BIO-INSPIRED BROAD-CLASS PHONETIC LABELLING

P. Gómez[1], J. M. Ferrández[2], V. Rodellar[1], R. Martínez[1], C. Muñoz[1], A. Álvarez[1], L. M. Fernández[1]

[1]Grupo de Informática Aplicada al Tratamiento de Señal e Imagen, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, 28660 Madrid, Spain
e-mail: pedro@pino.datsi.fi.upm.es
[2]Universidad Politécnica de Cartagena, Campus Universitario Muralla del Mar, Pza. Hospital 1, 30202 Cartagena, Spain

ABSTRACT

Recent studies have shown that the correct labeling of phonetic classes may help current Automatic Speech Recognition (ASR) when combined with classical parsing automata based on Hidden Markov Models (HMM). Through the present paper a method for Phonetic Class Labeling (PCL) based on bio-inspired speech processing is described. The methodology is based in the automatic detection of formants and formant trajectories after a careful separation of the vocal and glottal components of speech and in the operation of CF (Characteristic Frequency) neurons in the cochlear nucleus and cortical complex of the human auditory apparatus. Examples of phonetic class labeling are given and the applicability of the method to Speech Processing is discussed.

Index Terms— Adaptive Speech Processing, Phonetic Labeling, Bio-inspired Systems, Cognitive Audio.

1. INTRODUCTION

Bio-inspired Speech Processing is the treatment of speech following paradigms used by the human sound perception system, which has specific structures for this purpose. An open question is if bio-inspiration is convenient for specific tasks as Speech Recognition [12][9]. Bio-inspiration may offer alternative ways to implement specific functions in speech processing, helping to improve the performance of conventional methods. Cognitive Audio as a whole and Speech Understanding in particular are specific application areas requiring capabilities such as location and movement detection, source enhancement and separation, source identification, speaker's identification, recognition of discourse, detection of emotions in voice, etc. These are part of what is known as Cognitive Audio, in the sense that the understanding of the surrounding world by humans and machines is strongly dependent on the agile generation and management of Representation Scenarios and Spaces. The purpose of the present paper is to provide a hierarchical description of speech processing by bio-inspired methods discussing the fundamentals of speech understanding, helping to devise a general bio-inspired architecture for Cognitive Audio in the long range. For such a specific task as it is the Phonetic Labelling of Speech has been selected as an objective in improving ASR.

2. SPEECH PRODUCTION

Speech is produced by the combined action of different organs as simplified in the system diagram of Figure 1.

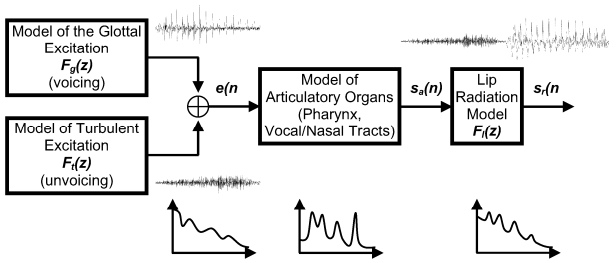


Figure 1. Speech Production Model. Voiced speech is due to the vibration of the vocal folds, whereas unvoiced speech is produced by turbulent flow of air in certain parts of the Vocal Tract. The Articulatory Organs produce specific modifications in the spectral density of the resulting signals which convey very specific message clues.

This means that speech may be divided in voiced and unvoiced segments, depending if vocal fold activity is present or not. Each one of them would imply a different representation under the spectral point of view, voiced sounds being dominated by the action of strong harmonic series colored by the changing vocal tract transfer function modified constantly by the articulation organs. This is so for the vocalic core of the syllables (except in the case of whispered speech). For unvoiced speech there is still a strong coloring of the sibilant sounds produced in plosives and fricatives resulting from the positions where air constrains leading to turbulence occur, as it may be easily perceived in the examples given in Figure 2 (The International Phonetic Alphabet as defined in [1] has been used for phonetic annotation throughout the paper).

