Articulatory Feature Detection based on Cognitive Speech Perception

Pedro Gómez, José Manuel Ferrández, Victoria Rodellar, Roberto Fernández-Baíllo, Agustín Álvarez, Luis Miguel Mazaira

Grupo de Informática Aplicada al Procesado de Señal e Imagen, Facultad de Informática, Universidad Politécnica de Madrid, 28600 Boadilla del Monte, Madrid, e-mail: pedro@pino.datsi.fi.upm.es

Abstract

Cognitive Speech Perception is a field of growing interest as far as studies in cognitive sciences have advanced during the last decades helping in providing better descriptions on neural processes taking place in sound processing by the Auditory System and the Auditory Cortex. This knowledge may be applied to design new bio-inspired paradigms in the processing of speech sounds in Speech Sciences, especially in Articulatory Phonetics, but in many others as well, as Emotion Detection, Speaker's Characterization, etc. The present paper reviews some basic facts already established in Speech Perception and the corresponding paradigms under which these may be used in designing new algorithms to detect Articulatory (Phonetic) Features in speech sounds which may be later used in Speech Labelling, Phonetic Characterization or other similar tasks.

Introduction

Bio-inspired Speech Processing is the treatment of speech following paradigms used by the human sound perception system, which is known to posses specific functions for this purpose. The question if bio-inspiration is a convenient strategy for devising specific tasks as Speech Recognition has been posed in the past, and it remains still open [8]. The generalized impression is that bio-inspiration may offer alternative ways to implement specific functions in speech processing, helping to improve the performance of conventional methods. The need for this approach is justified by the high level of complexity found in Language Processing. This is clearly stated by R. E. Cytowic: "… Language turned out to be far more complex than the grammar found in the textbooks. And yet it is routinely surpassed by what a six-year-old has in her head …" [2]. Many tasks in Speech Processing and Understanding remain unsolved yet, although they may benefit from using algorithmic methods mimicking the functionality shown by the human Auditory Perception structures. In the present paper a brief summary of some of them will be reviewed and commented.

Some Facts in Speech Production

Speech is produced by the combined action of different organs as simplified in the diagram of Figure 1 (Top). The energy for the production of Speech is provided by a set of muscles (the diaphragm among them). The airflow resulting from pressure build-up in the lungs induces the vibration of the vocal folds at a specific position (voiced speech component). Under certain conditions (when laminar flow becomes turbulent in the narrow constraints of the vocal tract) speech sounds are produced as modulated turbulence noise (unvoiced speech component). The Articulation Organs (V: velum, G: tongue, P: palate, A: alveoli, T: teeth and L: lips) produce specific modifications in the spectral density of the resulting sounds, conveying specific message clues. Glotal and turbulence are the two possible sources of excitation to the Vocal Tract, which may appear separated or joint, as symbolized in Figure 1 (Bottom). The Vocal Tract behaves is a Linear Time-Variant system enhancing or reducing certain frequencies at the resonant and anti-resonant positions of the equivalent acoustic pathways [4]. Thus speech may be divided in voiced and unvoiced segments, depending if vocal fold activity is present or not. The separation of the vocal and glottal components of speech is of capital interest for any speech analytical studies. Separation may be carried out using Adaptive Linear Prediction as described in [7]. The tool PhonSpec® running on MATLAB® is especially devised to implement the separation [9]. A view of its Graphic User Interface is given in Figure 2.

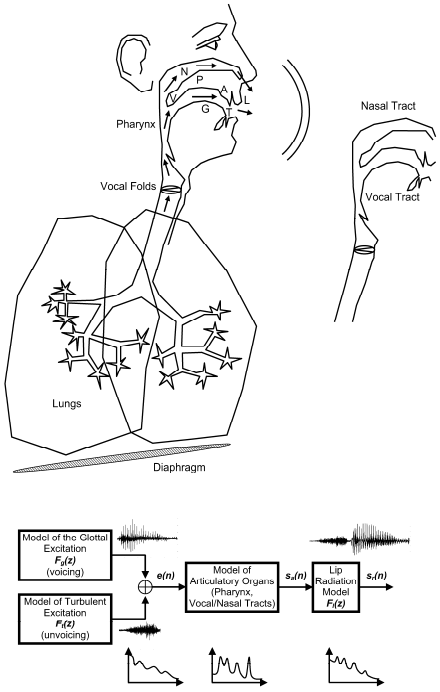

Figure 1. Top: Schematic section of the phonation apparatus (the vocal tract has been separated to its right). Bottom: Speech Production Model.

