IMPACT OF "ASCENDING SEQUENCE" AI (AUDITORY PRIMARY CORTEX) CELLS ON STOP CONSONANT PERCEPTION.

Eduardo Sá Marta, Luis Vieira de Sá
email: EMARTA@IT.UC.PT
Dep. Engenharia Electrotécnica, FCTUC (Universidade de Coimbra)
Instituto de Telecomunicações - Pólo de Coimbra

ABSTRACT

The existence of multiple information carriers for a single phonemic distinction is well evident in studies of auditory and visual information integration for speech perception. Given the highly non-homogeneous nature of the auditorily-represented information carriers, we are applying the same principle <u>within</u> the auditory domain. Based on psychophysical experiments we have hypothesized that firing of "ascending sequence" cells in the primary auditory cortex is a primary information carrier for LABIAL place in stop-consonant discrimination.

Partial implementation of a fuzzy-logic model for the firing of these cells, combined with a model for one other, secondary, information carrier, has yielded 1% errors in discrimination of /p/ vs. /t/ or /k/ in a "E-set", Portuguese research CV database. Exactly the same partial model, applied to /b/ vs. /d/ discrimination in an American English spelled letters database (ISOLET-1) yielded just 5% errors, providing strong evidence for the role of these cells in stop consonant discrimination across languages.

1 - INTRODUCTION

1.1 - Underlying research assumptions and objectives

The existence of multiple information carriers for a single phonemic distinction is simply a manifestation of a general perceptual strategy: to robustly discriminate between classes of objects (be them auditory, visual , or auditory <u>and</u> visual) the perceiver benefits in considering a small set of features instead of just one, because in some circumstances (e.g. noise, filtering such as that caused by the listener's head orientation) a particular feature might be obscured. In a given exemplar, not all the information carriers (working towards the identification of its correct class) need to be present; it suffices that one or a few be substantially represented, and that information carriers signaling conflicting classes be absent or very weak. One important characteristic of speech perception is the fleeting nature of the objects to be recognized. Phoneme sounds often are present for less than 50ms, and are succeeded rapidly by other phonemes, suggesting that phoneme recognition, as done at a cognitive level, must be a fast (and thus simple, not iterative) process and one heavily reliant on the information elements extracted peripherally; in other words, the perceptual algorithms for each phonemic contrast should have a reasonably simple expression in terms of the outputs of some cells in the auditory pathways. Neurophysiologists have in recent years been making available a wealth of information [1] about the patterns that excite the main types of cells in the auditory periphery, thus making possible an hypothesization endeavor about which of those cells have an impact on each particular phonemic contrast.

1.2 - *Centrifugal* role for the neurally-represented signals

We are currently assuming that signals from the cells, or peripheral processing structures, involved in each phonemic contrast have a *centrifugal* interpretation in perceptual heuristics. That is, strong firing from cells relevant to a particular phonemic contrast causes perception to migrate away from a point of indistinctiveness between the opposed phonemic classes. This is in contrast to a *centripetal* role, in which firing from cells would cause perception to get nearer to a (hypothetical) prototype for the phoneme class of the exemplar.

The main motivation for this assumption arises from the fact that everyday speech contains a high proportion of phonemic segments of low distinctiveness, barely sufficient or even insufficient to allow identification of its corresponding phonemes, where it not for lexical and/or semantical levels of recognition [2]. We speculate that human lexical recognition maintains its high performance in these circumstances because these segments are (correctly) interpreted by listeners as "indistinct" (while maybe still allowing for the identification of broader phonemic classes); that is, in relation to these segments, listeners do not attempt to force a categorical decision in their "phonemic recognition stage" (whatever the degree that this stage exists as a separable process). That is, listeners need also to be able to easily and correctly label many phone-like segments as indistinct. This is obviously easy to do if the neural information carriers have the *centrifugal* role: "indistinct" simply means low firing by all relevant cells. Other motivations, such as robustness to noise- and filtering-induced distortions, are detailed in [3].

1.3 - A new approach for uncovering the neural processing structures impacting phonemic distinctions

Our hypothesization process was also inspired by studies of natural utterances that achieve robustly correct recognition by human listeners in spite of highly unusual (that is, for the phoneme being uttered) acoustical patterns [4]. Achieving correct recognition by listeners implies substantial excitation of <u>at least some</u> of the neural structures working for the perception of the intended phoneme. Given the auditory-feedback basis of speech production acquisition, we might say that the speech patterns of an adult speaker are the ("published") result of a years-long study of phoneme perception by the speaker himself and by the body of listeners he communicated to. Of special interest are those "publications" that present novel data: that correct phoneme perception can be elicited by some original, not common, acoustical pattern.

