Neurocomputing 72 (2009) 820–830

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Time-frequency representations in speech perception

Pedro Gómez-Vilda [a,*], José M. Ferrández-Vicente [b], Victoria Rodellar-Biarge [a], Roberto Fernández-Baíllo [a]

[a] Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, 28660 Boadilla del Monte, Madrid, Spain
[b] Universidad Politécnica de Cartagena, Campus Universitario Muralla del Mar, Pza. Hospital 1, 30202 Cartagena, Spain

ARTICLE INFO

Available online 5 November 2008

Keywords:
Bio-inspired speech processing
Speech perception
Acoustic-phonetics
Phonetic boundaries and classes
Minimal semantic units

ABSTRACT

Nowadays applications demand a comprehensive view of voice and speech perception to build more complex and competitive procedures amenable of extracting as much knowledge from sound-based human communication as possible. Many knowledge-extraction tasks from speech and voice may share signal treatment procedures which can be devised under the point of view of bio-inspiration. The present paper examines a hierarchy of sound processing functionalities at the auditory and perceptual levels on the Auditory Neural pathways which can be translated into bio-inspired speech-processing techniques, their fundamental characteristics being analyzed in relation with current tendencies in cognitive audio processing. The pathways linking the peripheral auditory system (cochlear complex) with the brain cortex are briefly examined, with special attention to the study of neuronal structures showing specific capabilities under the point of view of formant analysis and the build-up of a semantic hierarchy from the time-frequency structure of speech to explore their capability of conveying semantics to speech processing and understanding from the minimal acoustic clues with elementary meaning or "sematoms". The replication of known biological functionality by algorithmic methods through bio-inspiration is a secondary aim of the research. Examples extracted from speech processing tasks in the domain of acoustic-phonetics are presented. These may find applicability in speech recognition, speaker's characterization and biometry, emotion detection, and others related.

1. Introduction

Bio-inspired speech processing is the treatment of speech following paradigms used by the human sound perception system, which is known to posses specific resources for this purpose. The question if bio-inspiration is a convenient strategy for devising specific tasks in speech processing has been a subject of great controversy, and it remains still open [1,2]. The generalized impression is that bio-inspiration may offer alternative ways to implement specific functions in speech processing, helping to improve the performance of conventional methods albeit at the cost of assuming bottom-up solutions, which most of the times are cumbersome to implement and do not grant better performance. Nevertheless the enormous gap still existing between natural and artificial systems in speech processing supports the belief that bio-inspiration is a rich approach still offering potential improvements. The complexity of human language processing is clearly stated by Cytowic: "… language turned out to be far more complex than the grammar found in the textbooks. And yet it is routinely surpassed by what a six-year-old has in her head …" [3].

At this point the present approach requires a semantic clarification. Under the term *language* a rich field of knowledge is hidden, of which this work will be mostly interested in acoustic-phonetics related with the productive, perceptual, neural processing and bio-inspired algorithmic design levels. Of course, due to the extension and complexity of the fields covered, the treatment will be lighter than desired, although the reader may find more detailed information in the references cited.

Many tasks in speech processing and understanding remain unsolved yet, which may benefit from using algorithmic structures mimicking the functionality shown by neuronal structures present in the auditory pathways between the peripheral auditory system and the auditory cortex. Based on this belief the approach to bio-inspired speech processing described in the present paper is disclosed as follows:

- A review on how speech is produced is covered in Section 2, to understand its physical and acoustical nature, with special emphasis in the speech production source-filter model.
- The review of speech perception, as described in Section 3 gives a general idea on how the time-frequency distribution of speech is treated by the human auditory system as a hierarchy of semantic categories or levels extracted from the dynamics of

* Corresponding author.
E-mail address: pedro@pino.datsi.fi.upm.es (P. Gómez-Vilda).

0925-2312/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2008.04.056

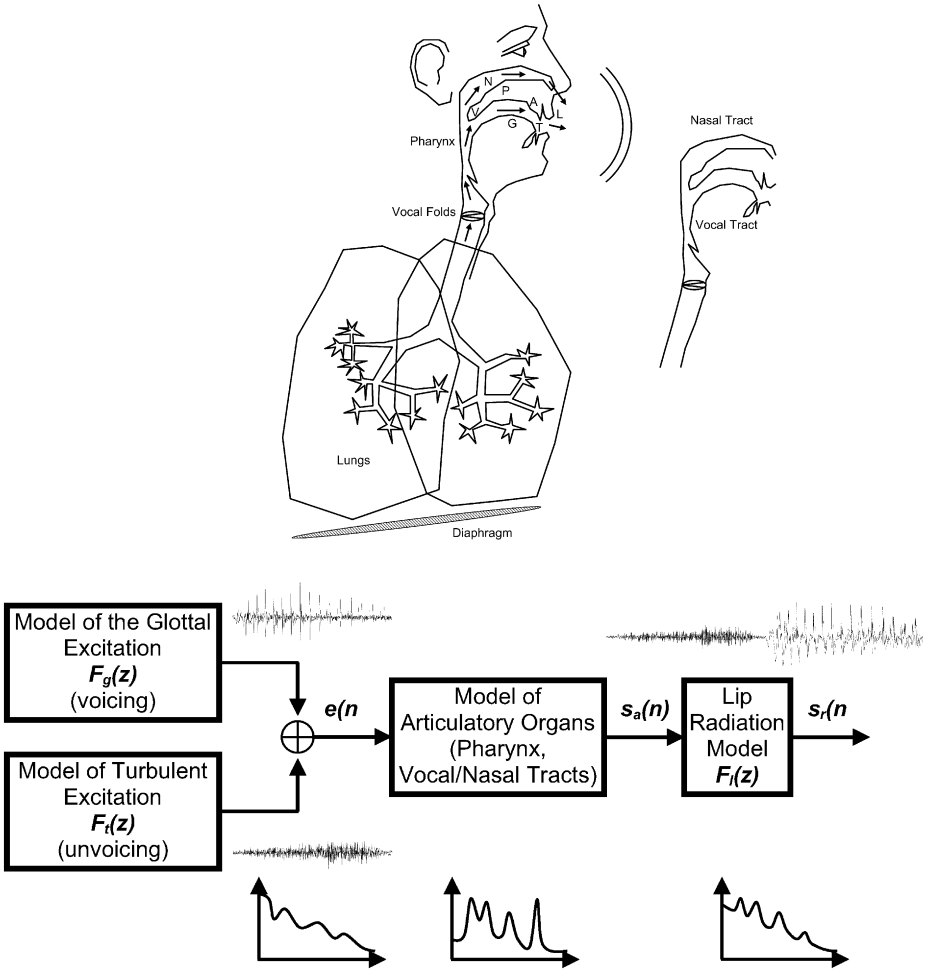

Fig. 1. Top: schematic section of the speech apparatus (the vocal tract has been separated to its right). Bottom: speech production model.

sequential acoustic-phonetic information. The lowest-level semantic units or "semantic atoms" will be reviewed in describing the generalized phoneme at the next semantic level.

- Time-frequency processing in the upper auditory system is briefly reviewed in Section 4 to determine which are the neuronal functionalities associated with the simplest semantic units determined in Section 3.
- Bio-inspired speech processing based on speech production, perception and representation is proposed in Section 5 using an image processing methodology to treat time-frequency representations of speech as images under a multimodal conception of cortical functionalities [4].
- Results from processing selected speech examples with elementary bio-inspired units are presented in Section 6.
- Conclusions are briefly reviewed in Section 7.

2. Fundamentals of speech production

Speech is produced by the combined action of different organs as simplified in the diagram of Fig. 1 (top). The energy for the production of speech is provided by a set of muscles (the diaphragm among them). The airflow resulting from pressure build-up in the lungs induces the vibration of the vocal folds at a specific position (voiced speech component). Under certain conditions (when laminar flow becomes turbulent in the narrow constraints of the vocal tract) sounds are produced also from the modulation of turbulence noise (unvoiced speech component). These are the two possible types of sources exciting the vocal tract, which may appear separate or joint, as symbolized in Fig. 1 (bottom). This means that speech may be divided in voiced and unvoiced segments, depending if vocal fold activity is present or not. The articulation organs (V: velum, G: tongue, P: palate, A: alveoli, T: teeth and L: lips) produce dynamic modifications in the spectral density of the excitation, conveying specific message clues, as the vocal tract behaves as an acoustic filter with time-frequency characteristics resulting from the specific configuration of the articulation organs at each moment [35]. Its role is that of a linear time-variant system enhancing or reducing certain frequencies at the resonant and anti-resonant positions of the equivalent acoustic sound way [5]. When voicing is present the resulting time-frequency representation (spectrogram) will be characterized by horizontal bands at the harmonics of the fundamental frequency of the vocal fold vibration, as shown in Fig. 2 (top), corresponding to an utterance of the C–V syllables /fa-θa-ʃa-χa/.[1] The resonances of the vocal tract enhance the energy of the nearby harmonics, producing a specific intensification of those bands at the harmonic position by the changing vocal tract transfer function modified constantly by the articulation organs as in Fig. 2 (middle), where the resonances manifest themselves as bright bands (the first two formants may be clearly appreciated in all cases as parallel bands near 800 and 1300 Hz). This is so for the vocalic core of the syllables

[1] The International Phonetic Alphabet as described in [6]. has been used for phonetic annotation throughout the paper.

