Neurocomputing 74 (2011) 1191–1202

Contents lists available at ScienceDirect

Neurocomputing

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

Neuromorphic detection of speech dynamics

Pedro Gómez-Vilda [a,*], José M. Ferrández-Vicente [b], Victoria Rodellar-Biarge [a], Agustín Álvarez-Marquina [a], Luis Miguel Mazaira-Fernández [a], Rafael Martínez Olalla [a], Cristina Muñoz-Mulas [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 16 October 2010

Keywords:
Neuromorphic computing
Auditory pathways
Phonetic labelling
Contextual speech information

ABSTRACT

Speech and voice technologies are experiencing a profound review as new paradigms are sought to overcome some specific problems which cannot be completely solved by classical approaches. Neuromorphic Speech Processing is an emerging area in which research is turning the face to understand the natural neural processing of speech by the Human Auditory System in order to capture the basic mechanisms solving difficult tasks in an efficient way. In the present paper a further step ahead is presented in the approach to mimic basic neural speech processing by simple neuromorphic units standing on previous work to show how formant dynamics – and henceforth consonantal features – can be detected by using a general neuromorphic unit which can mimic the functionality of certain neurons found in the upper auditory pathways. Using these simple building blocks a General Speech Processing Architecture can be synthesized as a layered structure. Results from different simulation stages are provided as well as a discussion on implementation details. Conclusions and future work are oriented to describe the functionality to be covered in the next research steps.

1. Introduction

Neuromorphic Speech Processing is an emerging field which has attracted the attention of many researchers looking for new paradigms helping to better understand the underlying brain processes involved in speech perception, comprehension and production [10,21]. This study can also be extended to cognitive audio (voice and sound processing by humans in general) when aspects as emotion or speaker recognition are concerned or in scene analysis [20,21,30]. The present paper is aimed to extend previous work on Neuromorphic Speech Processing [8] using a layered architecture of artificial Neuron-like Units derived from the functionality of the main types of neurons [14] found in the auditory pathways from the cochea to the primary and secondary auditory cortex [9]. In these early stages the typology a General Neuromorphic Computing Unit (GNCU) was defined using well-known paradigms from mask Image Processing [13]. It was also shown in the referred previous work [9] how one of these Mask Units can be adapted to model different processes as Lateral Inhibition to enhance Formant Detection. It was also shown how using different masks the GNCU could be configured to detect formant dynamics (ascending or descending resonance patterns appearing in certain speech sounds). The present work is intended to show how based on this GNCU a general layered architecture can be defined for the labelling of phonemes from formant positions and dynamics, advancing one step in the definition of a fully Bio-inspired Speech Processing Architecture. The paper is organized as follows: A brief description of formants and formant dynamics is given in Section 2. In Section 3 the different units found in the Auditory Pathway are defined accordingly to their functionality. The structure of the GNCU is shown to mimic the different units of interest for Speech Processing, and a Neuromorphic Speech Processing Architecture based on these units is presented. The purpose of Section 4 is to introduce plausible neural circuits to implement specific functions and comment results from simulations. Conclusions and future work are presented in Section 5.

2. Perceiving the dynamic nature of speech

Speech can be defined as the result of a complex interaction of the sound produced by either the vocal folds (pseudo-periodic vibration found in voiced speech) or the turbulent flow of air through constrictions along the vocal tract (broad-band a-periodic noise-like signal found in unvoiced speech). The articulation capabilities of the vocal and nasal tracts reduce or enhance the frequency contents of the resulting sound, which is perceived by the human auditory system as a flowing stream of stimuli

* Corresponding author. Tel.: +34 91 336 73 84; fax: +34 91 336 66 01.
E-mail address: pedro@pino.datsi.fi.upm.es (P. Gómez-Vilda).

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.07.023

distributed accordingly with the dominant frequencies present in it. An injection of complex spike-like neural stimuli is released from the cochlea to the auditory nerve fibres [1] which are then processed at the level of the Brain Stem and distributed to the auditory primary and secondary areas over the cortex. Speech perception is a complex process which results as a combination of different pattern recognition tasks carried out by neural structures hidden in these areas.

Two important observations may be highlighted in speech perception: That speech sounds are dominated by certain enhanced bands of frequencies called *formants* in a broad sense, and that the assignment of meaning is derived both from dominant frequency combinations as well as from the dynamic changes observed in these combinations in time. Therefore speech perception can be seen as a complex parsing problem of time–frequency features. The most meaningful formants in message coding are the first two, designated classically as f_1 and f_2 in order of increasing frequency. f_1 is the lowest, which for male voice may roughly lay in the range of 250–700 Hz, whilst f_2 sweeps a wider range, from 700 to 2300 Hz. To serve as a self-explaining example, as the present study is focussed on the dynamic features of speech, Fig. 1 shows the spectrogram of a voiced speech frame with rapid formant changes. Formants are characterized in this spectrogram by brighter bands, whereas the darker areas indicate energy valleys. Peak bands are especially interesting because they can be associated with formants or resonances of the articulation structures, and the perceptual processing of the auditory pathways work detecting dominant frequencies related to formants. What can be observed in the figure is that the first formant is oscillating between 350 and 650 Hz, whereas the second formant experiences abrupt fluctuations between 700 and 2200 Hz. Higher positions of the second formant point to front vowel-like sounds, as [ε, *i*, *j*], whereas low ones correspond to back vowel-like sounds as [*u*, ω]. The positions of [ε, *i*, *a*, *u*] correspond to the zones where the formant positions are stable or slightly changing, as around the peaks of f_2 (segments in time: 0.15–0.17, 0.45–0.50, 0.7–0.75, 0.85–0.95, 1.15–1.17, 1.25–1.30, 1.38–1.42) whereas the positions of [*j*, ω] correspond to the complementary intervals where strong dynamic changes of formant positions can be observed. When plotting f_2 vs f_1 formant trajectories appear as clouds of dots showing the vowel-like structure of the message. The vertices mark the positions of the extreme front [*i*], back [*u*] and middle [*a*] vowels. Stable positions produce clouds of dots where formant plots are denser, whereas dynamic or changing positions produce thin trajectories, appreciated in the figure as bead-like lines. Formant transitions from stable characteristic frequencies (CF) to new CF positions (or *virtual loci* [29]) are known as frequency modulation (FM) components.

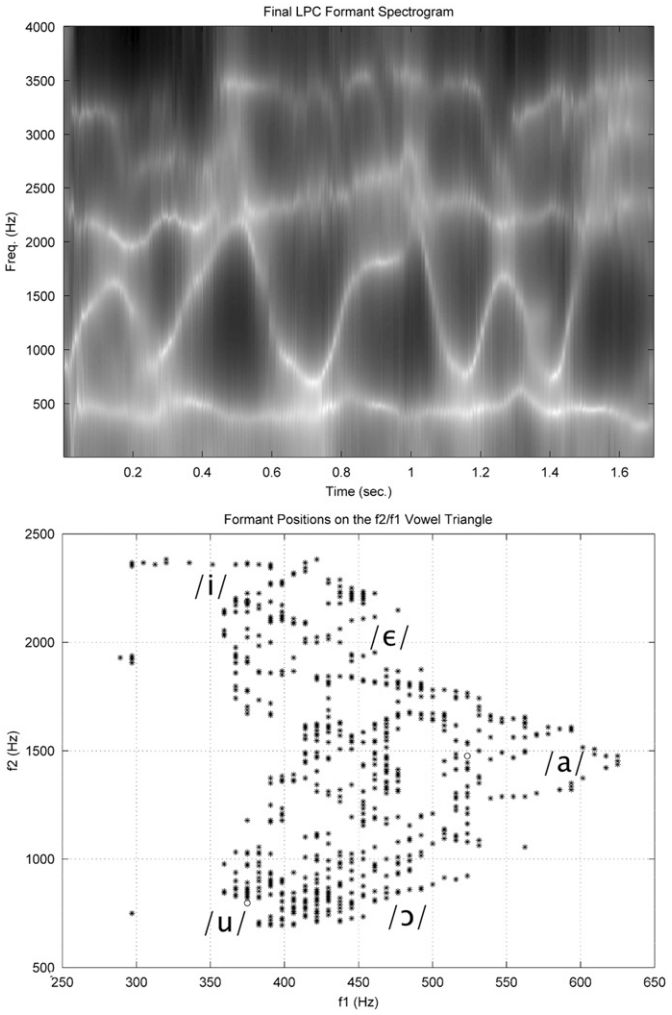

Fig. 1. Top: Adaptive Lineal Prediction (ALP) Spectrogram corresponding to the speech frame */Where were you while you were away/*, phonetically described as [$^{h}\omega\varepsilon^{\iota}\omega\varepsilon^{\iota}ju^{h}\omega aelju^{\omega}\varepsilon^{\iota}a\omega\varepsilon j$] uttered by a male speaker. The IPA has been used for annotation [3,4]. Bottom: Vowel triangle showing the five reference vowels in English framing the formant trajectories of the utterance.