Given the magnitude of the overall problem of characterizing the neuronal structures for phoneme perception, we focused on a circumscribed problem: that of stop consonant place discrimination under "E-set" conditions (that is, monosyllables in which the stop is followed by the vowel /i/). This has proved a very difficult problem, under speaker-independent conditions, for automatic recognizers - and thus the eventual success in achieving near human-like performance cannot be attributed to the low difficulty of the task. On the other hand, use of the vowel /i/ actually make the problem easier to model in neurophysiological terms. It is almost certain that most of the acoustic cues for stop discrimination, at least for the context of the vowel /i/, are found in the F2 region and above, which for the vowel /i/ means higher than 2K. In this frequency zone, it is agreed that the neurophysiological representation of sound components is based on a simple rate-place representation, thus avoiding all the potential intricacies of temporal and place-temporal models.

Also, we took into account the fact that stop consonant place discrimination is still very good when speakers and listeners belong to different native languages, if both languages have 3-way stop place categorization [5]. To gather acoustical patterns that are widely distinct from those of Portuguese stop-consonant CV's, but are nevertheless robustly identified by Portuguese listeners, we acquired CV's uttered by native speakers of French, Spanish, German, and Flemish.

2. THE NEURALLY - REPRESENTED INFORMATION CARRIERS FOR PLACE DISCRIMINATION IN STOP CONSONANTS

Extensive psychophysical experiments along the approach described above have resulted in the definition of a small number of information carriers for place discrimination; almost all have a simple interpretation in terms of known characteristics of cells in the cochlear nucleus or higher in the auditory pathways. Place discrimination is assumed to be supported by substantially the same mechanisms, be it among voiced or unvoiced stops. That is, information carriers relate to discrimination between LABIAL (/p/ or /b/), DENTAL (/t/ or /d/) and GLOTTAL (/k/ or /g/). Although the bulk of the /i/-set stops confusability is in the LABIAL-DENTAL and DENTAL-GLOTTAL 2-way discriminations, 3-way discrimination must be accounted for.

We have identified, for /i/-set conditions, the following information carriers (these information carriers also apply to CVs with other vowels; there may be slight differences in their perceptual value, however), presented in order of descending perceptual value (for each class):

For LABIAL (against DENTAL or GLOTTAL):

LABIAL-IC1 ascending sequence in the F2/F3 zone

LABIAL-IC2 ascending trajectory of the dominant low-frequency skirt in the F2-F3 zone

LABIAL-IC3 complete or near-complete absence of unvoiced energy prior to vowel onset

LABIAL-IC4 initial brief "vertical bar in the spectrogram" in the absence of the "thin-bandwidth, tone burst-like segments"

For DENTAL (against LABIAL or GLOTTAL):

DENTAL-IC1 initial tone-burst-like segments (>6-8ms) of formant(s) (aspirated or voiced) of very thin bandwidth; or, not exactly initial, but then of longer duration; or at least a very sharp lower-frequency skirt

DENTAL-IC2 initial high-frequency-dominated (>3.5KHz) distribution of energy (grossly in accordance with [6])

For GLOTTAL (against LABIAL or DENTAL):

GLOTTAL-IC1 descending sequence in the F2/F3 zone

GLOTTAL-IC2 abrupt and precocious offset of the initial energy in the F2-F4 zone.

GLOTTAL-IC3 descending trajectory of the dominant low-frequency skirt in the F2-F3 zone

The neurophysiological interpretation of LABIAL-IC1 and GLOTTAL-IC1 recently became clear with the finding that there exist, in the primary auditory cortex of (at least) primates, neurons that respond specifically to ascending sequences and others that respond specifically to descending sequences. It was found [7] that some neurons, while responding weakly to single-frequency tone-bursts, responded strongly to a descending sequence of two different-frequency tone bursts. The tuning to the Æt between the two tone bursts was found to be broad, not requiring a precise value for the time interval.

3 - A PARTIAL MODEL FOR THE RESPONSE OF THE ASCENDING-SEQUENCE CELLS (LABIAL IC1)

The basic assumption is that there are neurons capable of reacting specifically to a BEFORE-LOW -> AFTER-HIGH (BL->AH) sequence.

