INTEGRATION FOR EXTRACTION
What speech perception researchers can learn from Gibson and Marr

Jean-Luc Schwartz, Pierre Escudier

Institut de la Communication Parlée
CNRS UA 368 - INPG / ENSERG - Université Stendhal
INPG, 46 Av. Félix-Viallet, 38031 Grenoble Cedex, France / e-mail : schwartz@saphir.imag.fr

ABSTRACT

We try to make the case that integration mechanisms play a key part in information processing in the auditory system, because of the poor coding abilities of single channels, and the complexitity of the auditory image, but that integration mechanisms must be "clever", able to group in a complex way all the information relevant to a given ecologically relevant source. Gibson's intuitions about the "resonance of neural systems" on the "invariant of the physical environment", later reinforced by Marr's representational framework based on parallelism and specialization of intermediary processes should provide a very strong insight into the study of integration.

1. INTRODUCTION

Auditory processing basically begins by cochlear analysis which results in projecting the incident acoustic signal in one of the 50000 primary Type-I neurons of the auditory nerve. This first neural representation of the acoustic signal – the spatio-temporal pattern of discharges in the auditory nerve – exhibits properties which must strongly determine some of the main processing characteristics of further auditory centers :

(i) there is a great deal of redundancy in the neural outputs,
(ii) which means, in some sense, that there is too much information in the auditory nerve representation ;
(iii) however, a first decomposition of the acoustic signal is achieved, each unit looking more closely at a specific characteristic (the energy in the signal within a given frequency band),
(iv) but it is carried out by noisy channels with a poor ability to "represent" what they are looking for (because of classical limitations of neural cells).

Points (ii) and (iii) are related to one of the key concepts developed by Marr in his theory of vision [19], namely that the input – in his case, the image ; for us, the sound – carries too much information, and, more importantly, types of information that are specified at a number of different scales ; hence this input must be analysed in parallel by a number of different processing systems, which produce different representations, from which specific feature extraction can be achieved (property detectors). The pool of detectors grouped in this level of "intermediary processing" converges, according to Marr, towards what he calls a 2-1/2 D level, a level of "pure perception", which "provides the cornerstone for an overall formulation of the entire vision problem" (pp.269-272), and which corresponds, for the specialist of Cognitive Sciences Petitot, to the "morphological level" of the "phenophysics" (phenomenologically significant physics providing the ecological objective information about the real world) [23].

Points (i) and (iv) largely constraint what should be the key information processing mechanism for the elaboration of intermediary representations, namely *integration mechanisms*, able to build up the necessary *statistics* of the noisy poor-coder channels (iv), taking as best profit as possible of the *redundancy* of the auditory nerve representation (i).

Obviously, however, integration cannot be a kind of "smoothing" mechanism that would just give a gross insight into the content of the auditory nerve – it cannot compute a simple first-order statistics – but it must on the contrary be able to group in a complex way all what is relevant to a given feature, and filter out what is not relevant : it must be able to *extract* and *reveal*. This is highly reminiscent of Gibson's view about the "resonance on the ecological invariant" across the variability of the stimulus [8]. As Marr clearly shows, Gibson's brilliant intuition needs a computational framework ; integration could provide one of the cornerstones of this framework, considering that resonance can happen only if the resonant system has found time or space enough in order to filter out transients and build up its resonant behaviour.

We shall discuss here two possible examples of *clever* integration processes, which could help to recover important articulatory manoeuvers from auditory representations, i.e. (i) enable detection of acoustic events controlled in the timing of speech production, or (ii) extract a stable articulatory target "hidden" inside a continuously varying acoustic trajectory.

2. INTEGRATION ACROSS PLACE FOR EVENT DETECTION

Auditory perception needs a good instrument for estimations of temporal relationships between acoustic events. This is true for any ecological situation where timing provides key information on the structure of the objects that produced the sound, and specially for speech perception where the temporal organization of the glottal and supra-glottal gestures is finely controlled by the speaker in order to produce such phonological contrasts as voiced vs non-voiced plosives, simple vs double consonants (gemination), or tense vs lax vowels.