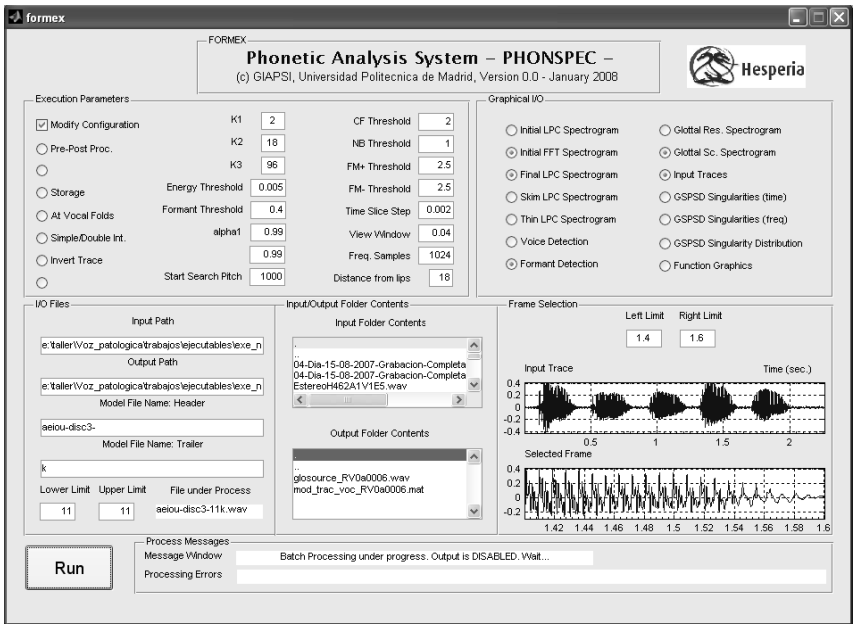

Figure 2. Graphic User Interface of PhonSpec®.

The formant templates and trajectories in the present paper have been produced using this tool. When voicing is present the time-frequency spectrogram will be characterized by regular horizontal bands at the harmonics of the fundamental frequency of the vocal fold vibration, as shown in Figure 3 (Left).

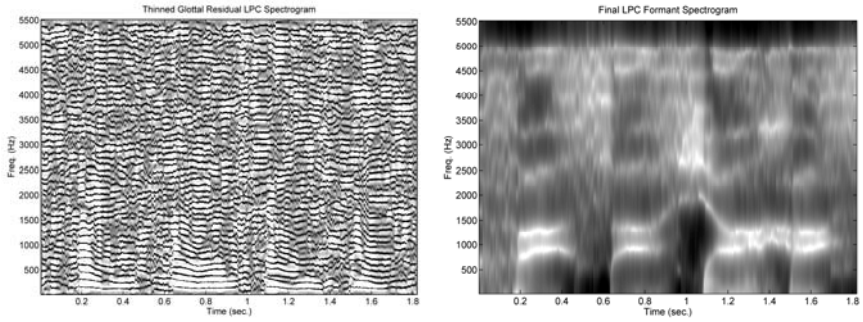

Figure 3. Spectrograms corresponding to the sequence /fa-θa-ʃa-χa/[1] (Spanish male speaker). Left: Harmonic structure of the Glottal Source. Right: Time-Frequency formant positions from Adaptive Linear Prediction (ALP).

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

The vocal tract resonances shown in Figure 3 (right) as bright bands pinpointing formant positions will enhance the energy of the nearby harmonics. Unvoiced speech shows also strong dominant bands which can not be considered formants in the strict sense, but that are due also to specific resonances when turbulence is produced at the back of the oral cavity, as is the case of /χa/ between 1.35-1.50 sec. This means that speech is perceived as sequences of harmonic series which may be preceded or followed by consonant-specific noisy bursts, coloured by the resonances of the vocal tract with characteristic onsets and trails, therefore formants will play a dominant role in speech perception.