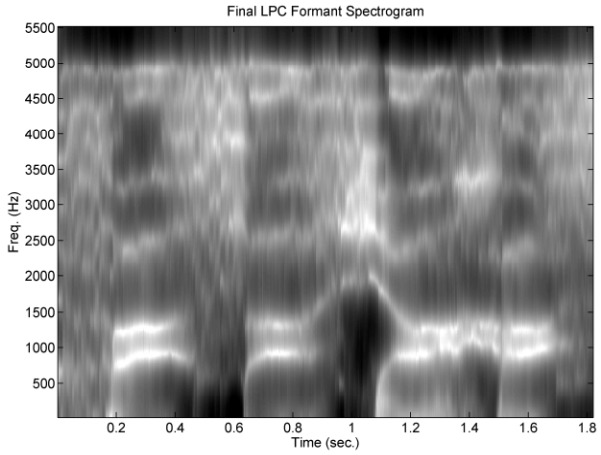


Figure 2. Adaptive Lineal Prediction (ALP) Spectrogram corresponding to the syllables /fa, θa, ʃa, χa/ uttered by a Spanish male speaker. The four unvoiced fricatives are marked by certain colored bursts of wideband noise preceding the vowel formants (the first two formants may be clearly appreciated in all cases as parallel bands near 800 and 1300 Hz). It may be seen that the fricatives are characterized by 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 unstable) maxima near the formants of /a/ for /χ/. The coloring of the spectrum with the formants of the vowel is more noticeable in this last consonant.

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 colored noisy bursts. Therefore harmonics and formants would play a dominant role in speech perception regarding timbre and meaning. In the present case, as phonetic class labeling is the aim of the work, formant positions, dynamics and unvoiced burst coloring are the matter of study in relation to relevant aspects for ASR applications.

3. SPEECH PERCEPTION

Speech is perceived by the Auditory System, which can be seen as a chain of different sub-systems depending on their relative role in the processing of information (see Figure 3).

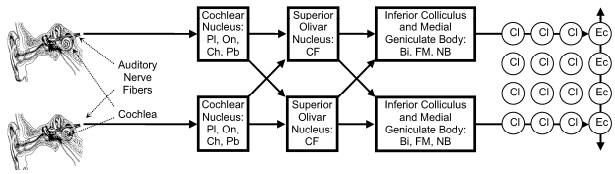


Figure 3. Speech Perception Model. The Cochlea produces time-frequency organized representations which are conveyed by the Auditory Nerve to the Cochlear Nucleus, where certain specialized neurons (Pl: Primary-like, On: Onset, Ch: Chopper, Pb: Pauser) are implied in temporal processing. Binaural information is treated in the Superior Olivar Nucleus, where tono-topic units (CF) have been identified. 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).

The most important organ of the Peripheral Auditory System is the Cochlea, which carries out the separation in frequency and time of the different components of Speech. Important processing is produced in the basilar membrane within the cochlea. Low frequencies produce maximum excitation in the apical end of the membrane, while high frequencies excite the basal area. Thus the excitation of transducer cells (hair-cells) responsible for the mechanical to neural transduction process is tono-topic. Electrical impulses propagate to higher neural centers through auditory nerve fibers of a different characteristic frequency (CF) responding to each of the spectral components (F_0, F_1, F_2...) of speech [20]. Within the cochlear nucleus (CN) different types of neurons are specialized in segmenting the signals (Ch: chopper units), detecting stimuli onsets (On: onset cells), delaying the information to detect temporal patterns (Pb: pauser units), or transferring the information (Pl: primary-like units). The Cochlear Nucleus feeds information to the Olivar Complex, where sounds are located by inter-aural differences, and to the Inferior Colliculus (IC), which is organized in spherical layers with orthogonal isofrequency bands. Delay lines are found in this structure to detect temporal features in acoustic signals (CF and FM components). The thalamus (Medial Geniculated Body) acts as a relay station (some neurons exhibit delays of a hundred milliseconds), and as a tonotopic mapper of information arriving to cortex as *ordered feature maps* [20]. The specific location of the neural structures in the cortex responsible for speech processing and understanding is not well established as the subjects of experimentation have been mainly animals. Neurons have been found in cats that fire when FM-like frequency transitions are present (FM elements) [15], while in macaque some neurons respond to specific noise bursts (NB components) [18]. Other neurons are specialized in detecting the combinations among these elements [21]. In humans, evidence exists of a frequency representation map in the Heschl circumvolution [19] and of a secondary map with word-addressing capabilities [17]. A comprehensive review of the structures involved and their functionality is given in [5]. As a summary the specific processing of speech by the Auditory System is based on the hierarchical detection and association of stable 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. 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). A recent review of these topics is given in [22]. At a higher hierarchy 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 [3]. From these studies a generalized model may be issued under a perceptual point of view as represented in Figure 4.