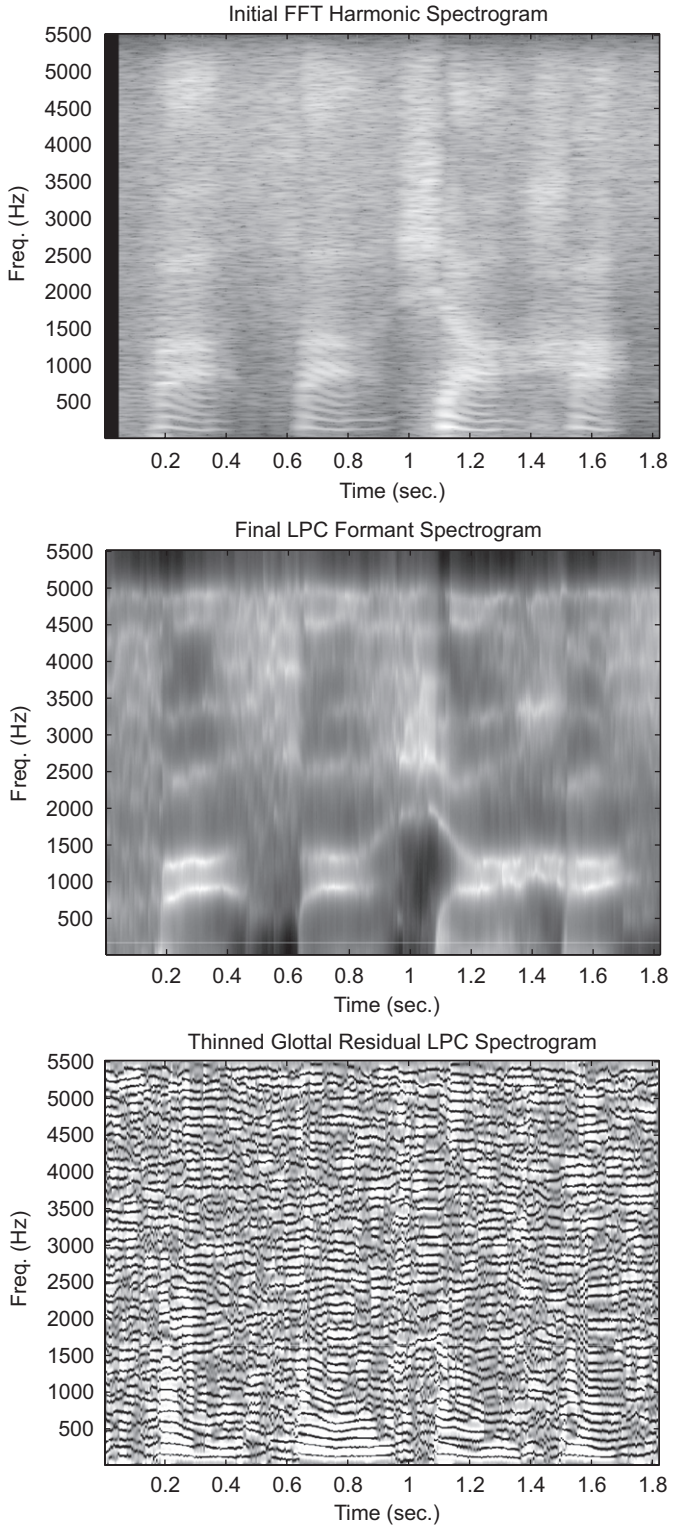

Fig. 2. Voiced and unvoiced speech. Top: FFT spectrogram corresponding to the syllables /fa-θa-ʃa-χa/ uttered by a Spanish male speaker. Middle: time-frequency formant positions (in bright) from the adaptive lineal prediction (ALP) spectrogram. Bottom: harmonic structure from glottal source spectrogram.

(except in the case of whispered speech). For unvoiced speech strong intensifications can still be perceived on the spectrum, as it may be appreciated in the same template around 1 s. The harmonic structure of the voice source is given in Fig. 2 (bottom).

It may be seen that the fricatives can be distinguished by their somewhat different spectra, corresponding to the articulation place, the spectrogram being more widespread for /f/ than for /θ/, showing maxima for /ʃ/ around 2000 and 2600 Hz, and marking clear (but less stable) maxima near the formants of /a/ for /χ/, this last phenomenon being in agreement with the results presented in Fig. 6 (bottom) for the articulation place of the velar /χ/.

The four unvoiced fricatives are characterized by certain bursts of wideband noise preceding the vowel formants. Due to this behavior normal speech may be perceived as sequences of harmonic series colored by the resonances of the vocal tract (formants) with characteristic onsets and trails, which may be preceded or followed by consonant-specific noisy bursts. Therefore harmonics and formants would play a dominant role in speech perception regarding timbre and meaning. Harmonics, their position, stability and evolution are mostly related with the identity of the speaker, while the first three formants convey articulation meaning to the message. In the present case, as the aim of the work is the study of time-frequency representation spaces of speech conveying the semantics of message, formant positions and dynamics, as well as unvoiced burst coloring are the matter of further study in relation with relevant aspects on how speech is perceived.

3. Speech perception

In the previous section it has been discussed how articulation (tongue position relative to lips, teeth, palate or velum, and the aperture of the constrictions) introduce resonances which change the perception of the sound produced and induce relevant changes in the time-frequency features of the acoustic signal to introduce semantic clues understandable by the auditory system as message coding patterns. The acoustic features acting as baseline semantic units are vowel tract resonances called formants, which give a good description of the message issued (acoustic-phonetic decoding) as well as the speaker's personality (see Fig. 3). Formants are labeled in order of increasing frequency, F_1 being the lowest in the range of 250–700 Hz in this case. F_2 sweeps a wider range, from 800 to 2200 Hz. Formants F_3, and higher may be also present in voiced speech, however, the lowest two formants give a good description of vowel-like phonemes. Each combination of the first two formants is decoded by the auditory system as a vowel, and assigned a different value accordingly to the phonologic structure of the target language (some languages as Arabic use a reduced vowel set on separate positions of the vowel triangle, others as Spanish or Japanese are based on five cardinal vowels, and many others posses a richer palette of vowels, as English or French). It is of most importance to emphasize here that the assignment of semantics to specific combinations of formant peaks is highly dependent on the specific language coding system, and therefore universal rules may be hardly applicable. Nevertheless formants are the lowest level semantic units, which when grouped in pairs result in second level meaningful units (vowels) accordingly with the phonologic rules of each specific language. Certain formants can be considered semantic units in themselves as their presence or absence may change the meaning assigned to a given articulation. In the present work as the phonologic rules will be based on Spanish an example based on the five cardinal vowels used in standard Spanish has been used for the example shown in Fig. 3.

The first two formants for /a/ appear between 750–860 and 1300 Hz, moving to 450–500 and 2200 Hz for /e/, 240–310 and 2400–2470 Hz for /i/, then to 500–560 and 850–980 Hz for /o/ and 340–390 and 750 Hz for /u/. This is seen in the vowel triangle represented in the bottom template of Fig. 3, plotting F_2 vs F_1. As a conclusion vowels are represented by relatively stable

