BIO-INSPIRED DYNAMIC FORMANT TRACKING FOR PHONETIC LABELLING

P. Gómez, J. M. Ferrández, V. Rodellar, R. Martínez, C. Muñoz, A. Álvarez, L. M. Mazaira

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

ABSTRACT

It is a known fact that phonetic labeling may be relevant in helping current Automatic Speech Recognition (ASR) when combined with classical parsing systems as HMM's by reducing the search space. Through the present paper a method for Phonetic Broad-Class Labeling (PCL) based on speech perception in the high auditory centers is described. The methodology is based in the operation of CF (Characteristic Frequency) and FM (Frequency Modulation) neurons in the cochlear nucleus and cortical complex of the human auditory apparatus in the automatic detection of formants and formant dynamics on speech. Results obtained in formant detection and dynamic formant tracking are given and the applicability of the method to Speech Processing is discussed.

1. INTRODUCTION

Bio-inspired Speech Processing is the treatment of speech following paradigms used by the human sound perception system, which has developed specific structures for this purpose. 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 [6]. For such, the dynamic tracking of formants has been selected as an objective in improving ASR. 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 filtered by the changing vocal tract transfer function modified constantly by the articulation organs. This stands also 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. Normal speech may be perceived as sequences of harmonic series filtered by the resonances of the vocal tract (formants) with characteristic onsets and trails, which may be preceded or followed by noisy bursts as shown in Figure 1 for four syllables of the sort V-C-V where C stands for a specific voiced approximant.

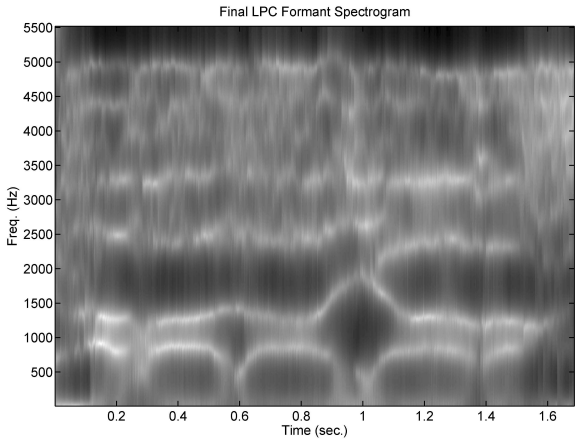

Figure 1. Adaptive Lineal Prediction (ALP) Spectrogram corresponding to the syllables /aβa-aða-aʒa-aɣa/ uttered by a Spanish male speaker. The IPA has been used for annotation [1].

This example has been selected for the fast dynamic movements of formants present in it, as dynamic formant tracking for phonetic class labelling is the aim of the work.

2. SPEECH PERCEPTION

Speech is perceived by the Auditory System described in Figure 2 as a chain of different sub-systems integrated by the Peripheral Auditory System (Outer, Middle and Inner Ear) and the Higher Auditory Centers. 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 Speech and their transduction from mechanical to neural activity. 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 different characteristic frequencies (CF) responding to the spectral components (F0, F1, F2...) of speech [10].

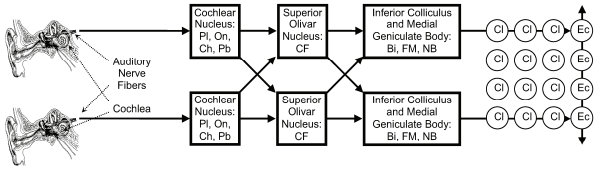

Figure 2. Speech Perception Model. The Cochlea produces time-frequency organized representations which are conveyed by the Auditory Nerve to the Cochlear Nucleus, where certain specialized neurons (Pl: Primary-like, On: Onset, Ch: Chopper, Pb: Pauser) are implied in temporal processing. Binaural information 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).

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 (Pb: pauser units), or acting as relay stations (Pl: primary-like units). 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 isofrequency bands. Delay lines are found in this structure to detect temporal features in acoustic signals (CF and FM components). The thalamus (Medial Geniculate Body) acts as a last relay station, and as a tonotopic mapper of information arriving to cortex as ordered feature maps.