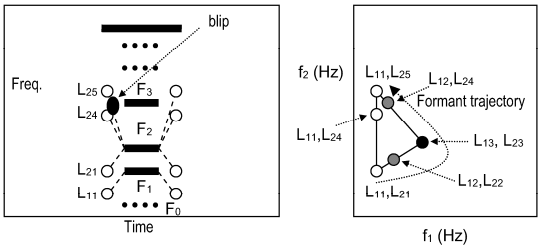


Figure 4. Generalized Syllable Model (GSM). Top: Time Domain Behavior. Bottom: Vowel Triangle Projection and Dynamic Behavior. White circles signal loci positions. The dark dot gives the position of the specific vowel modeled (/a/ in the present case).

4. BIO-INSPIRED SPEECH PROCESSING

From the study of the Generalized Syllable Model and the Auditory Speech Processing fundamentals exposed before, a Basic Neuron Set could be defined as an algorithmic structure operating both in the time and frequency domain modeling speech features, among these:

- 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 by the general structure shown in Figure 5.

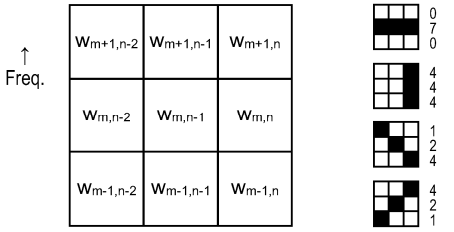


Figure 5. Basic Neuron Set for elementary operations on time-frequency representations of speech. Left: 3x3 weight mask. Right: Masks for feature detection on the formant spectrogram. Each mask is labelled with the corresponding octal code (most significant bits: bottom-right). Labels 070, 444, 124 and 421 correspond respectively with CfI, NB, NfM, PfM units.

The problem of feature detection in formant spectrograms is related to a well known one in Digital Image Processing [14]. A classical methodology is based on the use of reticule masks on the image matrix *X(m,n)*:

$$\tilde{X}(m,n) = \sum_{i=-1}^{1} \sum_{j=0}^{2} w_{i,j} X(m-i, n-j) \qquad (1)$$

where $\{w_{ij}\}$ is a 3x3 mask with a specific pattern and a set of weights. The basic cells for formant trajectory processing shown in Figure 5 have been derived from the neural structures of the auditory centers in brain, as presented in section 3. Weight adjustment may also be adaptive. In this last case a database of spectrograms and MLP structures for the training of each cell [11] are to be used. The objective of the present work is speech processing for broad class phonetic labeling. For such, the detection of formants is carried out through ALP algorithms producing all-pole spectral positions which keep track of the vocal tract resonances [4]. Precise formant positions may be obtained from these rough representations applying lateral inhibition between neighbor CF units using specific weight configurations of the mask in Figure 5 as shown in Figure 6.

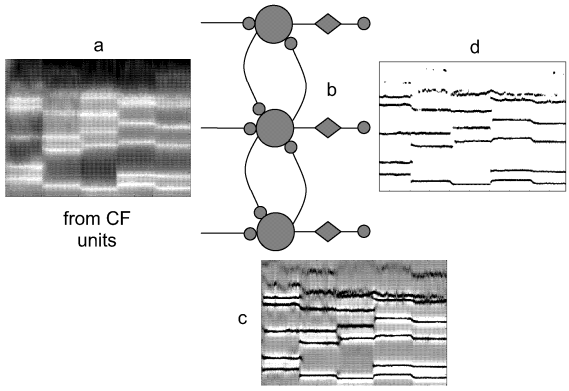


Figure 6. Formant Trajectory Profiling for the sequence /aeiou/ (Spanish, male speaker): The speech spectral density (a) as detected by CF units (see next section) is processed by columns of neurons implementing lateral inhibition (b), producing differentially expressed formant lines (c), which are transformed into narrow formant trajectories (d) after non-linear saturation.