3. Neuromorphic computing for speech processing

The structure responsible for speech perception is the auditory system, described in Fig. 2 as a chain of different sub-systems integrated by the peripheral auditory system (outer, middle and inner ear) and the higher auditory centres. The most important organ of the peripheral auditory system is the cochlea (inner ear), which carries out the separation in frequency and time of the different components of sound and their transduction from mechanical to neural activity [1]. Electrical impulses propagate from the cochlea to higher neural centres through auditory nerve fibres with different characteristic frequencies (CF) responding to the spectral components (or harmonics $f_0, f_1, f_2 \ldots$) of speech. Within the cochlear nucleus (CN) different types of neurons are specialized in specific processing [15] as described below. The cochlear nucleus feeds information to the Olivar Complex, where sound localization is derived from inter-aural differences, and to the inferior colliculus (IC) organized in spherical layers with orthogonal iso-frequency bands. Delay lines are found in this structure to detect temporal features in acoustic signals, which are of especial relevance for this study as will be seen in the sequel. The thalamus (Medial Geniculate Body) acts as a last relay station, and as a tono-topic mapper of information to the primary auditory cortex as ordered feature maps.

The functionality of the different types of neurons found in the auditory pathways is the following:

- Pl: Primary-like units. Reproduce the firing stream found at its input, acting as relay stages.
- On: Onset units. Detect the leading edge of a new firing stream, and separate the background activity from a new stimulus activity.
- Ch: Chopper units. Specialized in dividing a continuous stimulus into slices of different size.
- Pb: Pauser units. Act as delay lines, firing sometime after the stimulus onset.
- CF: Characteristic frequency units. Respond to different narrow bands of frequencies centred in a specific one and are tono-topically organized.
- FM: Frequency modulation units. Specialized in detecting changes in frequency. Their role is crucial in detecting dynamic speech features.

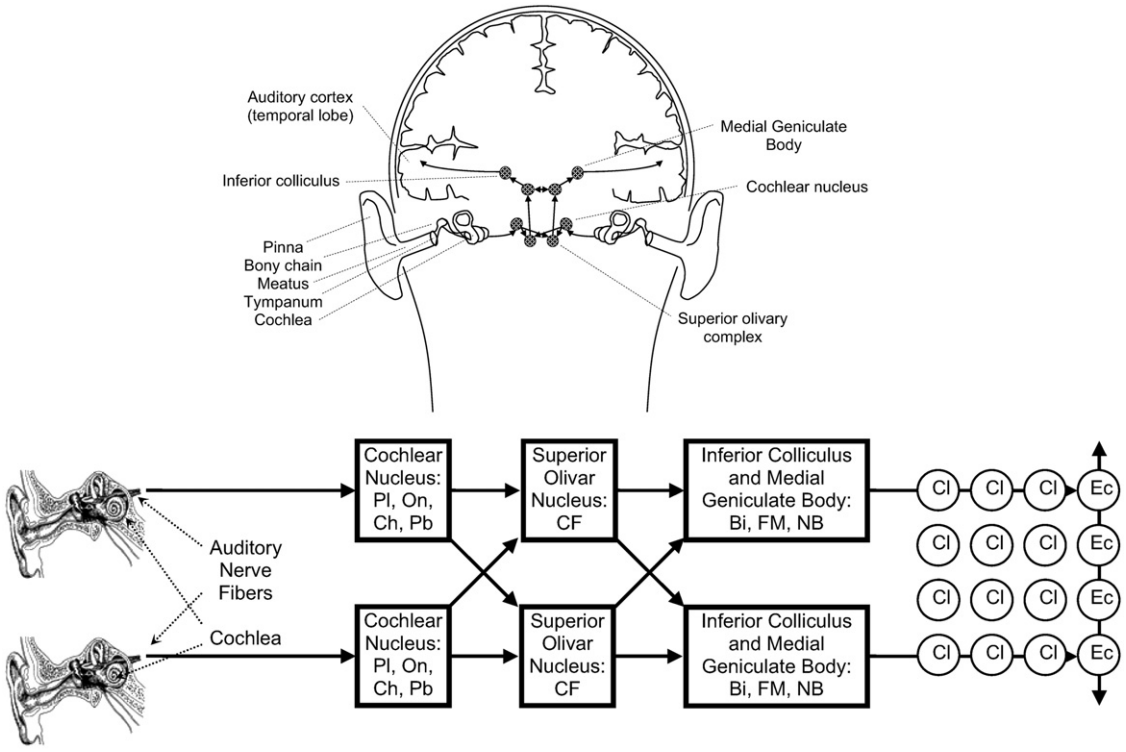

Fig. 2. Speech Perception Model. Top: main auditory pathways in the Peripheral and Central Auditory Centres (adapted from [8]). Bottom: simplified main structures found in the Auditory Pathways.

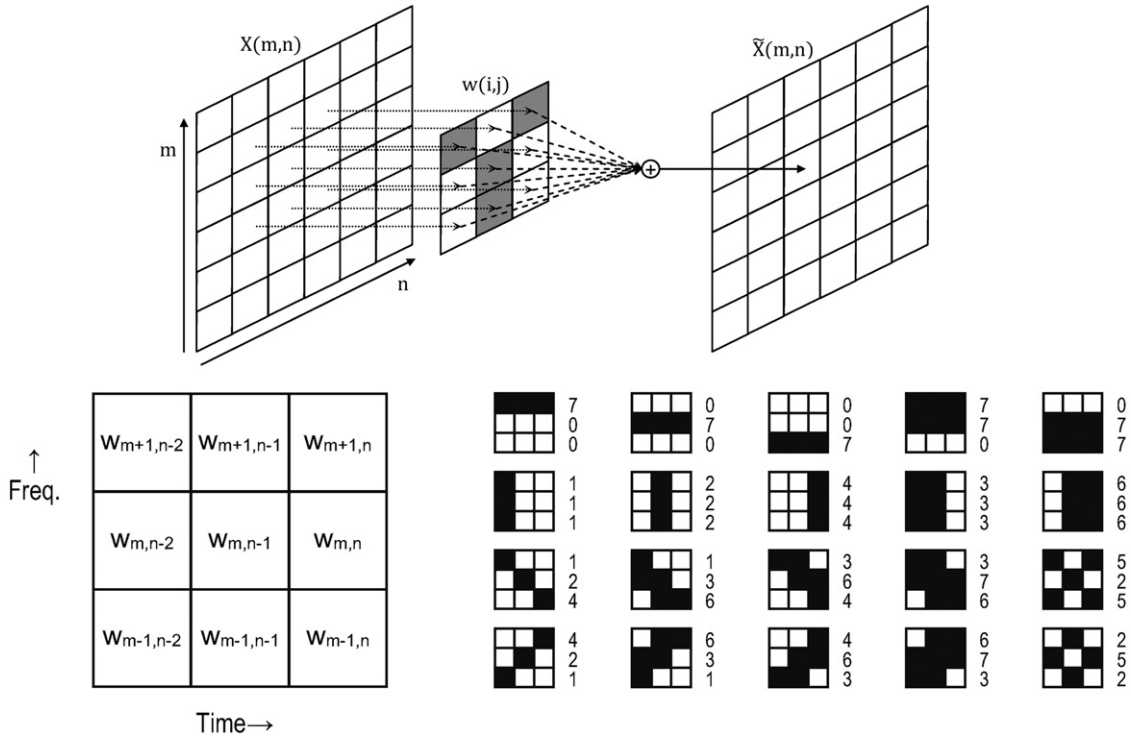

Fig. 3. Mask-based Neuromorphic Computing Units. Top: structure of a general unit. Bottom: 3×3 masks for feature detection on the formant spectrogram. Each mask is labelled with the corresponding octal code (most significant bits: bottom-right).

- NB: Noise burst units. React to broad-band stimuli, as those found in unvoiced consonants.
- Bi: Binaural units. Specific of binaural hearing by contrasting phase-shifted stimuli. They are found mainly in the inferior colliculus.
- Cl: Columnar units. Organized linearly in narrow columns through the layers of the auditory cortex. Their function may be related with short-time memory [19], although their role is to be further clarified.

