place of articulation classes are considered), which is rather satisfactory considering the quite difficult task of classifying plosive consonants using only two supervision points. A quite better performance (6% error-rate) were obtained in a speaker-dependent classification experiment where the same architecture resulted extremely robust even in a noisy 0dB signal-to-noise ratio condition. Moreover, both in the SD and in the SI experiment,

the simultaneous use of acoustic and articulatory parameters always improved the recognition performance with respect to the use of the acoustic parameters alone.

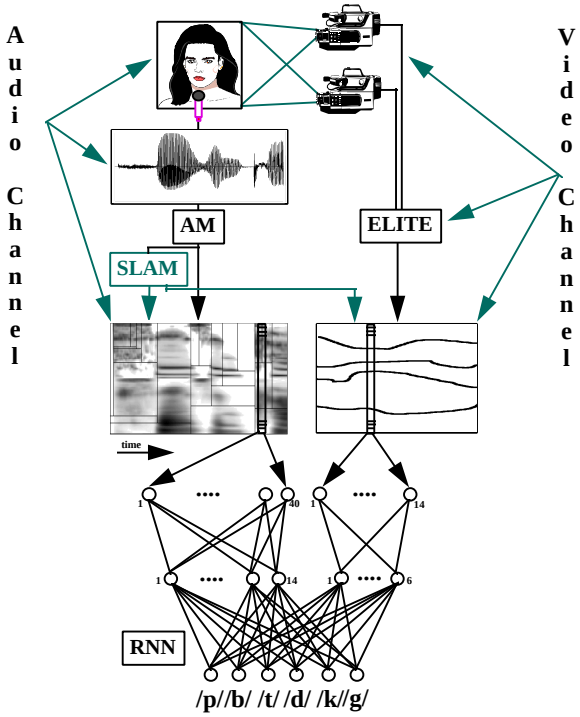

Fig 14. Block diagram of the bimodal recognition system utilized in the plosive classification task.

5. Conclusions

Various evidences suggest the effectiveness of the application of auditory speech processing techniques for speech analysis.

As for segmentation, considering both gross-errors (over-segmentation) and fine-errors (segmentation discrepancies) ASP parameters seem to constitute a very effective tool and a better alternative to other classical frame-based analysis parameters. The same applies to pitch extraction and noise enhancement.

As for recognition, much more experiments need to be performed to convince ASR people to adopt this kind of processing, but an increased interest on this matter seems to be already activated by preliminary results on comparing these new techniques against classical ones especially in noisy conditions.

Moreover, the well-known objection to auditory-based front-ends, i.e. the too high computational cost of such processing techniques, will be easily overcome in the future, due to the tremendous increase in speed and capacity characteristics, of new designed computer processing chips.

Despite the advances made in the last years, auditory modeling is still a young research field. Knowledge on human auditory functioning has been acquired during

the past and surely more discovering will be made in the future. When it will be possible to easily incorporate all this knowledge into effective real-time digital speech processing algorithms the automatic recognition of speaker independent fluent speech could become a reality.

7 Acknowledgments

This work has been made possible exclusively thanks to:
William E. Brownell, for the first section;
(brownell@bcm.tmc.edu, Ph.D. at the Department of Otorhinolaryngology and Communicative Sciences, Baylor College of Medicine, Houston, Texas, TX 77030.S.)
Malcom Slaney, for everything regarding the Lyon's model and its applications;
(malcolm@interval.com, Interval Research Corporation, 1801 Page Mill Rd. B.C, Palo Alto, CA 94304)
Stephanie Seneff and James .R. Glass, for the suggestions regarding the implementation of the joint Synchrony/Mean-Rate (S/M-R) model of Auditory Speech Processing, and the development of the MLS segmentation algorithm.
(seneff@goldilocks.lcs.mit.edu, glass@mit.edu,545 Tech. Square, NE43-643 Cambridge, MA 02139) .

References

1. Rabiner L.R. and Shafer R.W.: Digital Processing of Speech Signals. Prentice-Hall, Inc., Englewood Cliffs, New Jersey (1978).
2. Markel J.D. and Gray A.H., Jr.: Linear Prediction of Speech. Springer-Verlag, Berlin, Heidelberg, New York (1976).
3. Strube H.W.: Linear prediction on a warped frequency scale. JASA Vol. 68(4), Oct. (1980) 1071-1076.
4. Blomberg M., Carlson R., Elenius K. and Granstrom B.: Auditory Models and Isolated Word Recognition. STL-QPSR Vol. 4 (1983) 1-15.
5. Hermansky H., Hanson B.A. and Wakita H.: Perceptually Based Linear Predictive Analysis of Speech. Proc. IEEE ICASSP (1984) 509-512.
6. Hermansky H. and Junqua J.C.: Optimization of Perceptually Based ASR Front-Ends. Proc. IEEE ICASSP (1988) 219-222.
7. Davis S.B. and Mermelstein P.: Comparison of Parametric Representation of Monosyllabic Word Recognition in Continuously Spoken Sentences. IEEE Trans. ASSP Vol. 28(4) (1980) 357-366.
8. Furui S.: Speaker Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. IEEE Trans. on ASSP Vol. 34(1) (1986) 52-59.
9. Rabiner L.R., Wilpon J.G. and Soong F.K.: High Performance Connected Digit Recognition Using Hidden Markov Models. Proc. IEEE ICASSP (1988) 119-122.
10. Lee K.F.: Automatic Speech Recognition; The Development of the SPHINX System. Kluwer Academic Publisher, Boston (1989).
11. Rioul O. and Vetterli M.: Wavelets and Signal Processing. IEEE Signal Processing Magazine, October (1991) 14-38.
12. Delgutte B.: Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. JASA Vol. 68 (1980) 843-857.
13. Delgutte B. and Kiang N.Y.S.: Speech coding in the auditory nerve: I. Vowel-like sounds. JASA Vol. 75 (1984) 866-878.