The lateral inhibition filter produces sharp estimations of the spectral peaks (see Figure 6.b). The whitish bands surrounding the formants are due to the characteristic "mexican hat" response of the filter. The final formant distribution is given in Figure 6.d after adaptive saturation.

5. METHODOLOGY

Although there are good models [13] to deal with time-frequency stimulus separation, the present work will rely on 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 (avoiding the confusion and overlapping of the glottal formant and the lower vowel formants) following the iterative method represented in Figure 7.

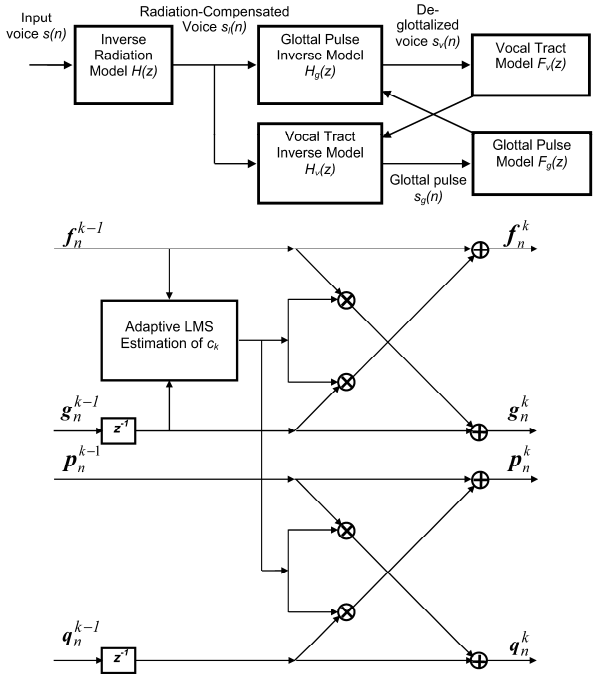


Figure 7. Top: Iterative implementation of glottal source and vocal tract separation using paired lattices (bottom).

Vocal tract inversion has been carried out using an adaptive lattice filter, allowing th easy extension of filter order, agile tracking of non-stationary speech, wave-propagation filter properties, glottal source and vocal tract decoupling, and easy vocal tract transfer function inversion and removal from speech. The precise reconstruction of the glottal source may be implemented by a lattice-ladder structure [4] or by a coupled pair of lattices [8]. Formant features are to be estimated from the resulting spectrogram, as:

$$X(m,n) = 20\log_{10}\frac{1}{\left|\sum_{k=1}^{p} a_{kn} e^{-jkm\pi\Omega}\right|} \qquad (2)$$

where a_{kn} are the coefficients of the equivalent ALP filter of order p estimated at each time instant n from the speech signal $x(n)$, 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). Many tools devised for image processing can be used for the detection of time-frequency features, as CF or FM patterns within a systemic framework given in Figure 8.

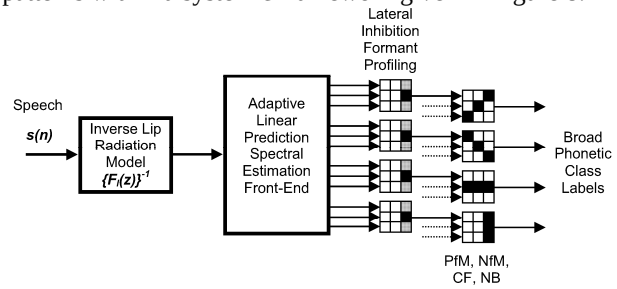
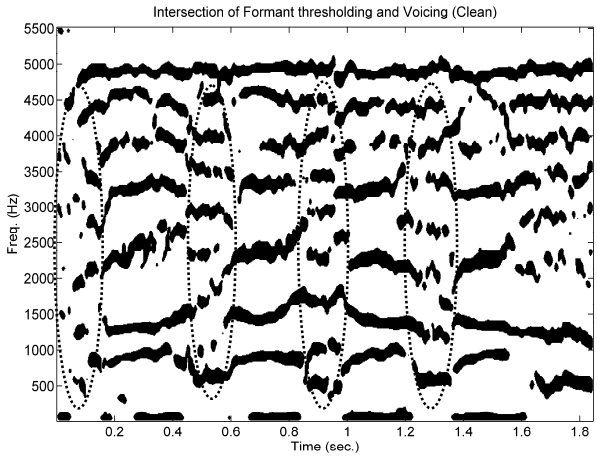


Figure 8. Bio-inspired Speech Processing Framework used in the study for a mono-aural channel.