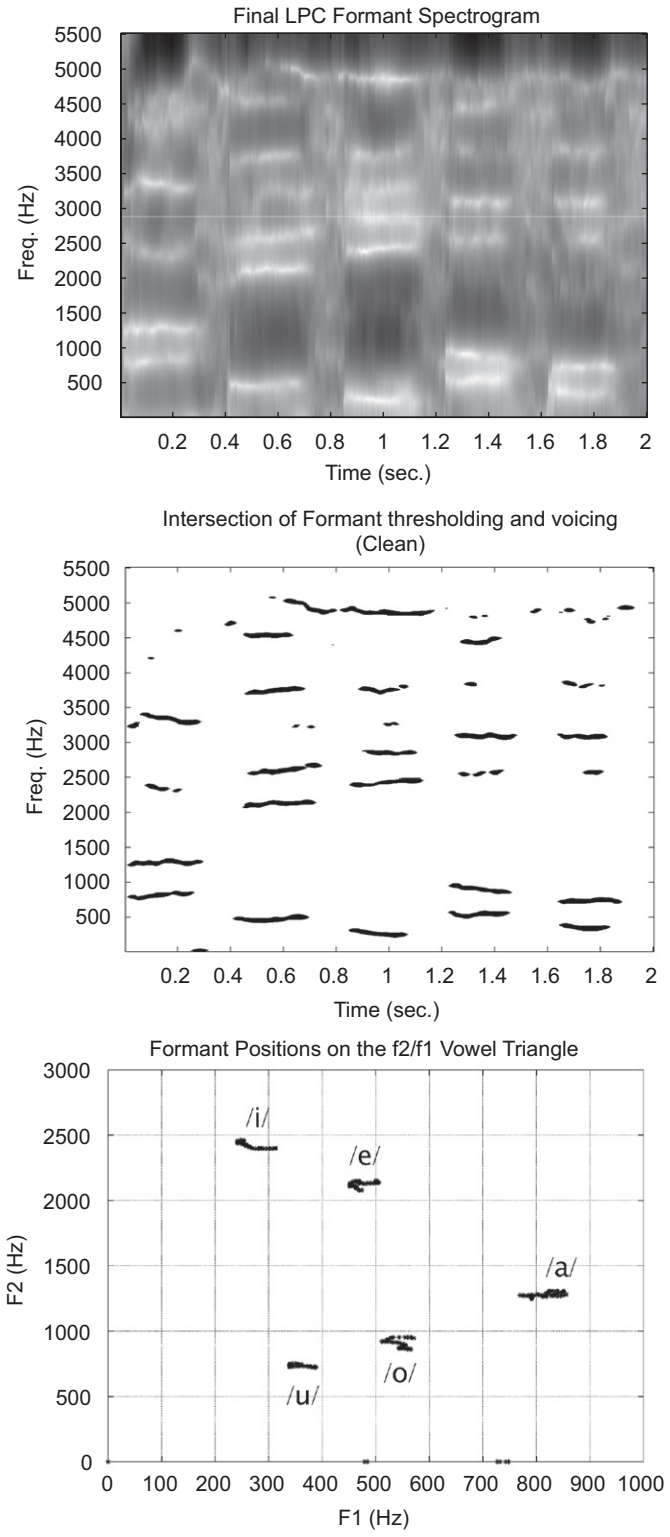

Fig. 3. Top: spectrogram of the five vowels in Spanish (/a/, /e/, /i/, /o/, /u/) from a male speaker, obtained by adaptive linear prediction. Middle: formant plot for the same recording. Bottom: vowel triangle showing the five vowel positions.

narrow-band patterns, known phonetically as formants, or perceptually as characteristic frequencies (CF).

Consonant behavior is rather different, as these are produced by dynamic constrictions of the articulation organs, resulting in vowel formant transitions, accompanied in many cases by turbulent sound resulting from air flow in the constrictions. The induced time-frequency behavior is less stable and known as co-articulation (in voicing). Formant transitions from pre- to post- CF positions are known as frequency modulation (FM) components, and can be considered as lowest-level semantic units as well, as opposition among themselves convey changes in meaning. The presence of wide band sounds as a result of turbulence, generally above 2000 Hz are known as noise bursts (NB) or as blips, and can also be considered lowest-level semantic units. These patterns (CF, FM and NB) bear important communication clues [7] with clear semantic interpretation, which for the purpose of description could be named semantic atoms, or "sematoms" from the information-bearing point of view, as they convey the smallest possible information features meaningful in themselves or when found associated to other "sematoms". In the human auditory pathways certain types of neuronal structures are specifically devoted to detect each one of them for their integration into a growing hierarchy of meaning. An example of these "semantic atoms" for consonant sounds is given in Fig. 4. The perception of vowels and dynamic consonants based upon changes of the steady positions of vowels is explained by the "loci paradigm". A locus is a theoretical formant position previous to the insertion of the consonant, marking a virtual place from where formants move to the situation defined by the stable consonant fragment established by phonological consensus. In Fig. 4 (bottom) the first formant moves from a lower locus to the

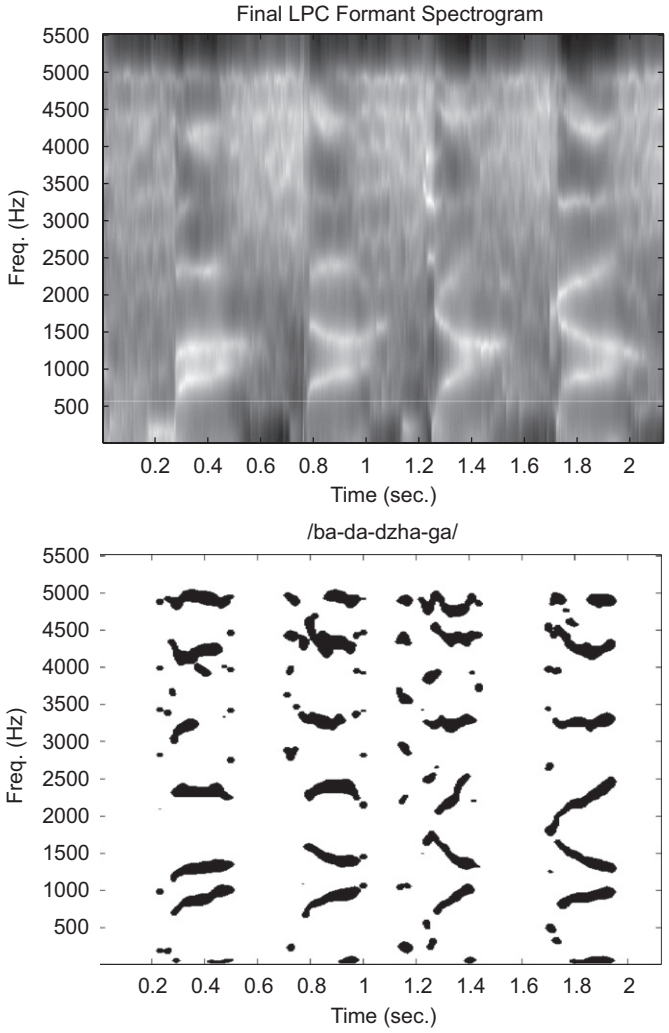

Fig. 4. Top: LPC spectrogram of the syllables /ba/, /da/, /ʄa/, /ga/ from the same speaker. Bottom: formant plots.

CF of /a/ for /ba/ (0.2–0.5 s), /da/ (0.7–1.0 s), /ɟa/ (1.1–1.4 s) and /ga/ (1.7–1.95 s). On its turn F_2 moves from three different loci: below its CF for /ba/, and above CF for /da/, /ɟa/ and /ga/. This dynamic (non-stationary) formant behavior is related to the character of the consonant perceived. The locus theory is of most importance to understand consonant production and perception [8]. It may be observed that F_1 climbs up in all cases from a virtual locus (800 Hz) to 1000 Hz, while F_2 descends from 1800 to 1400 Hz although at a different rate, which for /ɟa/ is the steepest one. Upper blips are clearly observed in this last consonant (1.22 s) at a frequency of 2500 and 3400 Hz, extending to 3800 Hz as a NB. Nasal blips around 200 Hz are also perceptible in all four cases, being lower for /ba/ and /ɟa/ than for /da/ and /ga/. Similar patterns may be observed for the unvoiced consonants /pa/, /ta/, /ca/ and /ka/, as shown in Fig. 5 (top). In this case the formant onsets do not show a clear nasal *blip*, but the general tendencies of the formats are similar to the cases of /ba/, /da/, /ɟa/ and /ga/. Another important "sematom" for the perception of the consonant in the case of /ca/ is the presence of a column of blips just before $t = 1$ s (at 1000, 1700, 2500, 3000, 3500, 4000, and 4500 Hz). Two blips are also present before the onset of the two first formants in the case of /ka/. This tendency is also evident in the case of the unvoiced fricatives given in Fig. 5 (middle). Finally, in Fig. 5 (bottom) the dynamic evolution of the formants from four nasals is presented, corresponding to the same articulation positions studied before. In this case, besides observing a similar dynamic behavior for the first two formants, a nasalization bar appears at a frequency between 200–300 Hz. Having in mind all these observations a generalized phoneme description integrated by "sematoms" of different nature may be abstracted from the facts described before as shown in Fig. 6 (top) where the temporal patterns of a typical phoneme are shown based on the nuclear vowel system and the virtual pre-onset and post-decay positions. The description is based on a vowel nucleus defined by formants F_1 and F_2, which can be considered as level-2 "sematoms" as their semantic value is conveyed in the association of two frequencies (level-1 "sematoms").

Formants are characterized by well defined relatively stable CF positions. The onset is marked by formant F_1 moving from a specific locus (L_{11}) to the final CF position (positive FM). The formant F_2 may move from a low frequency locus (L_{21}) (positive FM) or from high frequency ones (L_{24}, L_{25}) (negative FM) depending on the specific articulation place of the frontal consonant. Blips appear mainly in palatal articulations, and extend to wide-band patterns (with frequencies above 3000 Hz).