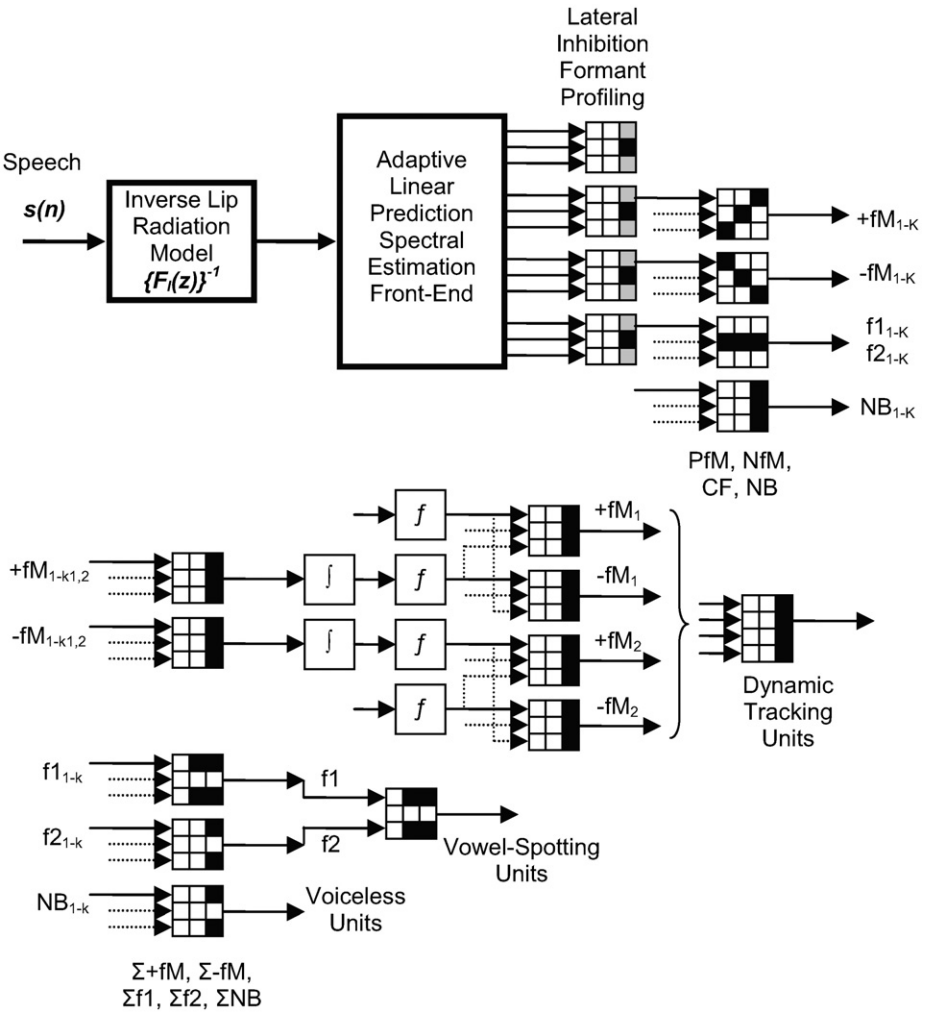

Fig. 4. Neuromorphic Speech Processing Architecture for a mono-aural channel. Each neuron is implemented as a GNCU of the sort given in Fig. 3, represented by its mask. Blocks labelled as ($\int$) and (f) stand for integrators and nonlinear thresholds.

Table 1
Nuclear set of consonant phonemes and some of their associated phonetic features. IPA: International Phonetic Alphabet. KB: Kirshenbaum Code. O/N: oral vs nasal. V/U: voiced versus unvoiced. DC: degree of closure (s: stop, f: fricative). AP: articulation place (bl: bilabial, a: alveolar, p: palatal, v- velar, ld: labiodental, d: dental, da: dentoalveolar, pa: palatoalveolar). O/R: oval vs round. FM1: dynamics of the first formant (a: ascending, d:descending). FM2: Idem of the second formant. CF: stable formants. NB: noise bursts (s: spread, mh: medium-high, h: high, ml: medium-low.

IPA	p	t	c	k	b	d	ĵ	g	f	θ	∫	x	β	ð	ζ	γ
KB	p	t	c	k	b	d	J	g	f	T	S	x	B	D	Z	G
O/N	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
V/U	u	u	u	u	v	v	v	v	u	u	u	u	v	v	v	v
DC	s	s	s	s	s	s	s	s	f	f	f	f	f	f	f	f
AP	bl	a	p	v	bl	a	p	v	ld	d	p	v	bl	da	pa	v
O/R	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
FM1	a	a	a	n	a	a	a	n	a	a	a	n	a	a	a	n
FM2	a	n	d	n	a	n	d	n	a	n	d	n	a	n	d	n
CF	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
NB	s	mh	h	ml	–	–	–	–	s	mh	h	ml	–	–	–	–

- Ec: Extensive connectors. The outer layers of the auditory cortex seem dominated by extensive connections among distant columns.

The present study is aimed to simulate some of the functionality of the auditory system to detect speech dynamics; therefore, time–frequency representations have to be taken as a source. Some possibilities are found in the literature to produce these representations, as filter banks, gammatones, FFT or LPC, among others.

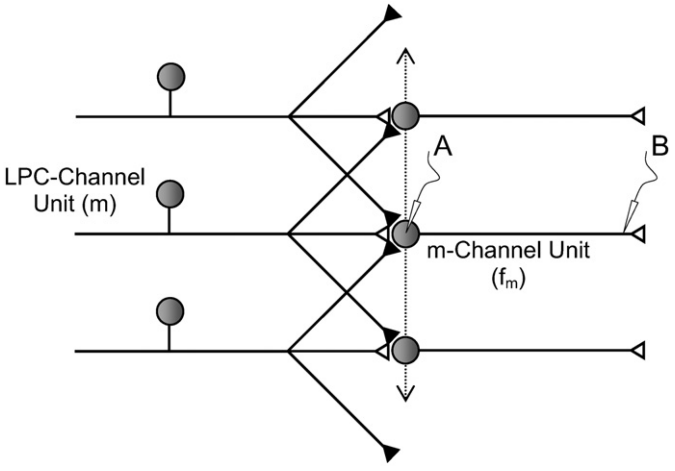

Fig. 5. Formant profiling from LPC broad-band spectrogram in Fig. 1 (top) by lateral inhibition units. Structure of feed-forward inhibition units and connections. Dark synapses mean inhibition, white ones stand for excitation. Pre-threshold (A) and post-threshold (B) spots have been clipped to monitor results.

The present work is based in formant-like pattern detection on LPC spectrograms [6] as the one in Fig. 1. One such spectrogram may be seen as a set of $1 \le m \le M$ frequency channels in time giving the envelope of the power spectral density of speech as

$$X(m,n) = 20\log_{10}\left|1-\sum_{p=1}^{P} a_{k,n}e^{-jmp\Omega\tau}\right|^{-1} \quad (1)$$

where $\{a_{k,n}\}$ is the coefficient set of a P-order predictor estimated at the time instant n from a speech signal $x(n)$, and τ and Ω are the resolutions in time and frequency. One of the m frequency-separated channels may be seen as the timely activity of an auditory fibre associated to a characteristic frequency f_m. The matrix $X(m,n)$ can be seen as a two-dimensional auditory image [18], describing the activity in time of a linear layer of CF units in frequency. Many tools devised for image processing can be used for the detection of time–frequency features, as CF or FM patterns [13], as for instance simple masks $w(i,j)$ operating on the auditory image as

$$\tilde{X}(m,n) = \sum_{i=-I}^{I}\sum_{j=0}^{J} w_{i,j}X(m-i,n-j) \quad (2)$$

where $\{w_{ij}\}$ is a $(2I+1)x(J+1)$ mask with a specific set of weights defined to mimic a specific function. The activity of this matrix may be represented by a Generalized Neuromorphic Computing Unit as the one shown in Fig. 3, where the output activity $\tilde{X}(m,n)$ is the result of applying a weight mask $w(i,j)$ to the inputs from $X(m,n)$, adding or subtracting the incoming stimuli (depending on their excitatory or inhibitory nature coded in the specific weight) and applying a threshold nonlinear function. These outputs will constitute a new layer of channels, where m is now a positional channel index (unit number) and n is the time index.

The lateral-inhibition filtering active in certain neuron associations in the inferior colliculus may be seen as a special case of (2) where the weights of columns $j=1,2$ are zeros and weights $w_{-1,0}=w_{1,0}=-1/2$ and $w_{0,0}=1$.