Perception of Speech Sounds based on Formant Dynamics

The acoustic features acting as baseline patterns in speech are the vowel tract resonances called formants, which give a good description of the message issued (acoustic-phonetic decoding) as well as the speaker's personality. An example of formant distribution for vowels is given in Figure 4.

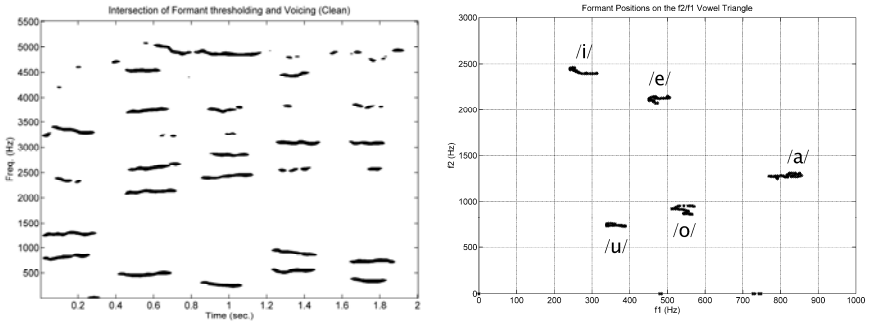

Figure 4. Left: Formant Spectrogram of the five vowels in Spanish (/a–e–i–o–u/) from a male speaker, obtained by Adaptive Linear Prediction using the tool PhonSpec®. Right: Formant plot for the same recording on F_2 vs F_1.

Formants are labelled in order of increasing frequency, F_1 and F_2 being the lowest ones. In the case shown the first two formants for /a/ appear between 750-860 and 1300 Hz, moving to 450-500 Hz and 2200 Hz for /e/, 240-310 and 2400-2470 Hz for /i/, then to 500-560 Hz and 850-980 Hz for /o/ and 340-390 Hz and 750 Hz for /u/. Formants F_3, and higher may be also present in voiced speech, albeit the lowest two formants are sufficient to give a good description of vowel-like phonemes. Each combination of the first two formants is decoded by the Auditory System as a vowel, and assigned a different value accordingly to the phonologic structure of the target language. It is of most importance to emphasize here that the assignment of semantics to specific combinations of formants is highly dependent on the specific language coding system, and therefore universal rules may be hardly

applicable. As a conclusion vowels are represented by relatively stable narrow-band patterns, known phonetically as formants, or perceptually as Characteristic Frequencies (CF). The Auditory System of mammals seems to be highly specialized in the detection of formants and formant groups as basic communication features [6]. Consonant behaviour is based on different clues, as these are the result of constrictions of the articulation organs, producing vowel formant transitions, preceded or accompanied in many cases by turbulent activity. Formant transitions from stable CF positions to new CF positions are known as FM (frequency modulation) components. An example of consonant FM components is given in Figure 5.

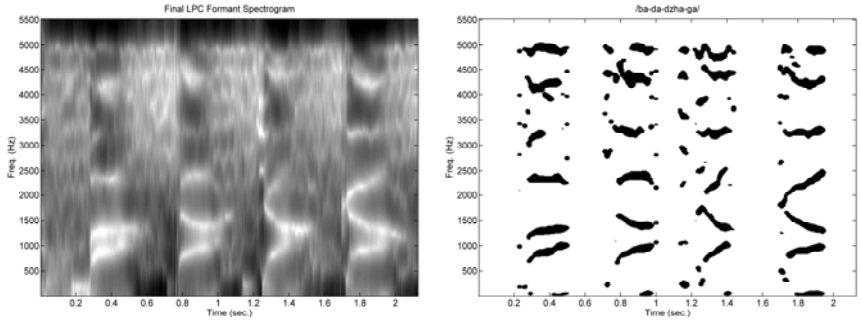

Figure 5. Left: LPC spectrogram of the sequence /ba–da–ʄa–ga/ from the same speaker. Right: Formant plots.