Loci in the decay side evolve to next vowel or consonant articulation places, and sometimes are associated to the following phoneme, in the sense that the articulation organs anticipate the position before the next phonation realization [9]. Nasalization appears as a low formant F_n which must not be confused with the glottal formant F_g. The number of formants above F_3 is variable and speaker dependent. If the first two formants were plotted on cartesian coordinates a specific consonantal system would be described by the dynamic trajectory shown in Fig. 6 (middle). Moving from the initial (onset) locus (L_{11}, L_{21}) for /ba/, (L_{11}, L_{24}) for /da/, (L_{11}, L_{25}) for /ɟa/ and (L_{12}, L_{22}) for /ga/ through the position of the vowel (/a/ in this case) ending in the final (decay) locus (CF positions) of the actual vowel or the next vowel or consonant with which it is being co-articulated. The positions of the formant nucleus and the virtual loci would be different for different vowels, accordingly to the articulation place of each consonant and coarticulated vowel. To illustrate these facts with a real example consider the formant trajectories produced by the group /aβa-aða-aʒa-aɣa/ shown in Fig. 11 (top) when plotted on the F_2/F_1 vowel triangle as given in Fig. 6 (bottom). The "sematomic" structure is given by the starting point (pre-onset locus), the

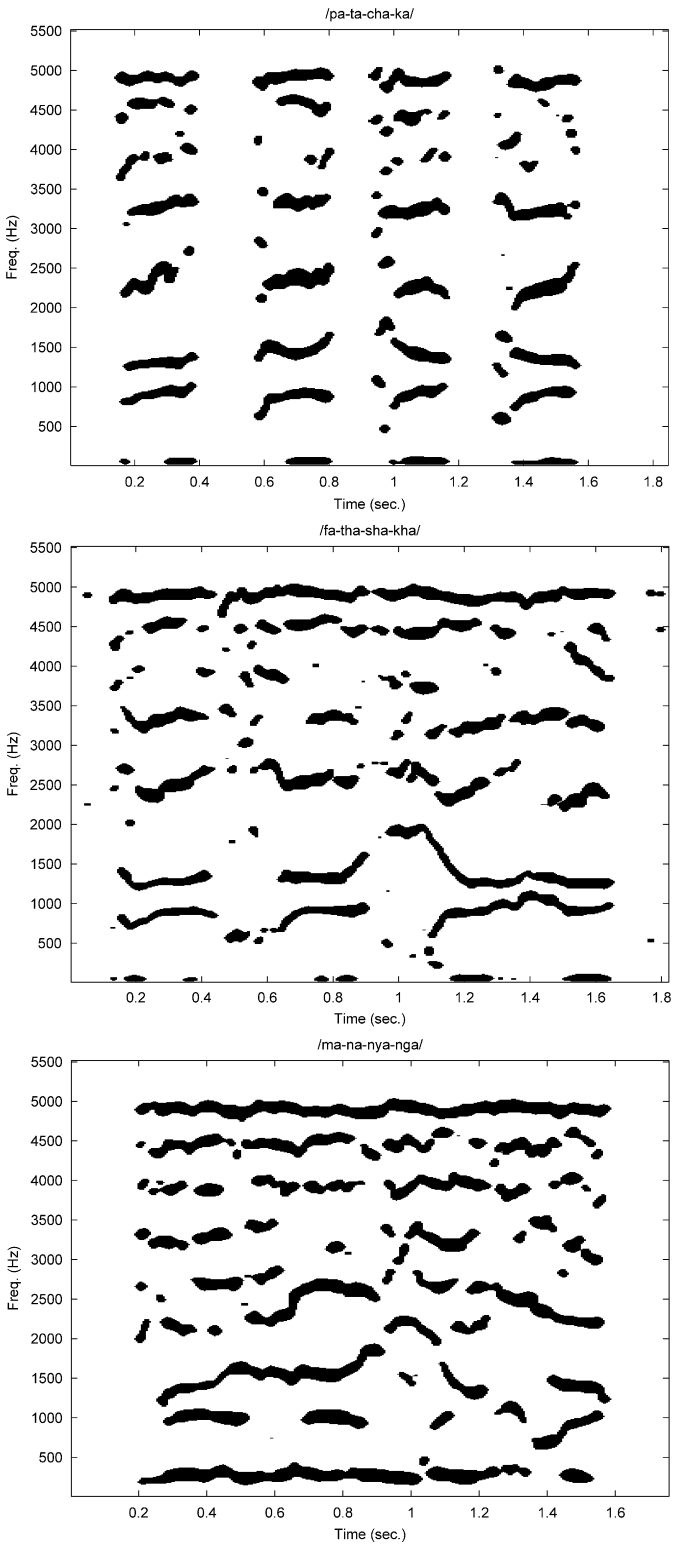

Fig. 5. Top: formant plots of the syllables /pa/, /ta/, /ca/, /ka/ from the same speaker. Middle: idem of /fa/, /θa/, /ʃa/, /χa/. Bottom: idem of /ma/, /na/, /ɲa/, /ŋa/.

trajectory to the vowel locus, and the vowel locus in itself. Incidentally the reader should perceive how these trajectories point to specific vowel triangle positions, and guess where those trajectories are aiming to: in other words, why is there any /i/ flavor in the onset of /ɟa/. To help with the riddle, compare the trajectory patterns against the vowel triangle in Fig. 3 (bottom). Now another puzzle for the interested reader from what has been

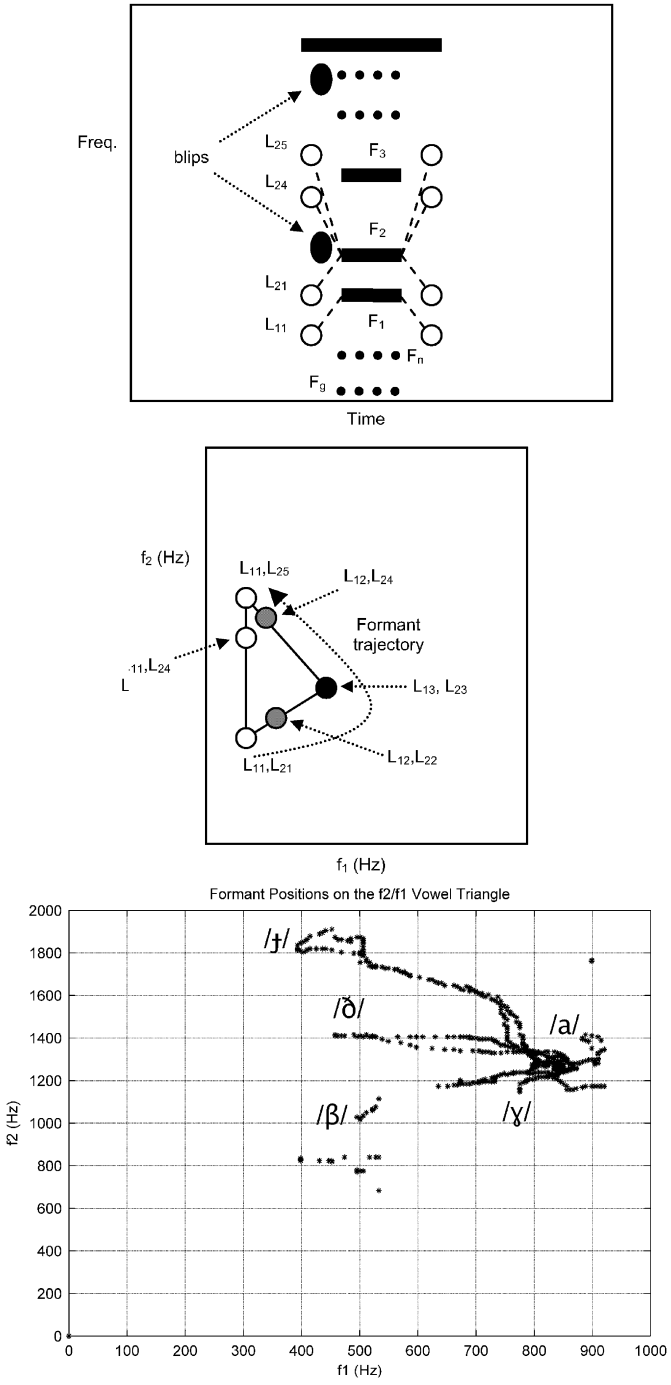

Fig. 6. Top: generalized phoneme description. Middle: loci of the GPD on the vowel triangle. White circles indicate the positions of the loci. The dark dot gives the position of the specific vowel modelled (/a/ in the present case). Bottom: real phonemic trajectories for the group /aβa-aða-aʒa-aɣa/ shown in Fig. 11.

discussed: try to infer how the trajectory of /ɛʒə:/ as in /pleasure/ looks like. Hint: visualize the locus of /ɛ/ close to /e/ and /ə/ near the center of gravity of the vowel triangle.

4. Time-frequency processing of speech-like sounds in the upper auditory system

The development of speech as based on the articulation and perceptual principles briefly examined has been parallel with the creation of representation spaces at the level of the mammal auditory system as is the notion of formant calls [10]. Nevertheless the determination of these representation spaces (static and dynamic) remains incomplete due to the enormous richness of neural phenomena in the auditory pathways as revealed from experimentation and to the problems in direct measurement of the activity in the auditory pathways [36]. The determination of the specific neural structures involved in speech processing and their precise functionality is still an open problem, as direct in vivo measurements is affected by the highly invasive nature of the testing techniques, which motivate that most of the neurophysiologic tests have been conducted on animal models [11–13]. Other sources of knowledge are based on indirect evidence from studying perceptual alterations in humans after brain damage (either induced by illness or by external injuries) [14,15]. Only recently functional magnetic resonance imaging (fMRI) and magneto-encephalography (MEG) have been introduced as real-time introspection tools, although with the limitations in resolution shown by these promising technologies [16,17]. Other sources of indirect data are perceptual tests, affected by other limitations [18].