Different neurons as the one defined in Fig. 3 organized in consecutive layers will mimic some of the speech processes of interest in the study (specifically formant profiling by lateral inhibition, positive and negative formant tracking, and first and second formant disambiguation by mutual exclusion). In the bottom part of the figure some examples of 3×3 masks are shown. To produce un-biased 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 3×3 mask, respectively. It is important to remark that the basic structure and functionality of the Generalized Neuromorphic Computing Unit defined is specifically based on the Hebbian Neuron [12]. The Neuromorphic Speech Processing Architecture proposed for dynamic formant tracking based in these units may then be described by the structure presented in Fig. 4. This architecture is composed by different layers of specific GNCU's

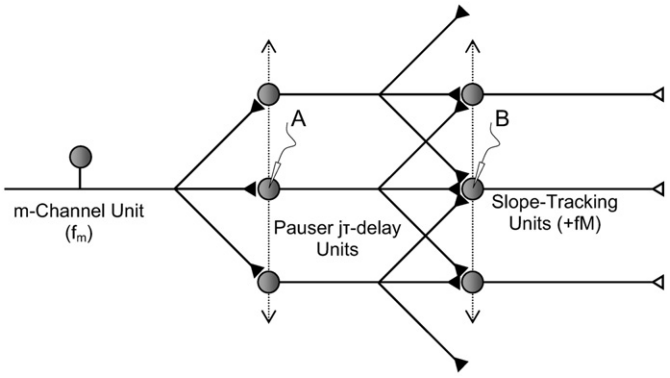

Fig. 7. Positive and Negative Slope Tracking Units. Pauser units (A) are activated by m-Channel Units. Pausers respond with a delay j time delay intervals (τ) different for each unit. These activate spatial summation units (B). The positive or negative slope-tracking capability of the unit is based on delay and channel configurations.

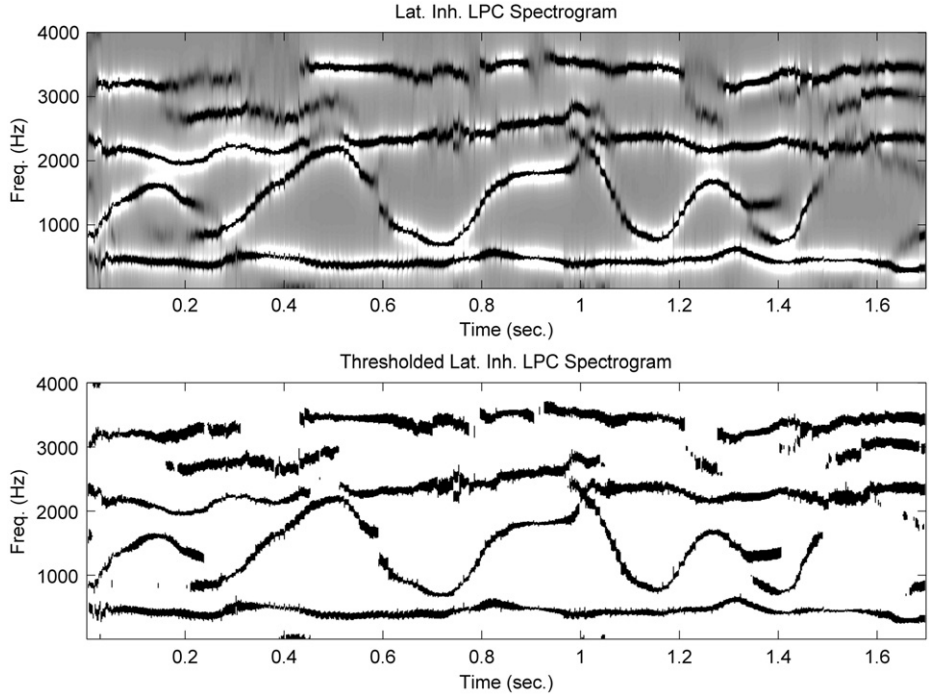

Fig. 6. Top: pre-threshold activity as hypothetically measured in (A). Bottom: post-threshold activity, as in (B). Compare the formant patterns produced against the input broad-band spectrum.

mimicking the physiological units found in the auditory pathways accordingly with the description given below as follows:

- LIFP: Lateral Inhibition Formant Profilers, reducing the number of fibres firing at a given time.
- $+f_{M1-k}$, $-f_{M1-k}$: Positive and Negative Slope Formant Trackers (K bands) detecting ascending or descending formant activity using masks {124–376} and {421–673}.
- $f1_{1-k}$, $f2_{1-k}$: First and Second Energy Peak Tracker, intended for formant detection mimicking CF neurons, using masks {700–077}.
- $+fM_{1-k1}$, $-fM_{1-k1}$, $+fM_{1-k2}$, $-fM_{1-k2}$: These are integrators or accumulators working on the inputs of previous formant tracker integration units on certain specific bands (350–650 Hz for the first formant, or 700–2300 Hz for the second formant).
- $+fM_1$, $-fM_1$, $+fM_2$, $-fM_2$: First and Second Formant Mutual Exclusion Units (positive and negative slopes), estimating the features $FM1$ and $FM2$ in Table 1.
- NB_{1-k}: Noise Burst Integration Units ({111–666}) for wide frequency activity working on the formant profiles.
- VSU: Voiceless Spotting Units. These integrate the outputs of different ΣNB's acting in separate bands to pattern the activity of fricative consonants.
- WSU: Vowel Spotting Units. These integrate the activity of $\Sigma f1$ and $\Sigma f2$ units to detect the presence of vowels and their nature.
- DTU: Dynamic Tracking Units. These integrate the activity of different dynamic trackers on the first two formants to detect consonant dynamic features.

4. Simulating FM Units

From what has been exposed a clear consequence may be derived: formant structure plays a major role in the vowel and consonantal structure of speech. Formant detection, tracking and grouping in semantic units must be a crucial role in speech understanding. Therefore the simulation of these functionalities by neural-like simple units may be of most importance for neuromorphic speech processing. In what follows some of the capabilities of these structures will be shown with emphasis in the detection of the most meaningful dynamic consonantal features. For such, some of the structures described in the Neuromorphic Speech Processing Architecture shown in Fig. 4 will be briefly reviewed and simulated with the aim of offering a brief review of the possibilities of this modelling, and the results obtained from their activity will be presented and discussed. These are the following:

- Lateral Inhibition Formant Profiling Units
- Positive Slope Formant-Tracking Units ($+f_{M1-k}$)
- Negative Slope Formant-Tracking Units ($-f_{M1-k}$)
- First-Formant Positive Detection Units ($+fM_1$)
- First-Formant Negative Detection Units ($-fM_1$)
- Second-Formant Positive Detection Units ($+fM_2$)
- Second-Formant Negative Detection Units ($-fM_2$)

As target speech a typical example illustrating fast formant dynamics as is the sentence – *Where were you while you were away* – will be modelled. The details of the architecture are the following: $K=512$ units are used as characteristic frequency outputs from LPC, defining a resolution in frequency of little less than 8 Hz for a sampling frequency of 8000 Hz. These 512 channels are sampled each 5 ms to define a stream of approximately 200 pulses/s per each.

4.1. Lateral Inhibition Formant Profiling Units

The first neuromorphic signal processing task simulated is formant profiling from the LPC broad-band spectrogram as shown in Fig. 5. In the figure a possible layered structure is represented where the activity expressed by Channel Units excite an output m-Channel Unit, and inhibit the neighbour ones. The results of sweeping the LPC spectrogram in Fig. 1 (top) with one such layer

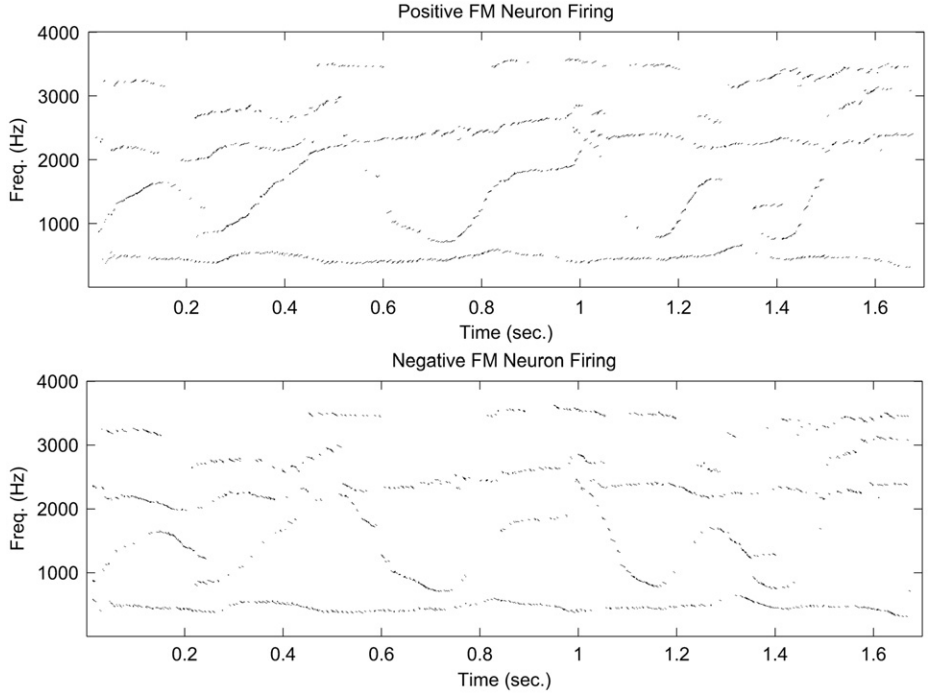

Fig. 8. Positive and Negative Slope Tracking Units. Top: activity of $+fM_{1-512}$ units detecting upwards formant trajectories. Bottom: activity of $-fM_{1-512}$ units for downwards formant trajectories.

produces the results shown in Fig. 6, where the pre-threshold (A: top) and post-threshold activity (B: Bottom) are presented. The pre-threshold activity shows the typical "Mexican Hat" behaviour.