Timing estimations obviously begin with a system able to *detect* acoustic events. The cochlear nucleus seems able to provide a biological equipment for such event detection, with its various kinds of "on" cells [30]. In our laboratory, Wu [28] proposed a simulation of "on" cells based on a model of neural adaptation that he had shown to produce good results at the level of primary neurons in the auditory nerve. At the output of each individual cell of a model of the auditory nerve, a strongly reinforced adaptation mechanism produces the "on" behaviour, with a very large component of temporal derivation, followed by a rather long (100 ms) forward masking effect. We showed that such a model can indeed cope with the acoustic consequences of glottal or supra-glottal articulatory events such as beginning of voicing, beginning of a vocalic state of the vocal tract, beginning of friction [29].

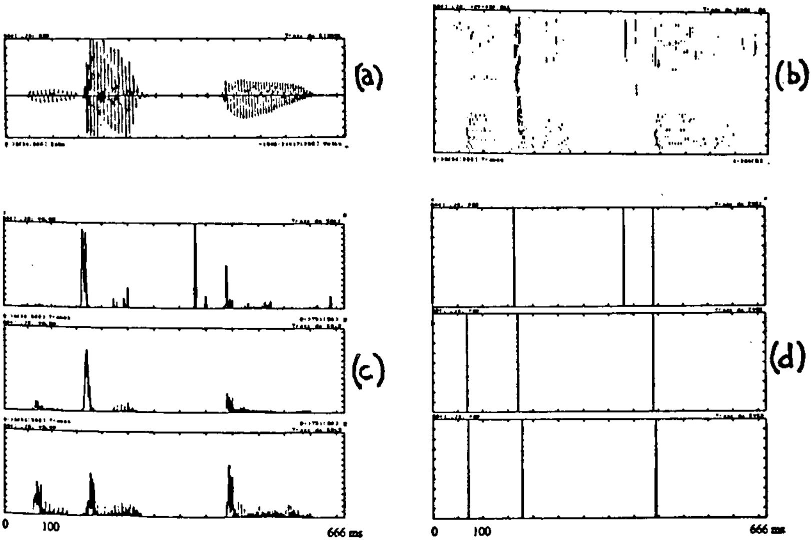

FIGURE 1 - "On" cells modelling and detection of acoustic events
(a) French logatome [baki] (b) response of 64 "on" cells (Characteristic Frequency CF increase from bottom to top) (c) from bottom to top : integrated responses in the low (100-300 Hz), middle (300-900 Hz) and high (900-4000 Hz) CF-regions (d) detected events in the three corresponding CF-regions (see [28] for a precise description of the detection algorithm).

On Fig.1 we show the modelled response of an array of "on"-cells to a French logatome [baki]. The events appear as the *synchronized* occurence of a strong maximum of discharge in a sub-array of cells covering a broad frequency range.

Event detection is based on a *summed response* in one of three sub-arrays, respectively in the low-, middle- or high-frequency region (Fig.1c, 1d). Hence, the functioning of the model relies on (i) strong responses to acoustic events in single frequency channels, due to reinforced adaptation, and (ii) summation of responses based on synchrony of behaviour in a number of channels. Summation of synchronized responses seems to be a very efficient way of signalling events localized in time, with a rather efficient behaviour in noise [1] : indeed in this case the response driven by an event in one frequency channel can be confounded with an amplitude fluctuation inherent to the noise, but, since the acoustic event cannot have a good temporal localization (small time spreading) without covering several frequency bands, it results in *synchronized fluctuations in a number of neighbour channels* : this helps disambiguate the "signal" (the acoustic event) from the noise.