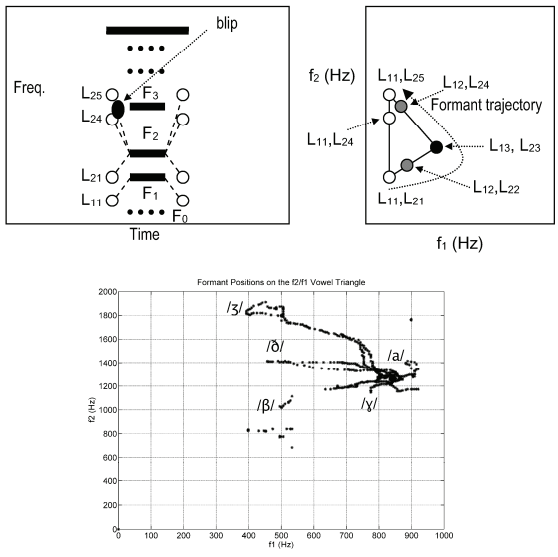

Figure 3. Generalized Phoneme Model. Top left: loci of the GPM on the vowel triangle. White circles indicate the positions of the loci. Top right: Dynamic trajectories on the vowel triangle. Bottom: Formant trajectories for the trace /aβaðaʒaɣa/ shown in Figure 1. The dark dot gives the position of the specific vowel modeled (/a/ in the present case).

Neurons have been found in the cortex that fire when FM-like frequency transitions are present (FM elements), while some others respond to specific noise bursts (NB components). Other neurons are specialized in detecting the combinations among these elements. In humans, evidence exists of a frequency representation map in the Heschl circumvolution and of a secondary map with word-addressing capabilities. A comprehensive review of the structures involved and their functionality is given in [4]. 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 a higher hierarchy dynamic changes in formants (onset times and slopes) and specific broadband signals present before the onset time define specific clues to the perception of syllables, seen as structures of consonants and vowels as in C-V, C-V-C, V-C-V, etc. The perceptual interpretation of such structures is well known since the works of Delattre et al. [2]. From these studies a Generalized Phoneme Model may be issued as represented in Figure 3. The static version of the model is based on formant positions and loci (places marking the starting and ending points of formant trajectories. The dynamic version is based on a projection on the vowel triangle (f2 vs f1).

3. BIO-INSPIRED SPEECH PROCESSING

From the study of the Generalized Phoneme Model and the Auditory Speech Processing fundamentals, 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) finite difference algorithms in the frequency domain profiling formants; Positive Frequency Modulation Units (PfM), detectors of up-hill formant displacements; Negative Frequency Modulation Units (NfM), detectors of down-hill formant displacements; Characteristic Frequency Units (CF), detectors of stable frequency positions; Vowel-Spotting Units (VS), detectors of stable or parallel-moving pairs of frequencies and 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 4.

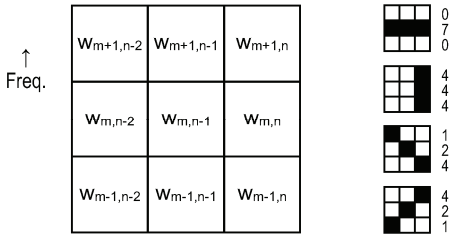

Figure 4. Basic Neuron Set for elementary operations on time-frequency representations of speech. Left: 3x3 weight mask. Right: Masks for feature detection on the formant spectrogram. Each mask is labelled with the corresponding octal code (most significant bits: bottom-right). Labels 070, 444, 124 and 421 correspond respectively with CfI, NB, NfM, PfM units.

In this way the problem of feature detection in formant spectrograms is related to a well known one in Digital Image Processing [8]. A classical method is based on the use of reticule masks on the spectrogram *X(m,n)*:

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

where $\{w_{ij}\}$ is a *IxJ* mask with a specific pattern and a set of weights, which may be adjusted adaptively. The spectrogram is built using ALP algorithms producing all-pole spectral positions which keep track of the vocal tract resonances [3]. 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 4 as shown in Figure 5.

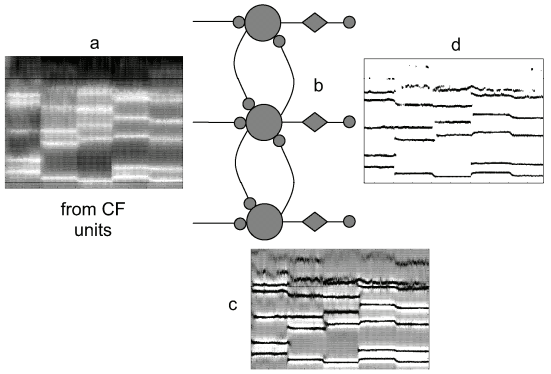

Figure 5. Formant Trajectory Profiling for the sequence /aeiou/ (Spanish, male speaker): The speech spectral density (a) as detected by CF units (see next section) is processed by columns of neurons implementing lateral inhibition (b), producing differentially expressed formant lines (c), which are transformed into narrow formant trajectories (d) after non-linear saturation.