The resource to lateral inhibition is a strategy well documented in natural neural systems [27], fulfilling several purposes. The transition from time–frequency detailed spatiotemporal structure of the responses of the auditory nerve to specific CF/CF and FM/FM responses found in the primary auditory cortex (AI) of the moustached bat by Suga [28] indicates that some powerful mechanism is applied to reduce spike firing rates and the number of fibres conveying information to higher auditory centres.

It is believed that the mechanism for this compression coding process is based in "specific lateral inhibition networks which may exist in the antero-ventral cochlear nucleus (CN), especially involving T-Stellate cells, which exhibit fast inhibitory surrounds and a robust representation of the input spectrum regardless of level" [25]. This belief is also supported by the strong reduction in spike firing rates found in the lower levels of the auditory pathways as compared with the firing rates in the AI areas, which suggest the presence of a strong compression mechanism both in the time and in the frequency domain [11].

4.2. Positive and Negative Slope Formant-Tracking Units

The Positive and Negative Slope Formant Trackers detecting ascending or descending formants by masks {124–376} and {421–673} in Fig. 3 (bottom) correspond to the cell columns to the uppermost right-hand side of Fig. 4, labelled as $+fM_{1-k}$ and $-fM_{1-k}$, where k is the respective order of the frequency bin bands being searched, and the sign + or – refers to the positive or negative sense of the slope. In the specific case shown in simulations throughout the paper the dimensions of the $+fM$ and $-fM$ units are 7×7, which means that the connectivity in frequency extends from +3 to −3 neighbour neurons, whilst the delay lines in the Pauser units responsible for the delay go from 0 to 30 ms, as 5 ms is the delay unit (corresponding roughly to a maximum firing rate of 200 spikes/s). Fig. 7 shows a possible morphology of the delay and detecting units, and their outputs to the same speech fragment considered above. Other techniques being studied for the determination of the mask coefficients are back-propagation NN's, although the results presented here are for pre-determined (firmware) masks.

The resulting activity as detected per each of the 512 channel units is given in Fig. 8. It may be seen that the strong activity compression produced from lateral inhibition results in a few units firing simultaneously at a given time instant.

In the general outcome, it may be said that the units detect the main episodes of formant ascent and descent with enough accuracy, although a certain amount of noisy artefacts may be present due to the glittering nature of formant detection in itself. Nevertheless these problems can be solved easily by massive integration (averaging) and thresholding, as will be seen in the sequel.

4.3. First/Second-Formant Positive/Negative Detection Units

The structure and operation of positive and negative formant slope detectors as the ones summarized in the middle level of Fig. 4 (Dynamic Tracking Units) will be discussed here and some results shown. Formant theory of speech perception is mainly based on psychophysical grounds. Its plausibility comes from the facts that vowel structures play the role of two-frequency robust primitive communication codes [7]. Therefore resources to distinguish vowel from non-vowel pitched sounds must be available at the level of auditory interpretation centres located in the auditory cortex. These must be linked to the structures providing information about formant displacements or trajectories when dealing with dynamic consonantal sounds of speech with share similarities with the detection of FM sweeps in the bat's auditory system [28]. As the possibilities are both for ascending or descending first and second formants, at least four different types of formant slope tracking units should be hypothesized. The real existence and the number of these structures present in the human auditory cortex remains as a question put forth to neurophysiologists [23]. In Fig. 9 the structure of two of such units interlocked for mutual exclusion is depicted.

As the frequency distribution is linear, the number of channels integrated for the first formant (for a band of 300–700 Hz) is around 52, whereas for the second formant (for a band of 700–2300 Hz) is around 205. The study of nonlinear (logarithmic) distributions based on Mel-scaling is left for future research. Conceptually formant ascent and descent are mutually excluding, therefore mutual exclusion mechanisms should be implemented through lateral inhibition. This is provided by the inter-locking inhibitory synapses running from each axon's output (B and D) to the bodies of the counterpart unit ($\sim$D to A and $\sim$B to C). In the simulations it is assumed that the stronger output inhibits the weaker.

In this way inconsistencies are removed from the resulting firing activity shown in the templates from Figs. 10–13.

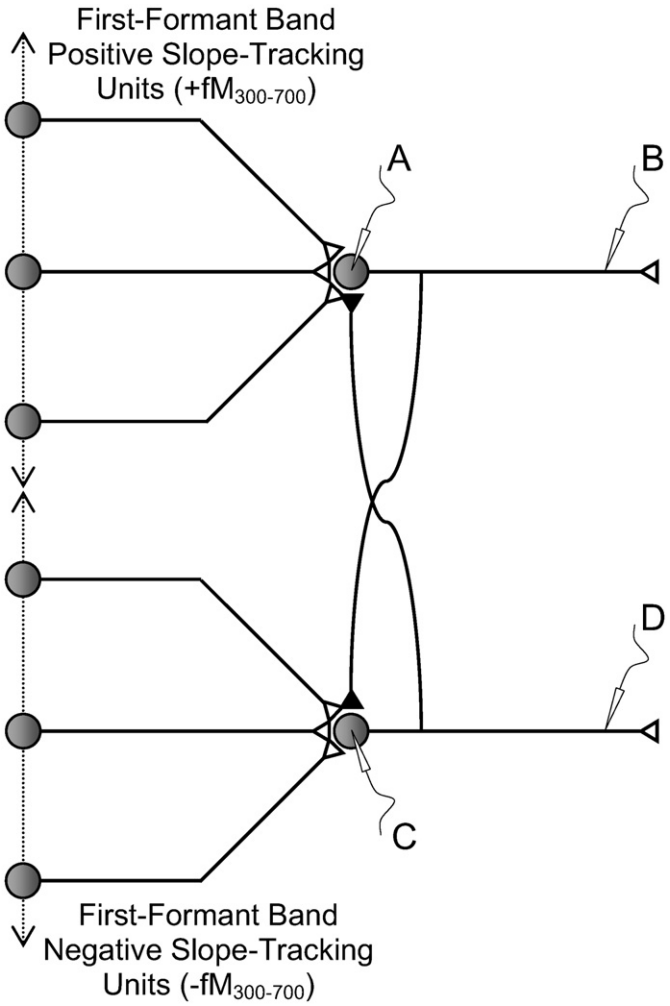

Fig. 9. Structure of $+f_{M300-700}$ and $-f_{M300-700}$ units coding the activity of the first formant f_1. A similar structure may be hypothesized as well for the second formant ($+f_{M700-2300}$ and $-f_{M700-2300}$). Dark synapses mean inhibition, white ones stand for excitation. The patterns detected at each of the spots (A, B, C, D) is given in Figs. 10 and 11, respectively.

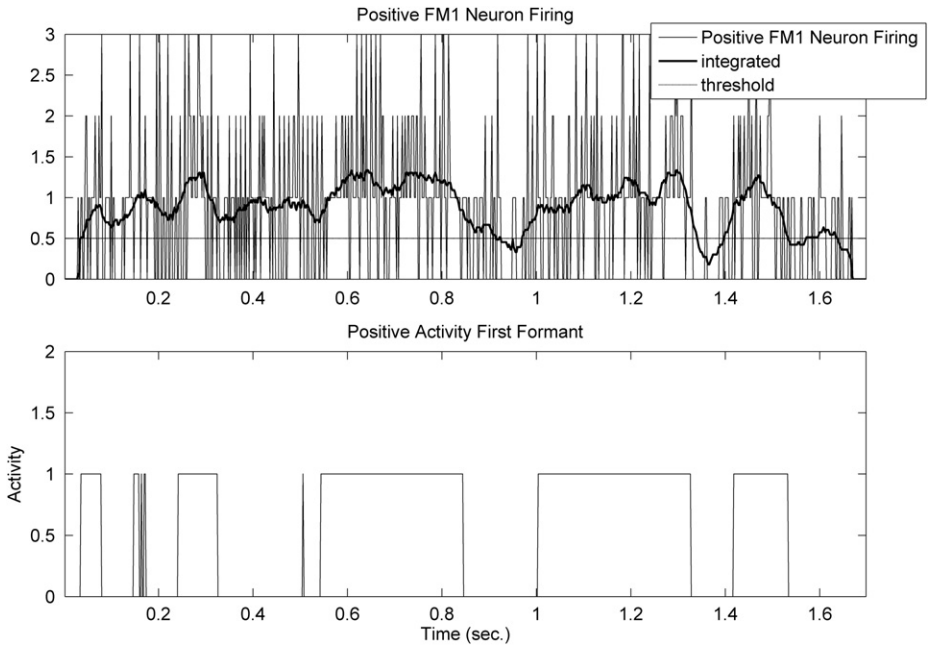

Fig. 10. Top: firing activity accumulated at the input (A) of the First-Formant Positive-Slope Unit (thin spiky pattern). Integration of the firing activity (B) at the input (bold line). The threshold is given as a reference. Bottom: activity of the First Formant Positive-Slope Integration Unit $+fM_1$ showing the time intervals where the first formant ascends.