The proposal of mechanisms for detection of coherence of response in a number of frequency channels is in line with an increasing number of psychoacoustic studies about signal detection in noise which show that a signal is easier to detect if inter-channel synchronies help separate the signal from the inherent noise fluctuations, and more generally if inter-channel coherences of the signal and the noise are as different as possible. Thus :

(i) if a noise band masking a pure tone signal is flanked by noise bands in temporal coherence with it, these adjacent bands improve the knowledge of the inherent noise fluctuations and increase the pure tone detectability (Comodulation Masking Release [11]);

(ii) coherent noise bands flanking a signal made of another noise band create less masking on the signal if their fluctuations are uncorrelated with those of the signal than if masker and signal band noises are coherent (Comodulation Detection Differences [18]);

(iii) a temporal gap applied in synchrony on several band noises is easier to detect if the band noises are not coherent, since the gap induces a coherent amplitude fluctuation (a local amplitude minimum) which emerges more when superimposed on uncoherent noise fluctuation patterns [10] ;

(iv) temporal gaps applied on several band noises are easier to detect if they are synchronized than if they are not [9].

All these data show, as Grose & Hall [10] notice, that "the auditory system is able to combine information across critical bands in order to improve temporal acuity" (pp.312), which "might act to facilitate a segregation between a target event (signal or gap) and the background pattern" (pp.313). *Hence, in the case of event detection – a key task for "ecological acoustics" – Gibson's "resonance on the invariant" should involve integration mechanisms finely tuned in time – probably after a stage of reinforcment of the response to temporal variations – and widely spread in place, which is exactly what is realized in Fig.1.* Or, put in Marr's framework, detection of temporal discontinuities – probably crucial for accessing to a possible morphological level equivalent to the vision 2-1/2 D level – should involve neurons owning specific "receptive fields" (see [12] for an extensive use of this concept in audition), which implement a derivation in time and an integration in frequency.

Notice finally that the concept of coincidence of neural discharges could be very general and seems to provide a powerful principle for neural integration in auditory processing [5].

3. RECOVERY OF NON-REACHED TARGETS BY INTEGRATION ALONG ACOUSTIC TRAJECTORIES

After the famous paper by Strange et al. [26] showing that vowels in consonantal context were better identified than isolated ones, a number of contradictory results about static vs dynamic specification of vowel identity were published in the next years : results of several studies from the Haskins Laboratories were in favor of the role of dynamic information, while a number of authors failed to replicate these data and obtained high identification performances for isolated vowels, sometimes better than with vowels in context, and never much lower [27].

Whatever the experimental details or linguistic conditions that could account for the divergence, it remains that formant variability of vocalic targets in function of consonantal or vocalic context, rate and speech conditions is a well-known and classical fact [13, 16, 20] which must be accounted for by human or machine vowel identification systems. Thus we obtained in our laboratory as large as 400 Hz F1- and 1000 Hz F2-variations for [a] and respectively 300 Hz and 600 Hz for F1 and F2 variations for [ε] in contexts [iVi] for a French speaker, depending on rate or focus conditions. This speaker displayed values going from (F1 = 800 Hz, F2 = 1250 Hz) for [a] and (F1 = 650 Hz, F2 = 1800 Hz) for [ε] in the best conditions (reached target) to (F1 = 400 Hz, F2 = 2250 Hz) for [a] and (F1 = 350 Hz, F2 = 2400 Hz) for [ε] in the worst ones (quick rate, no focus on the central vowel), while perceptual tests showed that listeners could easily recognize [a] in all cases, though [ε] was more difficult to recognize in the worst conditions.

Listeners' ability to integrate these variations was assessed in several studies [7, 17, 25] and a number of recognition systems incorporate such contextual effects by implicit rules, as in HMMs or MLPs, or explicit modelizations [2, 14].

In this framework it is important to remember that, in echo to Lindblom's initial considerations about modelling vowel reduction by a first-order attractor [16], all present models of speech production introduce at some level the crucial concept of *dynamic systems driven towards an attractor* for the specification of the gesture from one target to the other. Thus the pilot work from the Haskins Laboratories is based on second order dynamics in a target space, from which the articulator dynamics are "backwards" specified by classical principles of inverse dynamics [24]. The work in progress in our laboratory involves a low level of second-order articulatory gestures (mass-spring stiffness driven model [21]) controlled at a higher level by complex optimization principles implemented in a non-linear dynamic system [3]. The important point is that *as soon as a gesture is defined as a dynamic system with the target as an attractor, the equation that specifies the dynamic system provides a constant link between cinematic articulatory variables (such as position x and speed $\dot{x}$) and a limited set of control parameters with values fixed all along the gesture, that determine the gesture and hence the attractor (the "target", reached or not).*