With these conditionings in mind a brief review of the most relevant facts in the knowledge of the basic processing functions of specific neural tissue from neurophysiologic evidence are exposed. Needless to say, this overview will not intend to describe exhaustively the important time-frequency processing which takes place in the different neural structures, but to summarize some of the main phenomena of interest for speech processing under the perceptual point of view: tone intensity and pitch perception, harmonic and formant estimation, noise-like broadband signal perception, amplitude and FM, vowel onset, sustain and decay detection and its relation to consonant perception, etc. not to speak of sound source location and binaural hearing, which will not be treated here.

With this idea in mind the description will start from the point where speech-induced activity in the auditory nerve (AN) after time-frequency separation and transduction in the auditory peripheral system (cochlea) [19] takes place. Speech processing by the auditory system starts when acoustic signals arrive to the cochlea through the outer and middle ear. Frequency, time and space separation of signal components is produced in the basilar membrane, along the cochlea (see Fig. 7), operating as a filter bank. Low frequencies produce maximum excitation in the apical end of the membrane, while high frequencies produce maximum excitation towards the basal area. These locations code different frequency stimuli present in speech inducing the excitation of transducer cells (hair-cells) at different positions along the cochlea which will be responsible for the mechanical to neural transduction process to electrical impulse trains. This results in organized spike-like streams of stimuli coding frequency by place and phase locking, which are transferred form the cochlea to the first relay stage in the cochlear nucleus (CN) via the AN as depicted in Fig. 7 (Top). These propagate to higher neural centers along auditory nerve fibers, each one being specialized in the transmission of a different characteristic frequency (CF). CF fibers tend to respond to each of the spectral components found in the signal spectrum, although they respond also to nearby tones. Frequency is also encoded in the inter-peak intervals (phase locking) within each group of CF responses [21,22]. The firing rate codes the intensity of the stimulus as well. This information flow is transferred to the CN where different types of neurons specialized in elementary time-space processing are found, some of them segmenting the signals (Cp: chopper units), others detecting stimuli onsets in order to estimate inter-aural differences (On: onset cells), others delaying the information to detect hidden temporal relationships (Pb: pauser units), while others serve as information relay stages (Pl: primary-like units). The CN

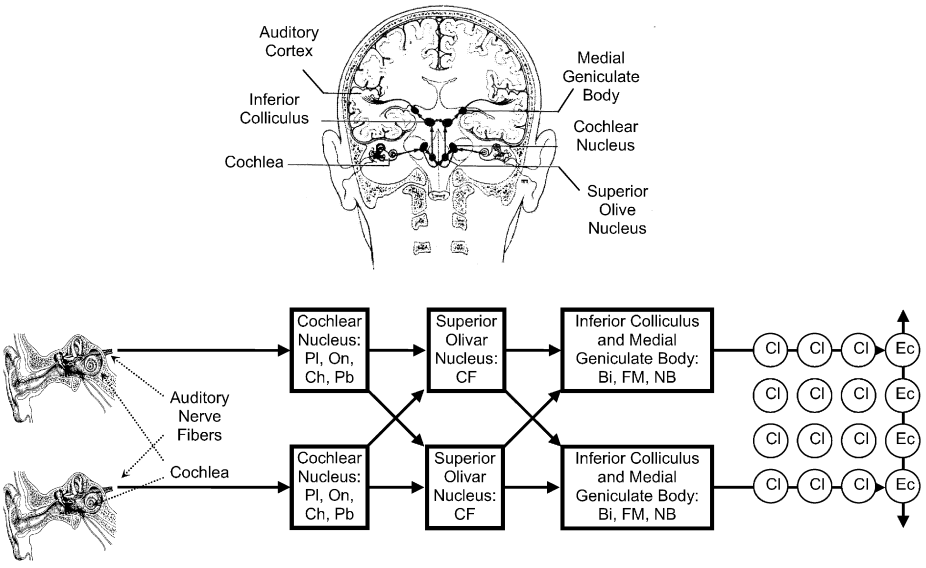

Fig. 7. Speech perception model. Top: main neural pathways in the peripheral and central auditory centers (taken from [20]). Bottom: simplified main structures. The cochlea produces time-frequency organized representations which are conveyed by the auditory nerve to the cochlear nucleus, where certain specialized neurons (Pl: primary-like, On: onset, Ch: chopper, Pb: pauser) are implied in temporal processing. Binaural information is treated in the superior olivar nucleus, where selective tono-topic units (CF) may be found. Other units specialized in detecting tonal movements (FM), broadband spectral densities (NB) and binaural processing (Bi) are found in the inferior colliculus and the medial geniculate body. The auditory cortex shows columnar layered units (Cl) as well as massively extensive connection units (Ec).

feeds information to the olivar complex, where sounds are located by interaural differences, and to the inferior colliculus (IC), which is organized in spherical layers with iso-frequency bands orthogonal to each other. Delay lines of up to 12 ms are found in its structure, their function being related with the detection of temporal elements coded in acoustic signals (CF and FM components). Fibers irradiate from this center to the thalamus (medial geniculate body) which acts as a relay station for prior representations (some neurons exhibit delays of a hundred milliseconds), and as a tonotopic mapper of information arriving to cortex, where high level processing takes place. It seems that the neural tissue in the brain is organized as ordered feature maps [15] according to this sensory specialization. The specific location of the neural structures in the cortex responsible for speech processing and understanding is not well defined in humans as the subjects of experimentation have been mainly animals. This fact and the need of using anesthesia to record single neuron responses in animal models lacking speech abilities puts some shades to the elaboration of theories on speech processing by the upper auditory system [13]. Although recent reports on speech brain center research using nuclear magnetic resonance have been published, these studies do not reach single neuron resolutions yet. Nevertheless some findings in neurophysiologic sound perception in animals may give interesting hints on which phenomena take place and where, which may be of interest for speech processing understanding. For example, in cats neurons have been found that fire when FM-like frequency transitions are present (FM elements) [11], while in macaque some neurons respond to specific (NB components) [12]. Other neurons as in bat's brain are specialized in detecting combinations among these elements (CF_1–CF_2 and FM_1–FM_2) [23]. In humans, evidence exists of frequency representation maps of this kind near the Heschl circumvolution [24] and of a secondary map with word-addressing capabilities [14]. These findings are of most interest for bio-inspired speech processing. A summarized description of the structures involved and their functionality is given in [7]. As a review it must be emphasized that the specific processing of speech by the Auditory System is based on the detection of stable frequencies, associations of frequencies, onset times, dynamic frequency changes, and tone bursts. At the first hierarchical level CF units are specialized in the detection of single tones associated among themselves or as running streams [25]. At a second hierarchical level associations of tones, in many cases separated by large frequency intervals are detected as specific semantic units (vowels being among these). Formant dynamics detection would be based on the capacity of the auditory system of extracting formants and associations between formants from relations among neighbor harmonic positions during short time intervals in using some left hemisphere mechanisms [17]. Specific sets of neurons are devoted to isolate formants from neighbor harmonic relations using association and lateral inhibition (see [8,26]). At a higher hierarchical level dynamic changes in harmonics (onset times and slopes) and specific broadband signals present before the onset time define specific clues to the perception of syllables, seen as associations of consonants and vowels as in C–V structures (other possibilities contemplate tri-phone structures of the kind C–V–C or V–C–V). The perceptual interpretation of such structures is well documented in literature [18].

5. Bio-inspired speech processing

The rich phenomena taking place in the Auditory Pathways exhibits a tremendous complexity to be exploitable in its whole extent, and in many cases the coding solutions found at the biophysical level may not be fully exportable to bio-inspired systems as they are. In other cases, as in Cochlear Implant design [27] strong adherence to the natural systemic coding is not only desirable, but mandatory, as the nature of stimuli induced on Auditory Centers should imitate the role of the real system as close as possible to optimize speech understanding. This needs not be the case in most applications related with Speech Processing, where global functional behavior is the target to translate natural process functionality to artificial structures and algorithms. The most interesting capabilities found conveying meaningful information to be imitated are frequency selectivity (pitch and formant detection), the discrimination between narrow and broad band signals, the onset and decay insertions, the

association of tones, the capability of pitch reconstruction—even in case of missing or virtual pitch, the tracking of moving frequencies (bird chirp), etc. From the study of the generalized phoneme description and the auditory speech processing fundamentals summarized before, a basic neuron set could be defined as an algorithmic structure operating both in the time and frequency domains modeling a generalized set of functions in the biological structures studied [23]. These structures need not be physiological units in themselves, rather than that it is the behavior of such functional "virtual" units what is of most interest under the systemic point of view to design a minimal set of "leaf-cells" which could form part of a higher hierarchy tree for artificial bio-inspired processing built following a bottom-up approach in the implementation of algorithms and hardware structures supporting the before mentioned "sematoms". A library of elemental "leaf-cells" would then include:

- Lateral inhibition units (LI), which can be seen as a finite difference algorithm in the frequency domain profiling formants from harmonics.
- Temporal derivative units (TD), or finite difference in the frequency domain (choppers, built-ups).
- Positive frequency modulation units (PfM), detectors of up-hill formant displacements.
- Negative frequency modulation units (NfM), detectors of down-hill formant displacements.
- CF integrating units (CfI), detectors of stable frequency positions.
- Broad formant set units (BfS), detectors of stable or parallel-moving pairs of frequencies.
- Noise-burst units (NB), detectors of wide-band noise-like signals.