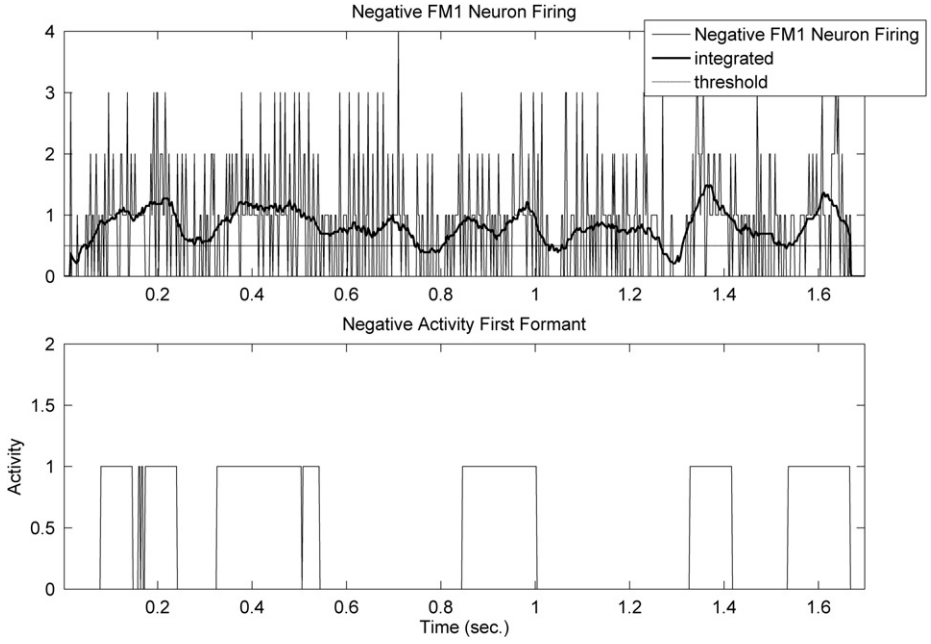

Fig. 11. Top: firing activity accumulated at the input (C) of the First-Formant Negative-Slope Unit (thin spiky pattern). Integration of the firing activity (D) at the input (bold line). The threshold is given as a reference. Bottom: activity of the First Formant Negative-Slope Integration Unit $-fM_1$ showing the time intervals where the first formant descends.

The top part of Fig. 10 shows the activity present at the input of the First-Formant Positive-Slope Tracking Unit as provided by 52 synaptic connections coming out from the Positive-Slope Tracking Units in the band 300–700 Hz, which corresponds to the band of frequencies where the average first formant can be found.

It may be seen that barely two or three of these synapses may be firing at a time. The Unit is based in the Mculloch-Pitts paradigm [17] including integration and threshold functionalities, symbolized in Fig. 4 as ($\int$), producing the output line in bold. When its value jumps over the threshold (f) the output (B) is activated high.

The correspondence between (A) jumping over the threshold and the activation of (B) is not straightforward, as the activation (C) of the First-Formant Negative-Slope Tracking Unit in Fig. 11 has to be taken also into account, because it will be trying to inhibit the Positive-Slope Tracking Unit at the same time. As a result, both (C) and (D) outputs will mark intervals where either one or the other output will be active, or both of them will remain inactive (when the formant remains stable, as in certain vowels). A similar structure activated with synapses in the band 700–2300 Hz must be built for the detection of second formant dynamics (not shown). The results produced by one such structure are given

in Figs. 12 and 13. In this case the input activity detected in the Unit soma is a little bit larger, as in some instants up to four synapses are active at a time. Nevertheless, the output will only be activated when the accumulated stimuli (integrated with a certain forgetting factor) jump over the threshold (signalled by a horizontal line). As before, the output activity of both interlocked Units does not correspond strictly to their inputs, as the mutual exclusion mechanism excludes the possibility of simultaneous positive outputs. A general criticism to this formant-oriented detection strategy is that formants are not so neatly separated by a given sharp boundary around 700 Hz in the cortex as neatly exposed here. It is important to consider that formant grouping may be carried out by large sets of spatial addition units picking-up information more or less randomly from Slope Tracking Units and interacting among them in some winner-takes-all kind of contests, this being part of the Speech Understanding Adaptation during Language Acquisition taking place in early infancy. Anyway, the intention of the research, which is essentially to show that small neuromorphic structures involving simple and biologically plausible resources can cope with the task, is more than fulfilled. Other similar structures could do the job as well for higher formants, although this aspect is not as neatly related to speech understanding. Its existence could be more in connection with other psychophysical listeners' abilities, as the capability to track the speaker's identity [24].

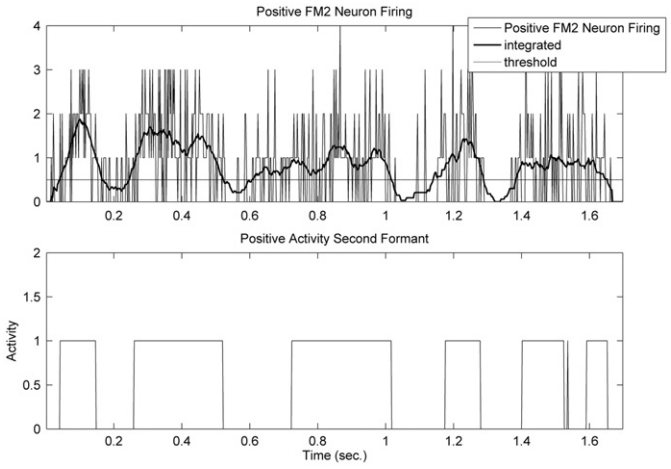

Fig. 12. Top: firing activity accumulated at the input of the Second-Formant Positive-Slope Unit (thin spiky pattern). Integration of the firing activity at the input (bold line). The threshold is given as a reference. Bottom: activity of the Second Formant Positive-Slope Integration Unit $+fM_2$ showing the time intervals where the second formant ascends.

4.4. Application to Neuromorphic Phonetic Labelling

An example on how specifically Speech Processing may benefit from Neuromorphic Computing will be given in the present section. Phonetic Labelling is a technique consisting in highlighting or spotting specific segments of speech accordingly with some property, as the presence of voicing, nasality, or even spotting vowels, specific phonemes and even words. It is very useful for certain applications as speech annotation, audio and video diarization, or forensic studies, among others. In phonetic labelling features as the ones referred in Table 1 are used as referencing marks for spotting. The first row of the table shows the IPA code of the corresponding phoneme, whereas the second row gives the corresponding ASCII-IPA equivalent, also known as *Kirshenbaum code* [3,4].

The specific example studied as a working case in the present paper was selected for the spotting of dynamic consonants and approximants as [*j*, ω], which have been set as targets within the speech frame used. The dynamic descriptors for the descending glides as found in [$\omega\varepsilon$] and [*ju*] are, respectively, NFM1 = '0', PFM1 = '1', NFM2 = '0', PFM2 = '1' and NFM1 = '0', PFM1 = '0', NFM2 = '1', PFM2 = '0'. The complete outcome to the LPC spectrogram in Fig. 1 (input) reproduced in the top part of Fig. 14 is given as the four outputs labelling the first and second formant ascents and descents as in Fig. 14 middle and bottom templates. It may be seen that the first and second negative and positive slope tracking outputs (NFM1, PFM1, NFM2, PFM2) overlap almost perfectly as complementary signals (when one is high its complementary is