The first basic operation on the LPC spectrogram will be to profile formant trajectories following the structure given in Figure 6. This is carried out using bio-inspired lateral inhibition, this mechanism being active in certain neuron associations in the Inferior Colliculus [20]. The proposed algorithm is expressed as:

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

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. Examples are given in Figure 9 for three consonantal phonetic classes articulated on vowel /a/ in a C-V syllabic structure using this methodology.



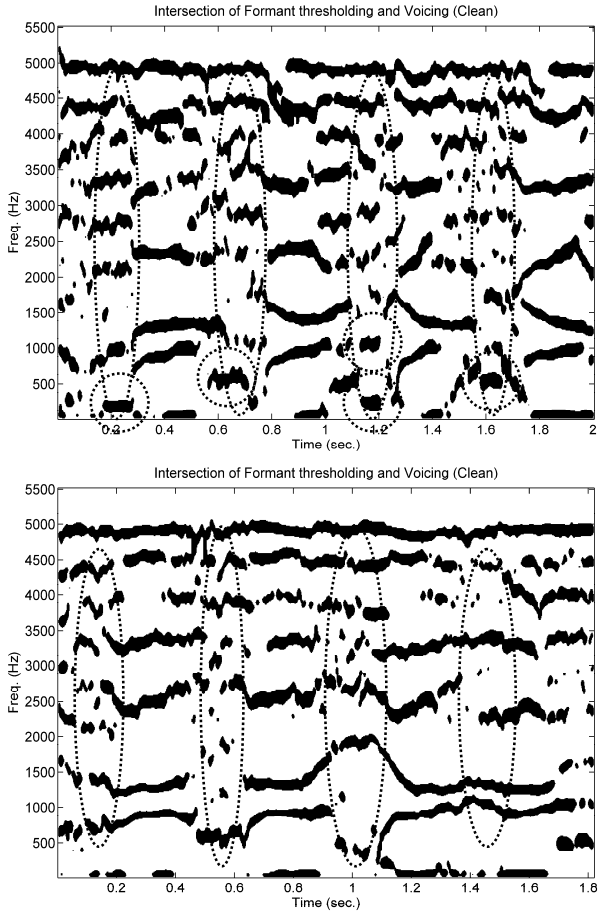


Figure 9. Detection of formants from the ALP spectrogram using lateral inhibition and nonlinear saturation. The patterns within dot circles among vowel formants are specific consonantal marks [3]. Top template: Formants corresponding to unvoiced stops /pa, ta, ca, ka/. Middle template: Formants corresponding to voiced stops /ba, da, ɟa, ga/. Bottom template: Formants corresponding to unvoiced fricatives /fa, θa, ʃa, χa/.

6. RESULTS

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 signaling ascending and descending formants. As an example results of processing four structures of the type V-C-V including four voicing fricatives are given in Figure 10. 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 msec. The dimensionality of the PfM, NfM, CF and NB units was also of 512 units per layer.