These elementary processing units could be implemented adequately programming the general cell shown in Fig. 8.

The problem of feature detection in formant spectrograms can be related to others similar in digital image processing [28]. A classical methodology in such field is based on operating the image matrix $X(m,n)$ using reticule masks

$$\tilde{X}(m,n) = \sum_{i=-1}^{M} \sum_{j=0}^{N} w_{ij} X(m-i, n-j) \quad (1)$$

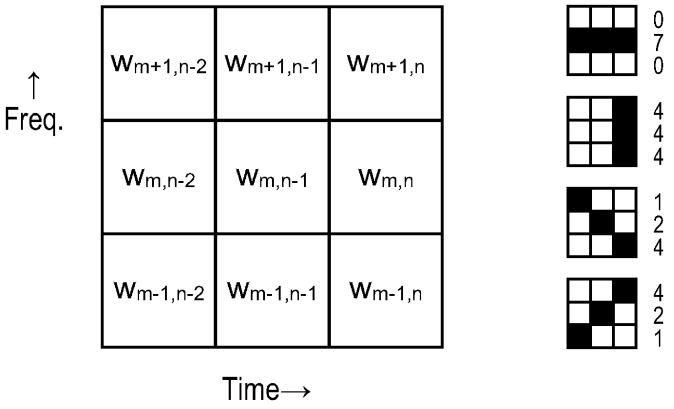

Fig. 8. Basic neuron set for elementary operations on time-frequency representations of speech. Left: 3×3 weight mask. Right: masks for feature detection on the formant spectrogram. Each mask is labeled with the corresponding octal code (most significant bits: bottom-right). Labels 070, 444, 124 and 421 correspond respectively to CfI, NB, NfM, PfM units.

where $\{w_{ij}\}$ is the set of weights associated to a $M \times N$ mask with a specific functional pattern, m and n being the respective frequency channel and time indices (a 3×3 one such neuron is given in Fig. 8 as an example). It may be shown that the generic filtering in (1) is equivalent to a liftering process [29]. There are two important concepts linked to mask design: the specific pattern in time-frequency, and the specific weight adjustment. The basic cells for formant trajectory processing shown in Fig. 8 have been derived from the neural structures of the auditory centers in brain, as presented in Section 4. To produce unbiased results, the weight associated to each black square is fixed to $+1/s_b$ and the weight associated to white squares is fixed to $-1/s_w$, s_b and s_w being the number of squares in black or white found in a $M \times N$ mask, respectively. Weight adjustment may also be adaptively fixed using a paradigmatic database of sample spectrograms and MLP structures for the training of each cell [30]. The results presented here are based on pre-assigned weights, their adaptive version being under production. The indexing in the time domain is intended to preserve the causality principle, whereas the indexing in the frequency domain is established to allow the detection of dynamic changes (as in FM units).

6. Selected results

To show the applicability of this methodology to speech processing tasks an example of broad class phonetic labeling is presented in this section. In this example the front end used in the detection of formants carried out in the biological case by elaborate methods plausibly involving correlation and integration of CF unit spike streams [22] is implemented through adaptive linear prediction (ALP) methods, the rest of the functions used in the example being fully bio-inspired. Although there are good models [31] to deal with time-frequency stimulus separation, the present work will rely on a more classical approach to carry out the following pre-processing: detect voiced and unvoiced sounds, separate the glottal source from the vocal tract transfer function to better detect formants, and estimate the power spectral density of unvoiced sounds. The separation of the vocal and glottal information is crucial for the robust detection of formants in voiced sounds, as it is well known that in certain sounds a low first formant may be easily confused with the glottal formant, resulting in defective formant assignment. The implementation of the vocal tract inversion has been carried out using an adaptive lattice filter [32] capable of following the inherent non-stationary nature of speech, allowing the estimation of the vocal tract transfer function and its removal from speech, leading to a precise reconstruction of the glottal source as described in [32]. ALP algorithms produce all-pole spectral positions which keep track of the vocal tract resonances [29].

Precise formant positions may be obtained from these rough representations applying lateral inhibition between neighbor CF as plausibly is done in the auditory pathways (some evidence points out to units in the CN, others to the IC [15] or even to auditory cortex [22]). Lateral inhibition may be implemented using a specific weight configuration of the mask in Fig. 8 as shown in Fig. 9.

It may be seen that the lateral inhibition filter produces sharp estimations of the spectral peaks (see Fig. 9c). The whitish bands surrounding the formants are due to the characteristic "mexican hat" response of such a filter. The final formant distribution is given in Fig. 9d after adaptive nonlinear saturation. This filtering may be seen as a special case of (1) where the weights of columns $j = 1,2$ have been filled with zeros. The whole systemic framework is given in Fig. 10. Once the power spectral density of the vocal tract transfer function has been decoupled from the glottal source,

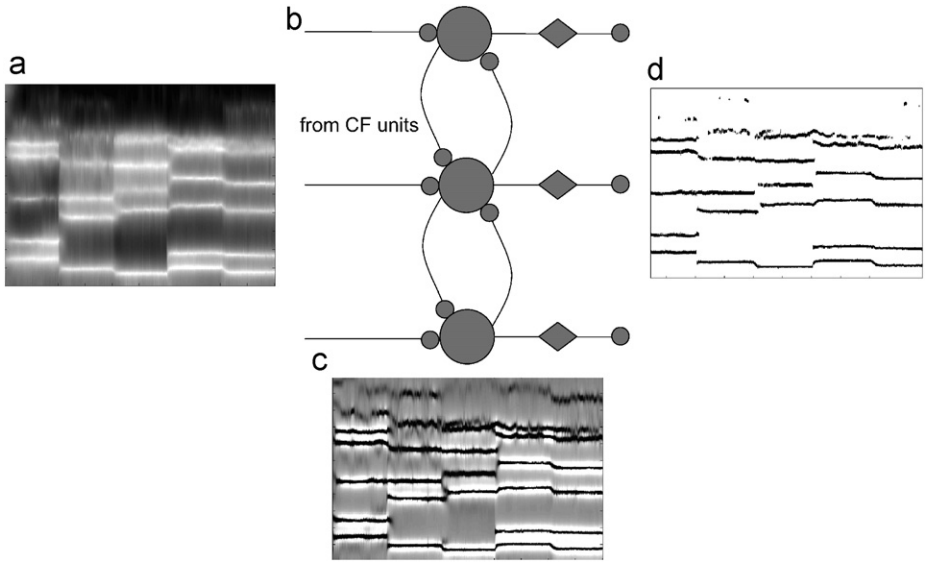

Fig. 9. Formant trajectory profiling for the sequence /aeiou/ (Spanish): the speech spectral density (a) as detected by CF units is processed by columns of neurons implementing lateral inhibition (b), producing differentially expressed formant lines (c), which are transformed into narrow formant trajectories and (d) after non-linear saturation.