Turbulence-induced broad band sounds are known as noise bursts (NB) or *blips*, and together with formants constitute minimal semantic units or "sematoms"[2]. CF, FM and NB "sematoms" bear important communication clues [5] either by themselves or when associated to other "sematoms". The perception of vowels and consonants based upon FM-like transitions from vowels is explained by the "loci paradigm" [3]. A locus is a formant position previous to the insertion of the consonant, marking a virtual place from where formants move to the situation defined by the stable vocal fragment. In Figure 5 (Bottom) the first formant moves from a virtual locus (800 Hz) to 1000 Hz for /ba/ (0.2-0.5 sec), /da/ (0.7-1.0 sec), /ʄa/ (1.1-1.4 sec) and /ga/ (1.7-1.95 sec). On its turn F_2 ascends from 1000 Hz to 1400 Hz (CF) for /ba/, and descends from 1800 Hz to 1400 Hz for /da/, /ʄa/ and /ga/ although at a different rate, which for /ʄa/ is the steepest one. Upper blips are clearly

[2] The term "semantics" is used here as the study of symbols and their combinations concerning meaning. Therefore a "sematom" would be any minimal perceivable sound pattern capable of conveying meaning *per se* under a phonetic point of view. The role played by a specific "sematom" in a specific linguistic system would become a phonological issue, and will not be treated here.

observed in this last consonant (1.22 sec) at a frequency of 2500 Hz and 3400 Hz, extending to 3800 Hz as a noise burst (NB). Nasal blips around 200 Hz are also perceptible in all four cases, being lower for /ba/ and /ɟa/ than for /da/ and /ga/. Similar patterns may be observed for the unvoiced consonants /pa/, /ta/, /ca/ and /ka/, as shown in Figure 6.

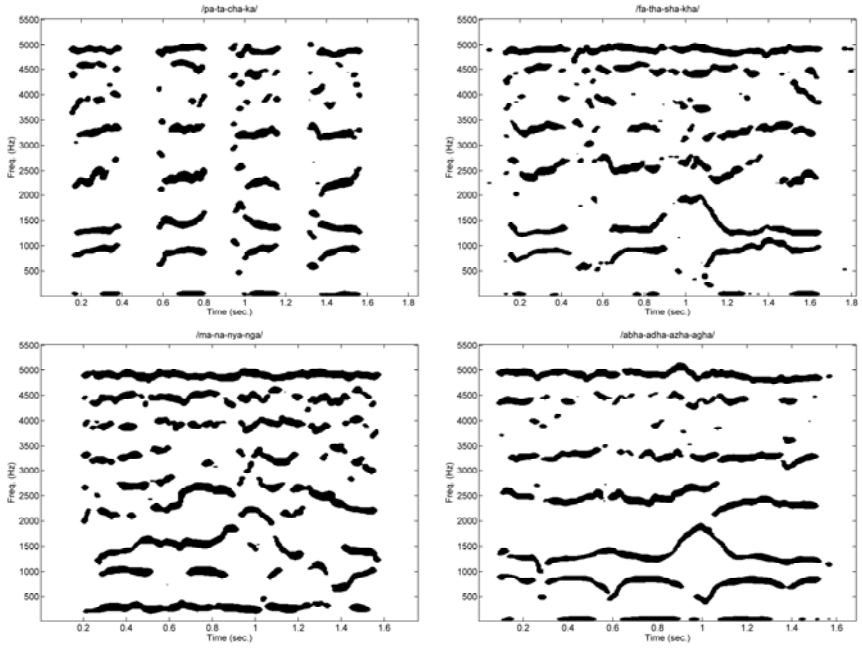

Figure 6. Top left: Formant plots of the syllables /pa–ta–ca–ka/ from the same speaker. Top right: Idem of /fa–θa–ʃa–χa/. Bottom left: Idem of /ma–na–ɲa–ŋa/. Bottom right: Idem of the V-C-V groups /aβaðaʒaɣa/.

Having in mind these observations a Generalized Phoneme Description integrated by "sematoms" of different nature may be abstracted as shown in Figure 7 (Top left) where the temporal patterns of a typical phoneme are shown based on the nuclear vowel system and the virtual pre-onset and post-decay positions. The description is based on a vowel nucleus defined by formants F_1 and F_2. The onset is marked by formant F_1 moving from a specific locus (L_{11}) to the final CF position (positive FM). The formant F_2 may move from a low frequency locus (L_{21}) (positive FM) or from high frequency ones (L_{24}, L_{25}) (negative FM) depending on the specific articulation place of the frontal consonant. Blips appear mainly in palatal articulations, and extend to wide-band patterns (with frequencies above 3000 Hz). Loci in the decay side evolve to next vowel or consonant articulation places. Nasalization appears as a low formant F_n which must not be confused