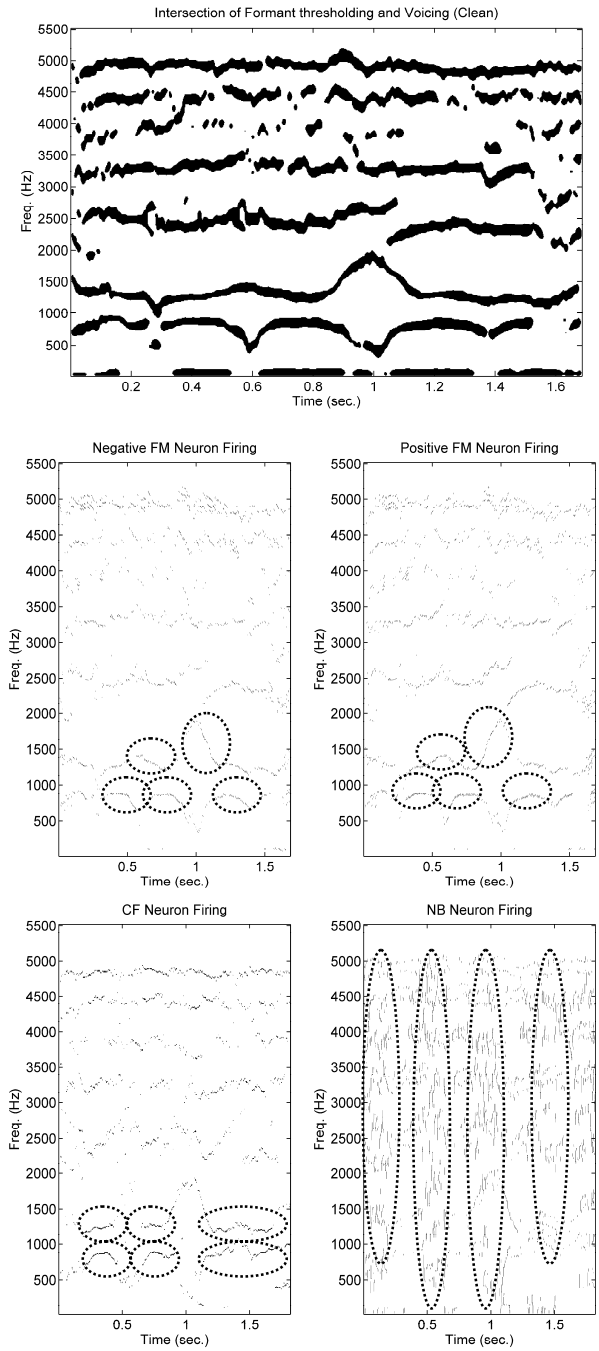


Figure 10. Detection of four basic formant patterns using PfM, NfM, CF and NB units for the V-C-V groups /aβa–aða–aʒa–aɣa/ of voiced fricatives in Spanish. Top: Formant profiled spectrogram. Middle left: Outputs from NfM units. Middle right: Id. from PfM units. Bottom left: Id. from CF units. Bottom right: Id. from NB units. The basic patterns detected are enclosed within dash circles.

The simple structures proposed for unsupervised consonantal feature detection are able of signaling stable,

ascending and descending formants, and noise bursts. These may help in labeling phonetic classes according to Table 1.

Table 1. Features of Broad Phonetic Classes

Classes	CF	FM	NB
Vowels	+	-	-
Voiced stops	+	+	+
Nasals	+	-	-
Glides	-	+	-
Unvoiced stops	-	+	+
Fricatives	-	-	+

This may be of great help in improving recognition rates in ASR as much as 26% (see [10]) by simplifying State-Transition Graph Search in HMM parsing reducing tri-phone ambiguity and computational needs. A deeper study is being conducted to improve the detection of glides and voiced stops by supervised Neural Networks [11].

7. CONCLUSIONS

Through the present work a simple architecture to detect and label broad class phonetic features has been presented. The results show the viability of bio-inspired phonetic feature detection using computationally inexpensive structures. More work is to be done to establish normalized thresholds and configuration parameters to improve robustness. The statistical performance of the methodology against large speaker databases in Spanish show improvements in labeling of around 6-10% against blind supervised labeling, although this study is far from being complete. The weaknesses in this approach are the need of adaptive threshold weight adjustments in (1) to optimize formant tracking. These questions remain the object of future study, as well as the role of higher-level phonetic class association, as well as the columnar organization of the Auditory Cortex [16] to include short-time memory and retrieval by Generalized Autoregressive Units.

8. 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.

9. REFERENCES

[1] Available from http://www.arts.gla.ac.uk/IPA/ipachart.html

[2] Culling, J. F. and C. J. Darwin, "Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0", *J. Acoust. Soc. Am.*, Vol. 93, pp. 3454-3467, 1993.