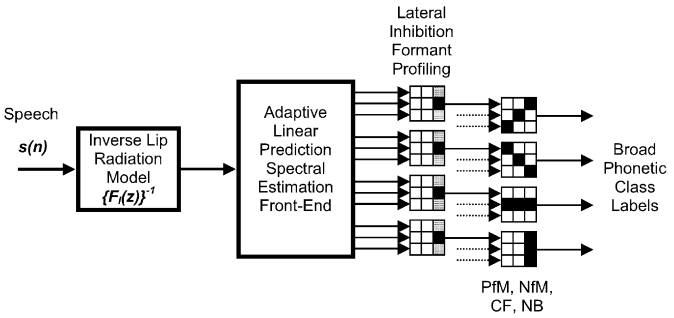

Fig. 10. Bio-inspired speech processing framework used in the study for a monoaural channel. Inverse radiation is removed, an ALP spectrogram is produced and profiled by lateral inhibition units. A layer of PfM, NfM, CF and NB units are in charge of broad phonetic class feature detection.

formants may be estimated from the resulting spectrogram, defined as

$$X(m,n) = 20\log_{10}\left|\sum_{k\in arg\{V\}} V(k)x(n+k)\mathrm{e}^{-\mathrm{j}mk\Omega\tau}\right| \quad (2)$$

where $x(n)$ is the speech signal, $V(k)$ is a specific framing window, and τ and Ω are the resolutions in time and frequency. The representation $X(m,n)$ can be seen as a two-dimensional image, indexed by time (n) and frequency (m). This means that many tools devised for image processing can be used for the detection of time-frequency features, as CF or FM patterns. The first basic operation on the LPC spectrogram will be to profile formant trajectories following the structure given in Fig. 9 using bio-inspired lateral inhibition as explained. The proposed algorithm is expressed as

$$\hat{X}(m,n) = \sum_{i=-1}^{1} w_i X(m-i,n) \quad (3)$$

where the respective weights are $w_{-1} = w_1 = -1/2$ and $w_0 = 1$. This filter has to be applied to each column of the ALP spectrogram. Once the formant trajectories have been profiled by the first layer of lateral inhibitory units a second layer of positive and negative formant dynamics detection units (PfM and NfM) are responsible of pinpointing ascending and descending formants. As an example results are shown from processing four structures of the type V–C–V in the utterance /aβa-aða-aʒa-aɣa/ containing four voiced approximants with articulation places at the lips, teeth, pre-palate and velum, their profiled formant plot being given in Fig. 11 (top). The dimensionality of the spectrograms produced by ALP is of 512 units for speech sampled at 11 KHz, resulting in a frequency resolution of 10.74 Hz. The order of the lattice filters used was 18 for the estimation of the vocal tract transfer function and 2 for the compensation of the glottal tilt. An adaptation step of 0.999 was used for spectral tracking, equivalent to a time constant around 10 ms. The dimensionality of the PfM, NfM, CF and NB units was also of 512 units per layer. Detectors with $M = N = 7$ were used in a hierarchy of different combined topologies for slope detection and grouping. The second template (middle-left) shows the activation of 512 NfM units in time, selecting the negative slope transitions of the formant profile (encircled). It was observed that each NfM unit is rather sensitive to frequency/time slopes, therefore combinations of faster and slower profiles forced to use 7×7 structures with delay units ranging from 5 ms up to 35 ms. A similar observation may be devoted to the middle-right template where the activity of PfM units is shown. It is of most importance to detect slow and fast frequency slopes excluding stable formant positions, as these would result in rather disturbing ambiguities.

Finally, the bottom left and right templates show the firing of CF and NB units responsible of stable formant detection and broad band activity. In this case the input profiled spectrogram corresponds to the utterance /fa-θa-ʃa-χa/ shown in Fig. 5 (middle). The CF units were programmed to select relatively stable frequency positions with a broader tolerance or low spontaneous firing rate (encircled). The same criterion was used in the detection of broadband activity by NF units. The activity exhibited by the different detection units helps in defining production rules which can be used in the labeling of phonetic classes as shown in Table 1.

Vowels and vowel-like sounds will induce mainly activity in CF units, NfM and PfM being almost "blind" to these sounds. C–V articulations including voiced stops {C(vs)–V} will induce activity in the three types of structures in synchrony with the stimulus, therefore the expected sequence would be NB(tg)–FM–CF with

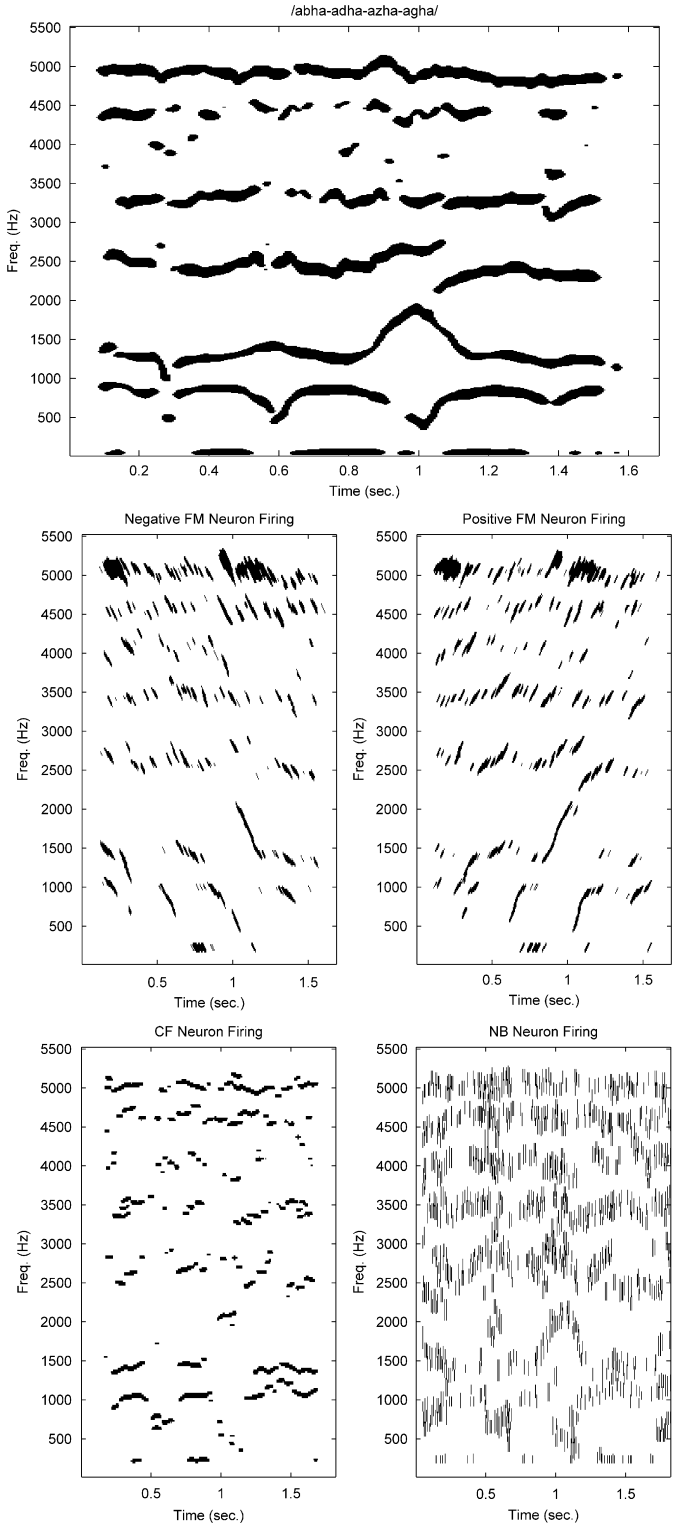

Fig. 11. Detection of four basic formant patterns using PfM, NfM, CF and NB units for the V–C–V groups /aβa-aða-aʒa-aɣa/ of voiced fricatives in Spanish uttered by a male speaker. Top: formant profiled spectrogram. Middle left: outputs from NfM units. Middle right: outputs from PfM units. Bottom left: outputs from CF units for the spectrogram in Fig. 2. Bottom right: outputs from NB units for the same set. The basic patterns detected are enclosed within dash circles.

some overlapping between turbulent and glottal activity (tg). Nasals would be characterized by low CF unit activity in sequences. Glides or approximants would show FM activity mainly. C–V articulations including unvoiced stops {C(us)–V} would produce NB(t)–FM–CF sequences in which glottal source activity would not be present during the NB phase. C–V articulations including voiced fricatives {C(vf)–V} would produce sequences with structure NB(tg)–CF and unvoiced fricatives with model {C(uf)–V} would induce NB(t)–CF activity. It is important to bear in mind that these situations correspond to neat syllabic structures rarely seen in fluent speech examples, where more intermingled situations are to be expected. Nevertheless this labeling may be of great help in certain tasks as ASR where recognition rates may be improved as much as 26% (see [33]) by simplifying state-transition graph search in phonetic HMM parsing, in reducing tri-phone ambiguity, and in lowering computational complexity. A deeper study is being conducted to label specific vowels by CF output association and the detection of glides and voiced stops from FM units output by using supervised neural networks [30].

Table 1
Features of broad phonetic classes.

	CF	FM	NB
Vowels	Yes	No	No
Voiced stops	Yes	+, –	Yes
Nasals	Yes	No	No
Glides	No	+, –	No
Unvoiced stops	No	+, –	Yes
Fricatives	No	No	Yes

7. Conclusions

Through the present work a review of basic concepts related with speech production, Perception and processing has been presented to define a framework for time-frequency representations of speech from neurophysiologic evidence. The results shown demonstrate the viability of bio-inspired phonetic feature detection using computationally inexpensive structures. The unsupervised detection of the basic consonantal features shown as examples indicate the presence of stable, ascending and descending formants (middle templates) characteristic of approximants, stable formants (bottom left) found in vowels, and noise bursts (bottom right) characteristic of both unvoiced and voiced consonants. More work is to be done to establish normalized thresholds and other configuration parameters which show to affect the robustness of the methodology exposed. Another open question is that of statistical performance using large databases of speakers, and their possible use in other tasks, as in speaker's identification. These questions remain the object of future study, as well as the exploration of higher levels of phonetic class association in hierarchies using broad-frequency formant structure units (BfS) and others similar. Bio-inspired phonemic parsing stands as another challenge, for which the columnar organization of the Auditory Cortex [34] to include short-time memory and retrieval systems by means of generalized autoregressive units remains as a study objective.