14. Delgutte B. and Kiang N.Y.S.: Speech coding in the auditory nerve: II. Processing Schemes for Vowel-like sounds. JASA Vol. 75 (1984) 897-907.
15. Delgutte B. and Kiang N.Y.S.: Speech coding in the auditory nerve: III. Voiceless fricative consonants. JASA Vol. 75 (1984) 887-896.
16. Delgutte B. and Kiang N.Y.S.: Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics. JASA Vol. 75 (1984) 897-907.
17. Young E.D. and Sachs M.B.: Representation of steady-state vowels in the temporal aspects of the discharge pattern of populations of auditory nerve fibers. JASA Vol. 66 (1979) 1381-1403.
18. Sachs M.B. and Young E.D.: Effects of nonlinearities on speech encoding in the auditory nerve. JASA Vol. 68 (1980) 858-875.
19. Miller M. I. and Sachs M. B.: Representation of stop consonants in the discharge patterns of auditory-nerve fibers. JASA Vol. 74 (1983) 502-517.
20. Sinex D.G. and Geisler C.D.: Responses of auditory-nerve fibers to consonant-vowel syllables. JASA Vol. 73 (1983) 602-615.
21. Kiang N.Y.S., Watanabe T., Thomas E. C. and Clark L. F.: Discharge patterns of single fibers in the cat's auditory-nerve fibers. Cambridge, MA, MIT press (1965).
22. Greenberg S. (ed.): Representation of Speech in the Auditory Periphery. Journal of Phonetics, Special Issue, Vo. 16(1), January (1988).
23. Seneff S.: A joint synchrony/mean-rate model of auditory speech processing. Journal of Phonetics, Special Issue, Vol. 16(1), January (1988) 55-76.
24. Zue V.W., Glass J., Philips M. and Seneff S.: Acoustic Segmentation and Phonetic Classification in the SUMMIT System. Proc. IEEE ICASSP (1989) 389-392.
25. Cosi P., Bengio Y. and De Mori R.: Phonetically-Based Multi-Layered Neural Networks for Vowel Classification. Speech Comm., Vol. 9, N. 1, Feb (1990) 15-29.
26. Cosi P., Frasconi P., Gori M. and Griggio N.: Phonetic Recognition Experiments with Recurrent Neural Networks. Proc. ICSLP (1992) 1335-1338.
27. Cosi P.: Auditory modelling for speech analysis and recognition. In: M. Cooke, S. Beet, M.Crawford (eds.): Visual representation of speech signals. Wiley & Sons Chichester (1993) 205-212.
28. Hunt M.J. and Lefebvre C.: Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model. Proc. IEEE ICASSP (1988) 215-218.
29. Jankowski C.R. Jr., Vo H-D. H. and Lippmann R.P.: A Comparison of Signal Processing Front Ends for Automatic Word Recognition. IEEE Trans Speech and Audio Processing, Volume SAP-3, N. 4, Jul (1995) 286-293.
30. Brownell W.E.: How the Ear Works - Natures Solutions For Listening. Volta Review, in press (1998). See also electronic publication: http://www.bcm.tmc.edu/oto/research/cochlea/Volta/index.html.
31. Pickles J. O.: An Introduction to the Physiology of Hearing. Academic Press, (1988).
32. Dallos P. , Popper A.N., Fay R.R. (eds.): The Cochlea. From the Springer Handbook of Auditory Research. Springer, New York (1996)
33. Lyon R. F.: A Computational Model of Filtering, Detection, and Compression in the Cochlea. Proc IEEE-ICASSP (1982) 1282-1285.
34. Lyon R. F.: Computational Models of Neural Auditory Processing. Proc. IEEE-ICASSP (1984) 36.1.1-36.1.4.
35. Slaney M.: Lyon's Cochlear Model. Tech. Rep. # 13, Apple Inc., Cupertino, Ca. (1988).
36. M. Slaney and R.F. Lyon, On the importance of time – a temporal representation of sound, in Visual Representation of Speech Signals, M. Cooke, S. Beet and M. Crawford (eds.), John Wiley & Sons Ltd, 1993, pp. 95-116.
37. Slaney M. and Lyon R.F.: A Perceptual Pich Detector. Proc. IEEE-ICASSP (1990)357-360.
38. Boll S.F.: Suppression of Acoustic Noise in Speech Using Spectral Subtraction. Trans. ASSP Vol. 27 (1979).