The lateral inhibition filter produces sharp estimations of the spectral peaks (see Figure 5.b). The final formant distribution is given in Figure 5.d after adaptive saturation. Other personalized neurons can be used for the detection of time-frequency features, as CF or FM patterns as shown in the systemic framework given in Figure 6. The first operation on the LPC spectrogram will be to profile formant trajectories using lateral inhibition as described. The rest of the structure works as follows: PfM and NfM are neurons specialized in detecting positive and negative movements formants, firing in response to different slopes ($+fM_{1\text{-}k}$, $-fM_{1\text{-}k}$); CF are neurons detecting the stable positions of formants firing when a given channel is active during a specific interval ($f1_{1\text{-}k}$ and $f2_{1\text{-}k}$ being the bands associated to the first two formants); $NB_{1\text{-}k}$ are neurons which fire when broad band activity is detected; Σ units (middle left) are specialized in adding formant dynamics, integrating channel activity ($\int$) and thresholding (f). Lateral inhibition is again used in eliminating possible ambiguities in the final detected activity in the first two formants (Dynamic Tracking Units +fM1, -fM1, +fM2, -fM2); finally Vowel Spotting Units and Voiceless Activity may be derived from f1, f2 and NB channels.

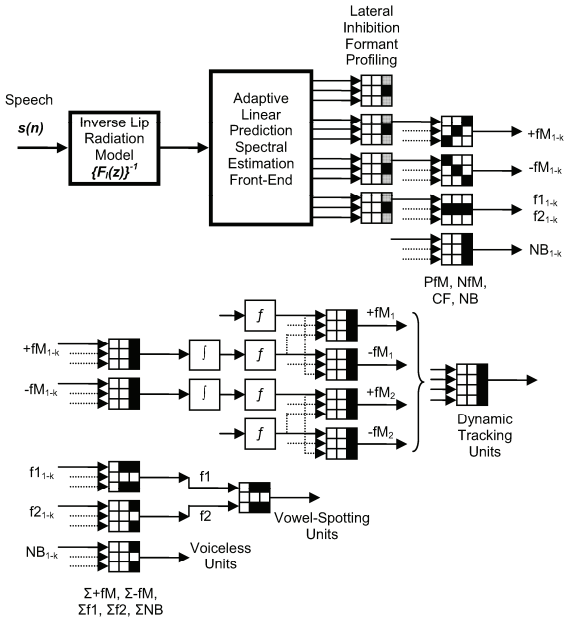

Figure 6. Bio-inspired Speech Processing Framework used in the study for a mono-aural channel.

4. RESULTS AND DISCUSSION

As an example Figure 7 illustrates the activity of Dynamic Tracking Units in processing a specific sentence as {es hábil un solo día}.

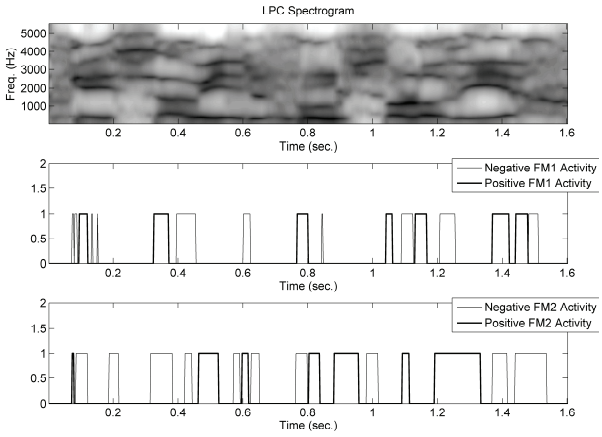

Figure 7. Detection of formant dynamics from the ALP spectrogram (top) using lateral inhibition and nonlinear saturation. The positive and negative slopes for f1 (middle) and f2 (bottom) have been detected from the sentence {es hábil un solo día} uttered by a male speaker (046).