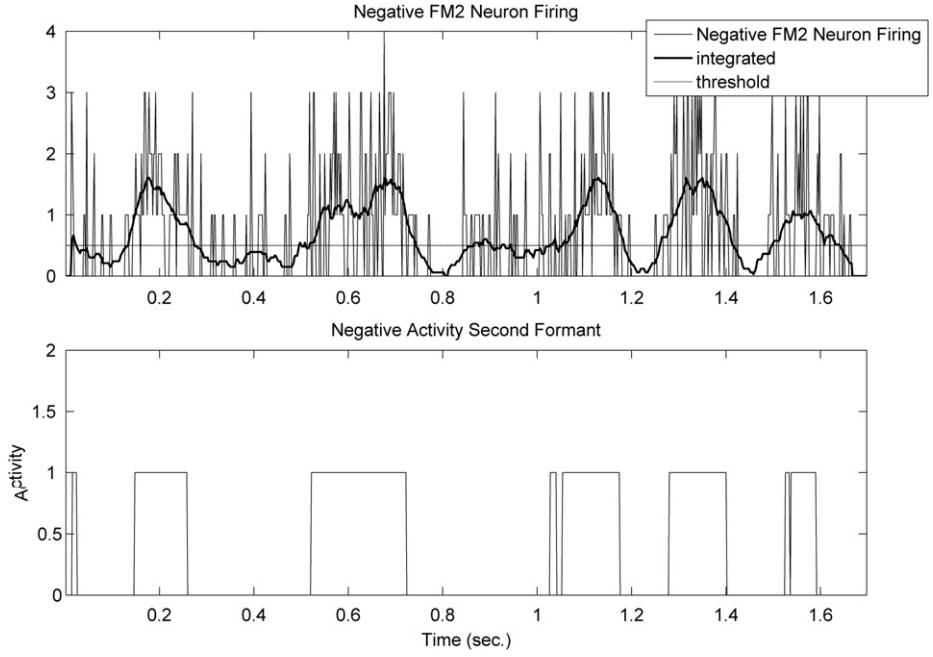

Fig. 13. Top: firing activity accumulated at the input of the Second-Formant Negative-Slope Unit (thin spiky pattern). Integration of the firing activity at the input (bold line). The threshold is given as a reference. Bottom: Activity of Second Formant Negative-Slope Integration Unit $-fM_2$ showing the time intervals where the second formant descends.

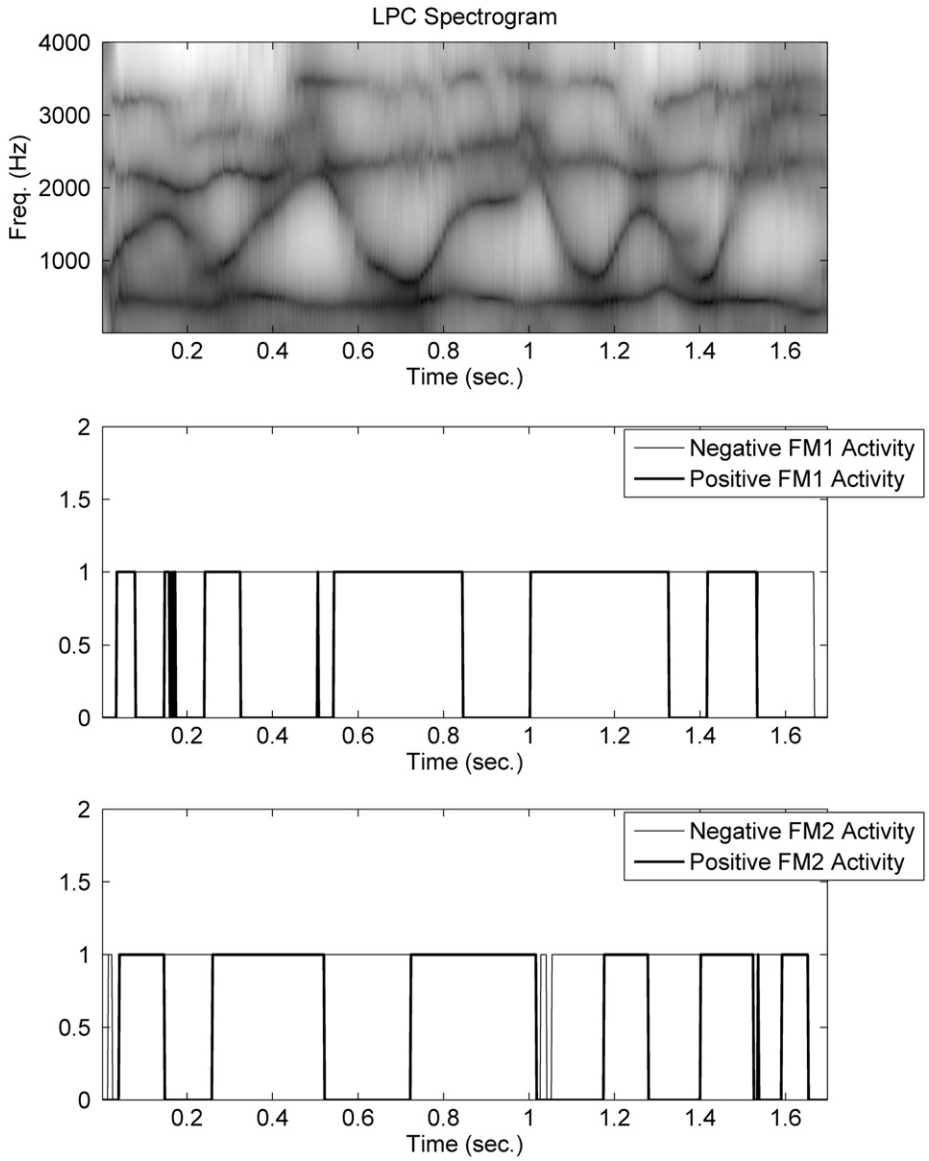

Fig. 14. Top: output of activity at the output of LIFP Units in the band of the second formant. Middle: output of the Second Formant Integration Unit *+fM2* reproducing the positive slope intervals in the second formant. Bottom: value of the slope detected on the given interval.

down, and vice versa). These four signals are used to designate the four possible states of rows FM1 and FM2 of Table 1. For instance, a situation where FM1 = '*a*' and FM2 = '*d*' as is the case in voiced phonemes /ĵ/, and /ζ/ would be signalled by NFM1 = '0', PFM1 = '1', NFM2 = '1', PFM2 = '0'. Specifically, for the speech frame being labelled, the presence of the phoneme [ω^e] is spotted by the combination PFM1 = '1' and PFM2 = '1', appearing in the intervals (0.04–0.09, 0.24–0.32, 0.72–0.84, 1.15–1.18 and 1.42–1.52). The reader may check that this is precisely the number of times the phonetic pattern targeted appears in the reference speech frame. These results show the way of implementing the paradigms in Table 1. The utility of this methodology is to be found in the automatic phonetic labelling of the speech trace, as shown in this study, as well as in typical tasks related with Cognitive Audio Processing [20].

5. Discussion and conclusions

Through the present paper it has been shown that formant-based speech processing may be carried out by well-known bio-inspired computing units. Special emphasis has been placed in the description of the biophysical mechanisms which are credited for being responsible of formant dynamics detection, as related to the perception of certain consonantal sounds.

A special effort has been devoted to the definition of a plausible neuromorphic or bio-inspired architecture composed of multiple modules of a general purpose computing unit. The use of such units in consonantal formant dynamics characterization as positive and negative frequency tracking and grouping has also been presented. The structures studied correspond roughly to the processing centres in the Olivar Nucleus and the inferior colliculus. The systemic bottom–up building of layered structures reproducing dynamic feature detection related to plausible neuronal circuits in the auditory cortex has also been introduced. Results from simulations explaining the behaviour of these layered structures have been presented as well, confirming that robust formant trackers built from simple Hebbian units may carry out important tasks in speech processing eventually related with the perception of dynamic consonants. This work may help in both understanding better how neural circuits may work in the brain, as well as in how speech processing can benefit from this understanding. But this is only a first step in the progress towards a