[3] Delattre, P., Liberman, A., Cooper, F.: Acoustic loci and transitional cues for consonants. *J. Acoust. Soc. Am.*, Vol. 27, pp. 769-773, 1955.

[4] Deller, J. R., Proakis, J. G., and Hansen, J. H. L.: *Discrete-Time Processing of Speech Signals*, Macmillan, NY, 1993.

[5] Ferrández, J. M.: Study and Realization of a Bio-inspired Hierarchical Architecture for Speech Recognition. Ph.D. Thesis (in Spanish), Universidad Politécnica de Madrid, 1998.

[6] Geissler, D. B and Ehret, G., "Time-critical integration of formants for perception of communication calls in mice", *Proc. of Nat. Ac. Sc.*, Vol. 99, No. 13, pp. 9021-9025, 2002.

[7] Gómez, P., Ferrández, J. M., Rodellar, V., Álvarez, A., Mazaira, L. M., "A Bio-inspired Architecture for Cognitive Audio", *Lecture Notes on Computer Science*, Vol. 4527, pp. 132-142, 2007.

[8] Gómez, P., Godino, J. I., Álvarez, A., Martínez, R., Nieto, V., Rodellar, V., "Evidence of Glottal Source Spectral Features found in Vocal Fold Dynamics", *Proc. of the ICASSP'05*, pp. 441-444, 2005.

[9] Gómez, P., Martínez, R., Rodellar, V., Fernández: J. M.. "Bio-inspired Systems in Speech Perception: An overview and a study case", *IEEE/NML Life Sciences Systems and Applications Workshop* (by invitation), National Institute of Health, Bethesda, Maryland, July 13-14, 2006.

[10] Gravier, G., Yvon, Y., Jacob B. and Bimbot, F., "Introducing contextual transcription rules in large vocabulary speech recognition", in *The integration of phonetic knowledge in speech technology*, William J. Barry and Win A. Van Domelen Eds, Springer series on Text, Speech and Language Technology, vol. 25, chapter 8, pp. 87-106, 2005.

[11] Haykin, S. *Neural Networks – A comprehensive Foundation*, Prentice-Hall, Upper Saddle River, NJ, 1999.

[12] Hermansky, H., "Should Recognizers Have Ears?", *ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-à-Mousson, France, 17-18 April, pp. 1-10, 1997.

[13] Irino, T., and Patterson, R. D., "A time-domain, level-dependent auditory filter: the gammachirp", *J. Acoust. Soc. Am.*, vol. 101, no. 1, pp. 412-419, 1997.

[14] Jähne, B., *Digital Image Processing*, Springer, Berlin, 2005.

[15] Mendelson J. R., Cynader, M. S., "Sensitivity of Cat Primary Auditory Cortex (AI) Neurons to the Direction and Rate of Frequency Modulation", *Brain Research*, Vol. 327, pp. 331-335, 1985.

[16] Mountcastle, V. B., "The columnar organization of the neocortex", *Brain*, Vol. 120, pp. 701-722, 1997.

[17] Ojemann, G. A., "Organization of language cortex derived from investigation during neurosurgery", *Sem. Neuros.*, Vol. 2, pp. 297-305, 1990.

[18] Rauschecker, J. P., Tian, B., Hauser, M., "Processing of Complex Sounds in the Macaque Nonprimary Auditory Cortex", *Science*, Vol. 268, 7 April, pp. 111-114, 1995.

[19] Sams, M., Salmening, R., "Evidence of sharp frequency tuning in human auditory cortex", *Hearing Research*, Vol. 75 pp. 67-74, 1994.

[20] Schreiner, C.E, "Order and Disorder in Auditory Cortical Maps", *Curr. Op. Neurobiol.*, Vol. 5, pp. 489-496, 1995.

[21] Suga, N., "Cortical Computational Maps for Auditory Imaging", *Neural Networks*, Vol. 3, pp. 3-21, 1990.

[22] Yin, P., Ma, L., Elhilali, M., Fritz J. and Shamma, S., "Primary Auditory Cortical Responses while Attending to Different Streams", in *Hearing: From Sensory Processing to Perception*, B. Kollmeier et al., eds., Springer, Heidelberg, pp. 257-265, 2007.