The long and intense climbing up and sliding down of f2 for /δía/ can be appreciated in the lowest template of the between 1.2-1.35 and 1.37-1.53 (separated in two different intervals in this last case). The combinations of these four signals (+fM1, -fM1, +fM2, -fM2) are a first broad labeling of the series of approximants studied. The estimates of the values of the two slopes of /δía/ for 8 male and 8 female speakers are given in Table 1.

Table 1. Second formant positive and negative slopes for /δía/

Male speakers (Hz/sec)			Female speakers (Hz/sec)		
Speaker	+fM2	-fM2	Speaker	+fM2	-fM2
046	6899	-8191	A23	7867	-6921
081	6603	-5955	A66	769	-2216
115	8450	-3445	B31	5696	-18079
126	5662	-4915	B97	6591	-2791
160	6065	-4688	C76	3166	-5191
208	7660	-5666	D45	19539	-14086
231	5779	-3095	D77	6497	-5088
304	6302	-18542	D87	6544	-2512

It may be seen that with certain exceptions the values of the slopes range from 5000 to 8000 Hz/sec. More research is to be conducted to determine the robustness of the estimates. Dynamic formant trajectory detection and characterization is important for forensic applications. The methodology presented may be used also for the detection of the statistical vertices of vowel triangles for different speakers, as given in Figure 8, derived from the formant trajectories of the same sentence {es hábil un solo día} produced by the same set of 8 male and 8 female speakers.

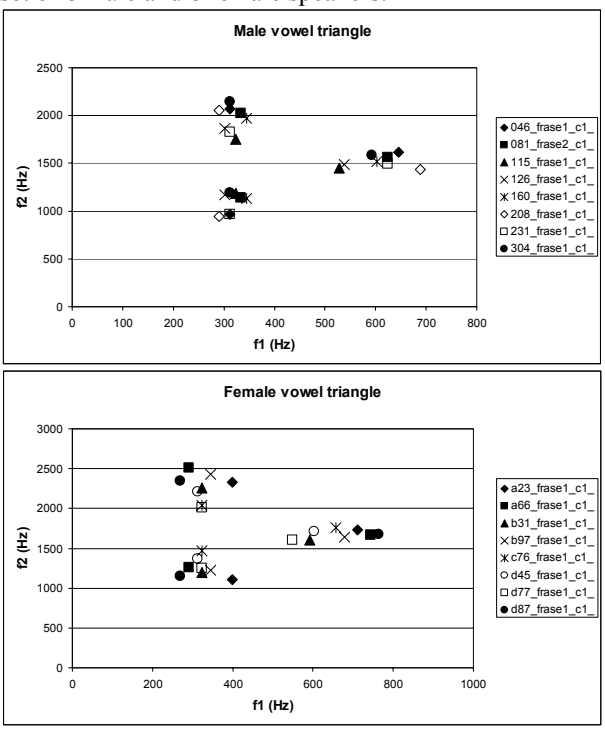

Figure 8. Detection of the vowel triangle centroids for eight male and eight female speakers using the bio-inspired methodology proposed. The positions of the vertices have been estimated using the lowest and highest quantiles of f1 and f2 statistical distributions for each speaker.

In general it may be observed that statistical spread is larger in female than in male, and that the upper left vertex is the one showing larger inter-gender differences.

5. CONCLUSIONS

Through the present work a hierarchical architecture to detect and label broad class phonetic features has been presented using replications of a Basic Neuron Set. The results show the viability of bio-inspired phonetic feature detection using combinations of these computationally inexpensive structures. The structures proposed are able of signaling stable, ascending and descending formants, and noise bursts. This may be of great help in improving recognition rates in ASR as much as 26% (see [7]) by simplifying State-Transition Graph Search in HMM parsing. More work is to be done to establish normalized thresholds and configuration parameters to improve robustness. Preliminary studies show that the statistical performance of the methodology show improvements in labeling of around 6-10% against blind supervised labeling, although this study is not complete yet. These questions remain the object of future study, as well as the role of the columnar organization of the Auditory Cortex [9] to include short-time memory and retrieval by Generalized Autoregressive Units.

ACKNOWLEDGMENTS

This work is being funded by grants 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] Available at http://www.arts.gla.ac.uk/IPA/ipachart.html

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

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

[4] 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.

[5] Goldstein, E. B., *Sensation and Perception*, Wadsworth, Belmont, CA., 2006.

[6] 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.

[7] 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.

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

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

[10] Shamma, S., "On the role of space and time auditory processing", Trends in Cognitive Sciences, Vol., No. 8, pp. 340-348, 2001.