Consider for example a linear second-order system with critical damping, completely described by 2 parameters, namely stiffness and equilibrium point. The determination of 3 close values of x and $\dot{x}$ allows the computation of 2 values of $\ddot{x}$, which then gives 2 linear equations involving the 2 control parameters. Resolution of this system provides us with an estimation of the 2 control parameters and, hence, of the target. Then this estimation can be iterated : each new value of (x, $\dot{x}$), in relation with the last two values of the previous set allows a new target estimation.

There are two major difficulties in this approach, one being the "inverse problem" of estimating articulator positions and speeds from formants, the other concerning the existence and exact nature of dynamic systems for speech gestures. However, these problems are not without solutions [5, 15, 22], though they cannot be discussed in detail here.

Our point is the following. As clearly stated by Elman & Mac Clelland [7], *variability in speech signals is lawful.* However, instead of the complex connectionist system they proposed, with connections variable in time as the "predictors" of the acoustic trajectories, what we perhaps need is *a true system of trajectory analysis, performing all along the trajectory an estimation of its target from estimations of formant positions and speeds, and finally, after integration of these estimations for stabilization of the final result and noise filtering, able to make explicit what was the constant target "hidden" in this trajectory, being it reached or not*. This research program, ambitious but not unrealistic and for which powerful tools are now available, could once more provide us with a "clever" integration mechanism perfectly fulfilling Gibson's requirement about the resonance (of the auditory system) on the invariant (of the speech gesture, namely its attractor specified by its dynamic equation).

CONCLUSION

Grouping (or "linking") of neural excitation is becoming one of the key problems for the neurophysiology of perception. It is our deep conviction that progress in the understanding of integration mechanism for speech perception will be possible only in a general framework, where it is acknowledged that (i) speech sounds are *produced*, and must be perceived as *acoustic consequences of articulatory gestures that obey some of the general* *laws of human gestures*, (ii) which requires *specialized processings, focussed on certain types of information*, at specific time-frequency scales, processings based on *elements of the biological information processing toolbox available in the auditory system*. James Gibson and David Marr seem to concentrate in a remarkable way some of the most profound advances in these two major directions.

ACKNOWLEDGMENTS

The framework presented in this paper was largely determined by highly interesting discussions with C. Abry, F. Berthommier, J. Holdsworth and J. Petitot, and based on works from and exchanges with Y. Arrouas, D. Beautemps, R. Laboissière and Z. L. Wu. Many thanks to all of them.

REFERENCES

[1] ARROUAS, Y., et al. (1991), "Représentations auditives de signaux acoustiques", Workshop on "Traitements et représentations du signal de parole", SFA, Le Mans.

[2] ASSMAN, P.F. (1982), "Vowel identification : orthographic, perceptual and acoustic aspects", *J. Acoust. Soc. Am., 71*, 975-989.

[3] BAILLY, G., et al. (1991), "Formant trajectories as audible gestures - An alternative for speech synthesis", to appear in *Journal of Phonetics*, special issue on Speech Synthesis.

[4] BERTHOMMIER, F. (1991), "Auditory processing based on temporal correlation between adjacent spikes trains", to appear in B. Ainsworth (ed.) *Advances in Speech, Hearing and Language Processing (Vol. 3)*, U.K. : JAI Press.

[5] BOE, L.J., et al. (1991), "What geometric variables of the vocal tract are controlled for vowel production ? Some proposals for inversion", *J. of Phonetics*, to appear.

[6] DI BENEDETTO, M.G. (1989), "Frequency and time variations of the first formant: properties relevant to the perception of vowel height", *J. Acoust. Soc. Am., 86*, 67-77.

[7] ELMAN, J.L., Mc CLELLAND, J.L. (1987), "Exploiting lawful variability in the speech wave", in J. S. Perkell and D. H. Klatt (eds.) *Invariance and Variability in Speech Processes* (pp.360-385), Lawrence Erlbaum Asociates.