systemic comprehension of what is language processing in the human brain [10,11,20,21,25,26]. The study of short-time memory-like structures found in the upper levels of the brain, and especially the columnar structures of the auditory cortex [16] using low order regressors is fundamental for phonemic parsing and deserve an extensive further attention. The lower and mid auditory pathways have been intensively and extensively researched, and a good deal of helpful and useful knowledge of use in Neuromorphic Speech Processing has been produced [21]. Nevertheless the cortical circuits involved in speech processing lack a similar detailed description, their functionality being a great challenge even nowadays. Through the works of Mountcastle [19] some kind of functionality could be inferred related mainly with the short memory capability of bidirectional linear structures reproducing cortical columns, essential for the parsing of phonetic sounds leading to the emergence of the word as a semantic unit. This idea was discussed during the celebration of a colloquium directed by Prof. Mira with some other colleagues in IWINAC07 at La Manga del Mar Menor (Spain) [8]. In that occasion Prof. Mira insisted in that the leading work of Rafael Lorente de No [16] should be revisited in relation with the topic of short memory phonetic parsing in the auditory cortex. Lorente de No was one of the most outstanding disciples of Santiago Ramón y Cajal [22], and lived and worked in the USA for some five decades till his death in Tucson, AZ in 1990. It seems that Lorente de No's work eventually inspired that of Mouncastle. During two other occasions in the early spring of 2008 Prof. Mira insisted in that Lorente de No's should be brought forth again to the interest of modern Neuromorphic Computing for the proposal of new mechanisms in speech processing and understanding. Shortly after, Prof. Mira passed away (August 2008) leaving this new challenge open for the attention of the neuromorphic speech research community. Since then, new steps have been given forward in pursuing a better comprehension on how speech is processed in the higher auditory paths. The ambitious Cajal-Blue Brain [5] has started its first steps through a collaboration programme between Instituto Ramón y Cajal and Universidad Politécnica de Madrid. One of its objectives is to get a better description of the neuronal structures by reverse engineering [2] which can eventually help in shedding new light on how neuronal circuits work in certain specific tasks. Sound and Image Processing are amongst the most challenging ones which may benefit from this long-run initiative. The programme to be covered is aimed to disentangle neural structures, explain biophysical and biochemical interactions, yield pace to systemic abstractions, and on a way-back, build-up a complete neuro-inspired structure explaining most of the functionalities observed both under the biophysical and the psychophysical points of view. The task is cumbersome and resource expensive, as must rely on high performance computing due to the complexity and low-level description required in many of the tasks. But the rewards can be enormous, as not only systemic behaviour is sought. A better understanding of simple and complex brain structures is expected both in behavioral and functional terms. This understanding may have a direct impact in developing advanced helps for the deaf, blind and sensory-motor impaired, improve speech and speaker recognition, language acquisition, and many others. This is a task in which early predecessors are to be acknowledged. In memoriam: Cajal, Lorente de No, Mira.

Acknowledgements

This work is being funded by grants TEC2006-12887-C02-01/02 and TEC-2009-14123-C04-03 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] J.B. Allen, Nonlinear cochlear signal processing and masking in speech perception, in: J. Benesty, M.M. Sondhi, Y. Huang (Eds.), Springer Handbook of Speech Processing, Springer Verlag, Berlin2008, pp. 27–60 (Chapter 3).
[2] J.I. Arellano, R. Benavides-Piccione, J. DeFelipe, R. Yuste, Ultrastructure of dendritic spines: correlation between synaptic and spine morphologies, Frontiers in Neuroscience 1–1 (2007) 131–143.
[3] Available from 〈http://www.arts.gla.ac.uk/IPA/ipachart.html〉.
[4] Available from 〈http://www.kirshenbaum.net/IPA/ascii-ipa.pdf〉.
[5] Available from 〈http://cajalbbp.cesvima.upm.es/〉.
[6] J.R. Deller, J.G. Proakis, J.H.L. Hansen, Discrete-Time Processing of Speech Signals, Macmillan, NewYork, 1993.
[7] D.B. Geissler, G. Ehret, Time-critical integration of formants for perception of communication calls in mice, Proceedings of the National Academy of Science 99-13 (2002) 9021–9025 B. Goldstein, Sensation and Perception, Wadsworth, Belmont CA, 1984.
[8] P. Gómez, J.M. Ferrández, V. Rodellar, A. Álvarez, L.M. Mazaira, A. Bio-inspired, Architecture for cognitive audio, Lecture Notes on Computer Science 4527 (2007) 132–142.
[9] P. Gómez, J.M. Ferrández, V. Rodellar, R. Fernández, Time–frequency representations in speech perception, Neurocomputing 72 (2009) 820–830.
[10] 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, 2006, pp. 3–17.
[11] S. Greenberg, W.H. Ainsworth, Speech processing in the auditory system: an overview, in: W.A.S. Greenberg (Ed.), Speech Processing in the Auditory System, Springer, New York, 2004, pp. 1–62.
[12] D.O. Hebb, in: The Organization of Behavior, Wiley Interscience, New York, 1949 (reprinted 2002).
[13] B. Jähne, Digital Image Processing, Springer, Berlin, 2005.
[14] E.R. Kandel (Ed.), Principles of Neural Science, McGraw-Hill, New York, 2000.
[15] A. Krishnan, Y. Xu, J. Grandour, P. Cariani, Encoding pitch in human brainstem is sensitive to language experience, Cognitive Brain Research 25 (2005) 165–168.
[16] R. Lorente de No, Cerebral cortex: architecture, intracortical connections, motor projections, in: J.F. Fulton (Ed.), Physiology of the nervous system, 3rd Edn, Oxford University Press, 1949, pp. 288–330 (Chapter 15).
[17] W. McCulloch, W. Pitts, A logical calculus of ideas immanent in nervous activity, Bulletin of Mathematical Biophysics 5 (1943) 115–133.
[18] G. Monaci, P. Vandergheynst, F.T. Sommer, Learning bimodal structure in audio–visual data, IEEE Transactions on Neural Networks 20 (2009) 1898–1910.
[19] V.B. Mountcastle, The columnar organization of the neocortex, Brain 120 (1997) 701–722.
[20] R. Munkong, B.H. Juang, Auditory perception and cognition, IEEE Signal Processing Magazine 98 (2008) 98–117.
[21] A. Palmer, S. Shamma, Physiological representation of speech, in: S. Greenberg, W. Ainsworth, A. Popper (Eds.), Speech Processing in the Auditory System, Springer, New York2004, pp. 163–230.
[22] S. Ramón y Cajal, (1899–1904) Textura del Sistema Nervioso del Hombre y de los Vertebrados, Madrid: Imprenta y Librería de Nicolás Moya, reprinted in English as: Histology of the Nervous System of Man and Vertebrates (Oxford University Press, 1995).
[23] J.P. Rauschecker, S.K. Scott, Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing, Nature Neuroscience 12-6 (2009) 718–724.
[24] P. Rose, Y. Kinoshita, T. Alderman, Realistic extrinsic forensic speaker discrimination with the diphthong/aI/, in: Proceedings of the 11th Australian International Conference on Speech Science & Technology 2006, pp. 329–334.
[25] S. Shamma, On the role of space and time auditory processing, Trends in Cognitive Sciences 5–8 (2001) 340–348.
[26] S. Shamma, Physiological foundations of temporal integration in the perception of speech, Journal of Phonetics 31 (2003) 495–501.
[27] Shepherd, G.M., The Synaptic Organization of the Brain (Oxford University Press, New York, 2004).
[28] N. Suga, Basic Acoustic patterns and neural mechanisms shared by humans and animals for auditory perception, in: S. Greenberg, W.H. Ainsworth (Eds.), Listening to Speech: An Auditory Perspective, Lawrence Erbaum Associates, 2006, pp. 159–181.
[29] H.M. Sussman, H.A. McCaffrey, S.A. Mathews, An investigation of locus equations as a source of relational invariance for stop place categorization, Journal of the Acoustical Society of America 90 (1991) 1309–1325.
[30] P. Yin, L. Ma, M. Elhilali, J. Fritz, S. Shamma, Primary auditory cortical responses while attending to different streams, in: B. Kollmeier et al.(Ed.), Hearing: From Sensory Processing to Perception, Springer, Heidelberg, 2007, pp. 257–265.

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.

Agustín Álvarez-Marquina was born in Madrid, Spain in 1969. He received the M.Sc. degree in Computer Science in 1994 and the Ph.D. degree in Computer Science from the Universidad Politécnica de Madrid, Madrid, Spain, in 1999. He is Associate Professor in the Computer Science and Engineering Department, at Universidad Politécnica de Madrid since 2000. His current research interests are speech recognition, speaker identification and architectures for digital signal processing

Luis Miguel Mazaira Fernández was born in Madrid, Spain in 1978. He received the M.Sc. degree in Computer Engineering in 2003, and the Certificate of Advance Studies (DEA) in 2005 from the Universidad Politécnica de Madrid. He is Assistant Professor in the Computer Science and Engineering Department, at Universidad Politécnica de Madrid since 2005 and is currently pursuing his Ph.D. degree with the GIAPSI research group. His current research interests are biomedical signal processing, speaker identification, cognitive speech recognition, pattern recognition.

Rafael Martínez-Olalla was born in Madrid, Spain in 1969. He received the M.Sc. degree in Communications Engineering in 1995 and the Ph.D. degree in Computer Science from the Universidad Politécnica de Madrid, Madrid, Spain, in 2002. He is Associate Professor in the Computer Science and Engineering Department, at Universidad Politécnica de Madrid since 2007. His current research interests are biomedical signal processing, speaker identification, and genomic signal processing.

Cristina Muñoz-Mulas was born in Madrid, Spain in 1982. She received the M.Sc. degree in Computer Science in 2006. She is Ph.D. student in the Computer science and Engineering Department at Universidad Politécnica de Madrid since 2007. Her research topic is about speaker identification and gender discrimination by speech signal processing.