with the glottal formant F_g. The number of formants above F_3 is variable and speaker dependent. If the first two formants were plotted on cartesian coordinates a specific consonantal system would be described by the dynamic trajectory shown in Figure 7 (Top right). Moving from the initial (onset) locus (L_{11}, L_{21}) for /ba/, (L_{11}, L_{24}) for /da/, (L_{11}, L_{25}) for /ɟa/ and (L_{11}, L_{25}) for /ga/ through the position of the vowel (/a/ in this case) ending in the final (decay) locus (CF positions) of the actual vowel or the next vowel or consonant with which it is being co-articulated. The example given in Figure 7 (Bottom) produced with PhonSpec® illustrates this modelling for a real case. The template consists in an F_2 vs F_1 plot of the formant trace given in Figure 6 (Bottom Right) for an utterance of /aβaðaʒaɣa/ with no stops in between. It may be seen clearly how the dynamic behaviour of the sequence is organized around the locus of the vowel (compare positions with the five vowel framework in Figure 4), from where trajectories are fired to each respective locus, which happen to be those of: /ɔ/ for /aβa/, /ə/ for /aða/ and /ɛ/ for /aʒa/. The case of /aɣa/ is rather interesting, as the articulation of this consonant does not seem to differ much with that of the vowel, as signalled in Figure 3 (left) for its unvoiced counterpart /χ/ whose colouring is almost the same than that of /a/, therefore its trajectory is less extended and close to the vowel nucleus. This study, conducted around the vowel nucleus of /a/ can be extended to other vowels as well.

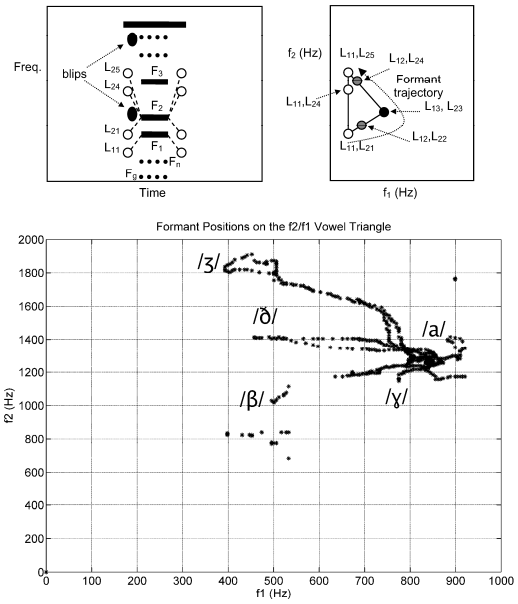

Figure 7. Generalized Phoneme Description. Middle: Loci of the GPD on the vowel triangle. White circles indicate the positions of the loci. The dark dot cluster gives the position of the specific vowel modelled (/a/ in the present case). Bottom: Formant trajectories for the trace /aβaðaʒaɣa/ shown in Figure 6.

Conclusions

Through the present work a review of basic concepts related with Speech Production, Perception and Processing has been presented to define a framework for Time-Frequency Representations of Speech from perceptual and graphical evidence. A structured bottom-up approach from the concepts of "sematom" and the Generalized Phoneme Description using minimal semantic units is shown with graphical examples. These observations open questions which will be the object of future study, as the exploration of higher levels of phonetic class association in hierarchies and their association to Representation Spaces in the Upper Auditory System.

Acknowledgments

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

References

[1] From http://www.arts.gla.ac.uk/IPA/ipachart.html
[2] Cytowic, R. E., "The man who tasted shapes", Abacus, 1993, pg. 178.
[3] Delattre, P., Liberman, A., Cooper, F.: Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Am., Vol. 27 (1955) 769-773.
[4] Fant, G., Theory of Speech Production, Mouton, The Hague, Netherlands (1960).
[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 the Nat. Ac. of Sc., Vol. 99, No. 13, pp. 9021-9025, 2002.
[7] 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 (2005) 441-444.
[8] 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 (1997) 1-10.
[9] Project MAPACI: http://www.mapaci.