[8] GIBSON, J. J. (1966), *"The senses considered as perceptual systems"*, New-York, Boston : Houghton-Mifflin.

[9] GREEN, D. M., FORREST, T. G. (1989), "Temporal gaps in noise and sinusoids", *J. Acoust. Soc. Am., 86*, 961-970.

[10] GROSE J.H., HALL, J.W. (1988), "Across-frequency processing in temporal gap detection", in H. Duifhuis et al. (eds.) *Basic Issues in Hearing* (pp.223-231), London : Academic.

[11] HALL, J.W., et al. (1984), "Detection in noise by spectro-temporal pattern analysis", *J. Acoust. Soc. Am.*, 76, 50-56.

[12] HOLDSWORTH, J., et al. (1991), "A multi-representation model for auditory processing of sounds", to appear in 9th Int. Symp. on Hearing.

[13] KOOPMANS - van BEINUM, F. J. (1980), *"Vowel contrast reduction. An acoustic and perceptual study of Dutch vowels in various speech conditions"*, Doct. Thesis, Univ. of Amsterdam.

[14] KUWABARA, H. (1985), "An approach to normalization of coarticulation effects for vowels in connected speech", *J. Acoust. Soc. Am., 77*, 686-694.

[15] LABOISSIERE, R., et al. (1991), "Motor control for speech skills : a connectionist approach", in D. Touretzky et al. (eds.) *Proceedings of the 1990 Connectionist Models Summer School.*, San Mateo, CA : Morgan Kaufmann.

[16] LINDBLOM, B. (1963), "Spectrographic study of vowel reduction", *J. Acoust. Soc. Am., 35*, 1773-1781.

[17] LINDBLOM, B., STUDDERT-KENNEDY, M. (1967), "On the role of formant transitions in vowel recognition", *J. Acoust. Soc. Am., 42*, 830-843.

[18] MAC FADDEN, D. (1987), "Comodulation detection differences using noise-band signals", *J. Acoust. Soc. Am., 81*, 1519-1527.

[19] MARR, D. (1982), "*Vision*", San Francisco : W. H. Freeman and Company.

[20] NORD, L. (1986), "Acoustic study of vowel reduction in Swedish", *STL-QPSR, 4*, 19-36.

[21] PERRIER, P., et al. (1989), "Vers une modélisation des mouvements du dos de la langue", *J. Acoustique, 2*, 69-77.

[22] PERRIER, P., et al. (1991), "Modelling of speech motor control and articulatory trajectories", this Congress.

[23] PETITOT, J. (1990), "Le physique, le morphologique, le symbolique - remarques sur la vision", *Revue de Synthèse, IV/1-2*, 139-183.

[24] SALTZMAN, E. L., MUNHALL, K. G. (1989), "A dynamical approach to gestural patterning in speech production", *Hask. Lab. SR, 99/100*, 38-68.

[25] STEVENS, K. et al. (1966), "Acoustic description of syllabic nuclei : an interpretation in terms of a dynamic model of articulation", *J. Acoust. Soc. Am., 40*, 123-132.

[26] STRANGE, W., et al. (1976), "Consonant environment specifies vowel identity", *J. Acoust. Soc. Am., 60*, 213-222.

[27] WOLF, J. J., KLATT D. H. eds. (1979), *"Speech Communication Papers presented at the 97th Meeting of the Acoust. Soc. of Am."* (pp.15-32), New-York : Acoust. Soc. Am.

[28] WU, Z. L. (1990), *"Peut-on "entendre" des événements articulatoires ? Traitement temporel de la parole dans un modèle du système auditif"*, Doct. Thesis, INP Grenoble.

[29] WU, Z. L., et al. (1991), "Physiologically-plausible modules and detection of articulatory-based acoustic events", to appear in B. Ainsworth (ed.) *Advances in Speech, Hearing and Language Processing (Vol. 3)*, U.K. : JAI Press.

[30] YOUNG, E. D. (1984), "Response characteristics of neurons of the cochlear nuclei", in C. I. Berlin (ed.) *Hearing Science, Recent Advances* (pp. 423-460), San Diego : College Hill Press.