39. Cosi P., Pasquin S. and Zovato E.: Auditory Modeling Techniques for Robust Pitch Extraction and Noise Reduction. Proc. of ICSLP (in press) (1998) paper 1053.
40. Slaney M., Naar D. and Lyon R.F.: Auditory Model Inversion for Sound Separation. Proc. IEEE ICASSP (1994) II. 77-80.
41. Cosi P. and Zovato E.: Lyon's Auditory Model Inversion: a Tool for Sound Separation and Speech Enhancement. Proc. of ESCA Workshop on 'The Auditory Basis of Speech Perceprion', Keele University, Keele (UK), 15-19 July (1996) 194-197.
42. Seneff S.: A computational model for the peripheral auditory system: application to speech recognition research. Proc. IEEE ICASSP (1986) 37.8.1-37.8.4.
43. Glass J.R.: Finding Acoustic Regularities in Speech: Application to Phonetic Recognition. Ph. D Thesis, May, MIT press (1988).
44. Glass J.R. and Zue V.W.: Multi-Level Acoustic Segmentation of Continuous Speech. Proc. IEEE ICASSP (1988) 429-432.
45. Seneff S. and Zue V.W.: Transcription and Alignment of the TIMIT Database. Unpublished manuscript to be distributed with the TIMIT database by NBS (1988).
46. Cosi P.: La Segmentazione delle Occlusive dell' Italiano mediante SLAM. Proc. XXVI Convegno A.I.A, Torino, 27-29 Maggio (1998).
47. Rumelhart D.E., Hinton G.E. and Williams R.J.: Learning Internal Representation by Error Propagation. In: Rumelhart D.E., Hinton G.E. and Williams R.J: Parallel Distributed Processing: Exploration in the Microstructure of Cognition. MIT Press (1986) 318-362.
48. Bourlard H. A. and Morgan N.: Connectionist Speech Recognition. A Hybrid Approach. Kluwer Academic Publishers (1994).
49. Plaut D.C. and Hinton G.E.: Learning Sets of Filters Using Back Propagation. Computer Speech and Language. Vol. 2 (1987) 35-61.
50. Bourlard H. and Wellekens C.J.: Multilayer Perceptron and Automatic Speech Recognition. Proc. IEEE ICNN (1987) IV-407-416.
51. Watrous R.L. and Shastri L.: Learning Phonetic Features Using Connectionist Networks. Proc. IJCAI (1987) 851-854.
52. Cosi P., Bengio Y. and De Mori R.: Phonetically-Based Multi-Layered Neural Networks for Vowel Classification. Speech Communication, Vol. 9, No. 2, (1990) 15-29.
53. Cosi P., De Mori R. and Vagges K.: A Neural Network Architecture for Italian Vowel Recognition. Proc. VERBA-90, Rome, 22-24 January (1990) 221-228.
54. Pineda F.J.: Generalization of Back-Propagation to Recurrent Neural Networks. Physical Review Letters, Vol. 59, n. 19, November (1987) 2229-2232.
55. Sejnowsky T.J. and Rosemberg C.R.: NETTalk: a Parallel Network that Learns to Read Aloud. Technical Report JHU/EECS-86/01 (1986).
56. Waibel A., Hanazawa T., Hinton G.E., Shikano K. and Lang K.: Phoneme Recognition Using Time-Delayed Neural Networks. A.T.R. Tech. Report TR-I-0006, October (1987).
57. Waibel A.: Modularity in Neural Networks for Speech Recognition. Proc. NIPS (1988).
58. Gori M. , Bengio Y. and De Mori R.: BPS: A Learning Algorithm for Capturing the Dynamical Nature of Speech. Proc. IEEE-IJCNN (1989) II-417-432.
59. Fourcin A.J., Harland G., Barry W. and Hazan W. (eds.): Speech Input and Output Assessment, Multilingual Methods and Standards. Ellis Horwood Books (1989).
60. Cosi P., Mian G.A. and Contolini M.: Speaker Independent Phonetic Recognition Using Auditory Modelling and Recurrent Neural Networks. Proc. ICANN (1994) 925-928.
61. Storke D.G. and HennekeM.E. (eds.): Speechreading by Humans and Machine: Models, Systems and Applications. NATO ASI Series, Series F: Computer and Systems Sciences, Vol. 150, Springer-Verlag (1996).
62. Cosi P., Magno Caldognetto E., Vagges K., Mian G.A. and Contolini M.: Bimodal Recognition Experiments with Recurrent Neural Networks. Proc. IEEE ICASSP (1994)
63. Cosi P., Magno Caldognetto E., Ferrero F.E., Dugatto M. and Vagges K.: Speaker Independent Bimodal Phonetic Recognition Experiments. Proc. ICSLP (1986) 54-57.