Acknowledgments

This work is being funded by grants TIC2003-08756 and TEC2006-12887-C02-01/02 from Plan Nacional de I+D+i, Ministry of Education and Science, by grant CCG06-UPM/TIC-0028 from CAM/UPM, and by project HESPERIA (http://www.proyecto-hesperia.org) from the Programme CENIT, Centro para el Desarrollo Tecnológico Industrial, Ministry of Industry, Spain.

References

[1] H. Hermansky, Should recognizers have ears? in: ESCA–NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels, April 17–18, Pont-à-Mousson, France, 1997, pp. 1–10.
[2] P. Gómez, R. Martínez, V. Rodellar, J.M. Fernández, Bio-inspired systems in speech perception: an overview and a study case, in: IEEE/NML Life Sciences Systems and Applications Workshop (by invitation), National Institute of Health, Bethesda, Maryland, July 13–14, 2006.
[3] R.E. Cytowic, The Man Who Tasted Shapes, Abacus Press, London, England, 1993.
[4] A. Pascual-Leone, R. Hamilton, The metamodal organization of the brain, Prog. Brain Res. 134 (2001) 427–445.
[5] G. Fant, Theory of Speech Production, Mouton, The Hague, Netherlands, 1960.
[6] From 〈http://www.arts.gla.ac.uk/IPA/ipachart.html〉.
[7] J.M. Ferrández, Study and realization of a bio-inspired hierarchical architecture for speech recognition, Ph.D. Thesis, Universidad Politécnica de Madrid, 1998 (in Spanish).
[8] P. Delattre, A. Liberman, F. Cooper, Acoustic loci and transitional cues for consonants, J. Acoust. Soc. Am. 27 (1955) 769–773.
[9] E. Martínez-Celdrán, A.M. Fernández-Planas, Manual de fonética española, Ariel, Barcelona, 2007.
[10] D.B. Geissler, G. Ehret, Time-critical integration of formants for perception of communication calls in mice, in: Proceedings of the National Academy of Sciences, vol. 99(13), 2002, pp. 9021–9025.
[11] J.R. Mendelson, M.S. Cynader, Sensitivity of cat primary auditory cortex (AI) neurons to the direction and rate of frequency modulation, Brain Res. 327 (1985) 331–335.
[12] J.P. Rauschecker, B. Tian, M. Hauser, Processing of complex sounds in the macaque nonprimary auditory cortex, Science 268 (1995) 111–114.
[13] S. Greenberg, W.H. Ainsworth, Auditory processing of speech, in: S. Greenberg, W.H. Ainsworth (Eds.), Listening to Speech: An Auditory Perspective, Lawrence Erbaum Associates, Mahwah, NJ, 2006, pp. 3–17.
[14] G.A. Ojemann, Organization of language cortex derived from investigation during neurosurgery, Sem. Neuros. 2 (1990) 297–305.
[15] C.E. Schreiner, Order and disorder in auditory cortical maps, Curr. Op. Neurobiol. 5 (1995) 489–496.
[16] B.R. Buchsbaum, G. Hickok, C. Humphries, Role of left posterior superior temporal gyrus in phonological processing for speech perception and production, Cognitive Sci. 25 (2001) 663–678.
[17] D. Poeppel, The analysis of speech in different temporal integration windows: cerebral lateralization as "asymmetric sampling in time," Speech Commun. 41 (2003) 245–255.
[18] A. Palmer, S. Shamma, Physiological representation of speech, in: S. Greenberg, W. Ainsworth, A. Popper (Eds.), Springer, New York, 2004, pp. 163–230.
[19] J.B. Allen, Cochlear modeling, IEEE ASSP Mag. January (1985) 3–29.
[20] B. Goldstein, Sensation and Perception, Wadsworth, Belmont, 1984.
[21] H. Secker, C. Searle, Time domain analysis of auditory-nerve fibers firing rates, J. Acoust. Soc. Am. 88 (1990) 1427–1436.
[22] S. Shamma, Speech processing in the auditory system II: lateral inhibition and central processing of speech evoked activity in the auditory nerve, J. Acoust. Soc. Am. 78 (1985) 1622–1632.
[23] N. Suga, Cortical computational maps for auditory imaging, Neural Networks 3 (1990) 3–21.
[24] M. Sams, R. Salmening, Evidence of sharp frequency tuning in human auditory cortex, Hearing Res. 75 (1994) 67–74.
[25] P. Yin, L. Ma, M. Elhilali, J. Fritz, S. Shamma, Primary auditory cortical responses while attending to different streams, in: B. Kollmeier et al. (Eds.), Hearing: From Sensory Processing to Perception, Springer, Heidelberg, 2007, pp. 257–265.
[26] J.F. Culling, C.J. Darwin, Perceptual separation of simultaneous vowels: within and across-formant grouping by F0, J. Acoust. Soc. Am. 93 (1993) 3454–3467.
[27] G. Clark, Cochlear implants, in: S. Greenberg, W. Ainsworth, A. Popper (Eds.), Springer, New York, 2004, pp. 422–462.
[28] B. Jähne, Digital Image Processing, Springer, Berlin, 2005.
[29] J.R. Deller, J.G. Proakis, J.H.L. Hansen, Discrete-Time Processing of Speech Signals, Macmillan, New York, 1993.
[30] S. Haykin, Neural Networks—A Comprehensive Foundation, Prentice-Hall, Upper Saddle River, NJ, 1999.
[31] T. Irino, R.D. Patterson, A time-domain, level-dependent auditory filter: the gammachirp, J. Acoust. Soc. Am. 101 (1997) 412–419.
[32] P. Gómez, J.I. Godino, A. Álvarez, R. Martínez, V. Nieto, V. Rodellar, Evidence of glottal source spectral features found in vocal fold dynamics, in: Proceedings of the ICASSP'05, 2005, pp. 441–444.
[33] G. Gravier, Y. Yvon, B. Jacob, F. Bimbot, Introducing contextual transcription rules in large vocabulary speech recognition, in: W.J. Barry, W.A. Van Domelen (Eds.), The Integration of Phonetic Knowledge in Speech Technology, Springer Series on Text, Speech and Language Technology, vol. 25, 2005, pp. 87–106.
[34] V.B. Mountcastle, The columnar organization of the neocortex, Brain 120 (1997) 701–722.
[35] D. O'Shaughnessy, Speech Communication, IEEE Press, Park Avenue, New York, 2000.
[36] N. Suga, Basic acoustic patterns and neural mechanism shared by humans and animals for auditory perception: a neuroethologist's view, in: Proceedings of the Workshop on the Auditory bases of Speech Perception, ESCA, July, 1996, pp. 31–38.

Pedro Gómez-Vilda was born in Burgo de Osma, Spain in 1952. He received the M.Sc. degree in Communications Engineering in 1978 and the Ph.D. degree in Computer Science from the Universidad Politécnica de Madrid, Madrid, Spain, in 1983. He is Professor in the Computer Science and Engineering Department, at Universidad Politécnica de Madrid since 1988. His current research interests are biomedical signal processing, speaker identification, cognitive speech recognition, and genomic signal processing. Dr. Gómez Vilda is a member of the IEEE, ISCA and EURASIP.

J. Manuel Ferrández Vicente was born in Elche, Spain in 1969. He received the M.Sc. degree in Computer Science in 1995, and the Ph.D. degree in 1998, all of them from the Universidad Politécnica de Madrid, Spain. He is currently Associate Professor at the Department of Electronics, Computer Technology and Projects at the Universidad Politécnica de Cartagena and Head of the Electronic Design and Signal Processing Research Group at the same University. His research interests include bioinspired processing, neuromorphic engineering and cognitive speech recognition.

Dr. Victoria Rodellar-Biarge was born in Huesca, Spain. She received the M.Sc. and the Ph.D. degree in Computer Science from the Universidad Politécnica de Madrid, Madrid, Spain. She is Associate Professor in the Computer Science and Engineering Department, at Universidad Politécnica de Madrid. Her current research interests are biomedical and genomic signal processing and reconfigurable logic designs for DSP. Dr. Rodellar-Biarge is a member of the IEEE.

Roberto Fernández-Baillo was born in Madrid (Spain) in 1973. He got a grade in Speech Therapy from Universidad Complutense of Madrid in 1995. In 1997 he graduated in Neurolinguistics at Universidad Complutense and the University Hospital "Gómez Ulla". In 2005 he obtained the grade of Orofacial Alterations from Universidad Pontificia of Salamanca. He has been working in the Laboratory of Locomotive Apparatus Biomechanics in the Medicine School of Universidad de Alcalá de Henares. Since 2006 his professional activity is developed in the Laboratory of Speech Communication of the Universidad Politécnica de Madrid.