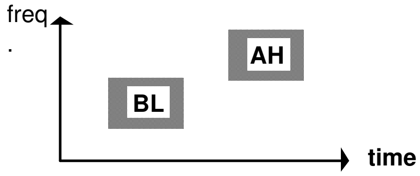

Figure 1 - Ascending sequence

The ascending (2-step) sequence capable of exciting the cell is not characterizable by two fixed-duration time slots. Rather what is suggested (by the psychophysical experiments we have conducted, as well as the neurophysiological studies in [7]) is that the resulting percept is elicited as long as there is detection of a BL event followed by the detection of a AH event. The psychophysical experiments also suggested that the most important characteristic of either event is its onset. The onsets of these two events may be separated by a range of intervals; too short, or too long an interval ceases to be perceptually effective.

In actual speech signals, whenever this sequence exists (if at all) the speech sound components which are candidates for the BL and AH are themselves surrounded by other speech components. Even in simple stop-consonant/vowel monosyllables (where the BL, if it exists, must be located at the start of the sound - in the initial unvoiced segment, or at the exact beginning of the vowel segment), it must be considered that significant energy at higher frequencies might exist preceding it, or simultaneously. This will plausibly degrade the perceptual adequacy of the BL candidate, but some measure of this high frequency energy must be tolerated. Likewise, some lower frequency energy may exist in the AFTER phase, at even lower frequencies than the best candidate for BL in the sound under study; again, some degree of (graded) tolerance for this must be considered. So, in speech sounds the components of the ascending sequence may be degraded in their "goodness" (for causing the firing of the cells) by various factors. Indeed, a contrary (descending) sequence (construed when a AFTER-LOW component appears that is actually lower in frequency than the candidate for BEFORE-LOW) might even exist to varying degrees.

We considered the following factors:

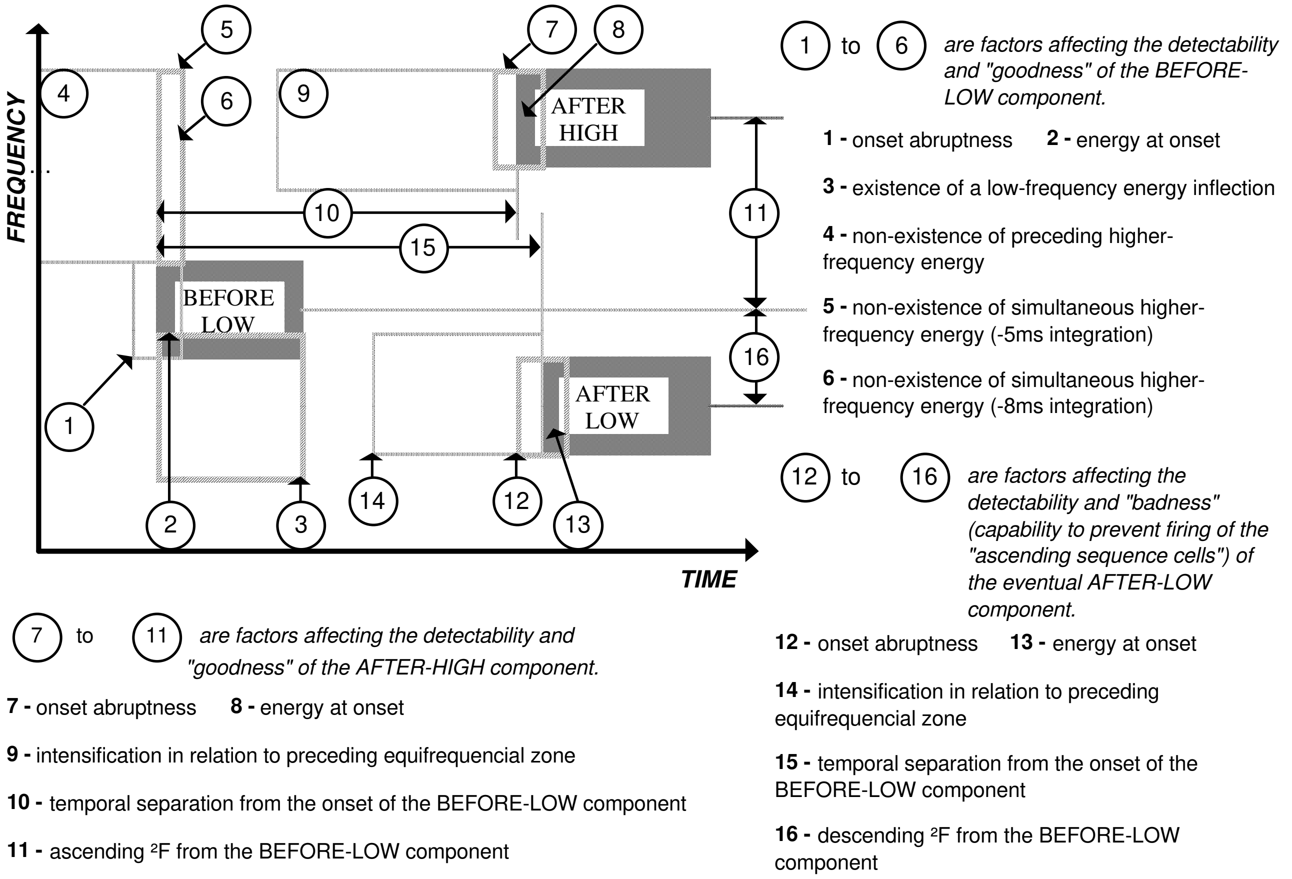

Figure 2 - Intersective factors for LABIAL IC1

The impact of each of these factors on the firing of the "ascending sequence cells" was expressed by a *fuzzy membership function* of the simple Γ type (for factors that propitiate firing) or L type (for factors that tend to prevent firing). The parameters **L1** and **L2** for each membership function were estimated by psychophysical experiments with edited natural sounds in which the underlying physical parameter was changed through filtering, frequency-shifting, and temporal splicing operations.

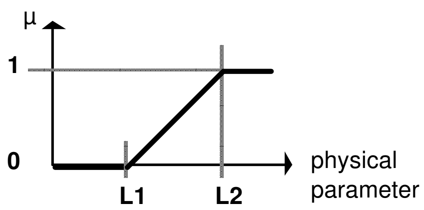

Figure 3 - Γ-type function

The set of membership functions was subject to fuzzy intersection to yield a final membership function for the estimated possibility of "substantial firing rate" by the cells. Intersection was optimistic (minimum) in the cases where it was judged that factors were not cumulative in their effect towards preventing excitation of the cells, and pessimistic (product) in the cases where it was thought to exist a cumulative effect.

4 - EXPERIMENTS AND RESULTS

For the experiments we combined the results from implementations of LABIAL IC1 (as described in the preceding section) with LABIAL IC3 (much simpler, and not described here). At first thought, it should not be expected (this being a very partial implementation of the perceptual structure that we characterized as underlying /pi/-/ti/-/ki/ discrimination) a very high discriminative score of /pi/ against /ti/ or /ki/. For instance, a particular exemplar of /pi/ might be "perceptually construed" resorting mostly to LABIAL IC2 or LABIAL IC4 (which are not covered in the present partial implementation), and not so much to LABIAL IC1 and LABIAL IC3; or, a speaker uttering /ti/ might indulge a weak degree of LABIAL IC1, but supplanting it with a strong presence of DENTAL IC1 (not covered here, either).

We reasoned that by being very optimistic about LABIAL IC1 we might capture even exemplars of /pi/ which resort primarily to LABIAL IC2 or LABIAL IC4. This would be so because it is plausible that articulatory imprecision would be an important cause of some exemplars (of a given class) ending up by being "perceptually reliant" on a more secondary information carrier; thus, even in those cases, the acoustical pattern will be close to exciting the cells reacting to the more primary information carrier, and should be captured by a (unrealistically) optimistic implementation of these cells' behavior. The optimistic implementation was obtained by not checking for the existence and adequacy of the AFTER-HIGH event; however, if the BEFORE-LOW event occurs before vowel onset, it normally meets a AFTER-HIGH event caused by excitation of F3 and higher formants at vowel onset.

On the other hand, the possibility of articulatory imprecision should motivate speakers uttering non-labial stops to ensure a margin of safety against (accidental) excitation of (cells responding to) the primary information carriers working towards LABIAL.

The sequence of experiments was as follows:

(i) - We refined parameters for the fuzzy membership functions using a set of /pi/, /ti/ and /ki/ utterances by 33 Portuguese speakers. A threshold was defined which resulted in 2% errors in discriminating /pi/ vs. /ti/ or /ki/.

(ii) - We applied the resulting models to /bi/-/di/ discrimination in the 1st set of 30 (American English) speakers in the CSLU/OGI ISOLET database. It emerged a suggestion for a refinement: since the cells responding to ascending sequence are influenced by 2 events separate in time, it is to be expected that they somehow use the outputs of other, more peripheral, cells which are adequate for detection of event onset. Onset cells in the cochlear nucleus are known to respond preferentially to wide-bandwidth stimuli, and some respond very weakly to tones. The refinement we introduced was then to suppress the contribution of very thin-bandwidth spectral peaks to candidates for the BEFORE-LOW event. The resulting error rate dropped to 5%, and the remaining errors were individually examined and all were found to be explainable in terms of information carriers not covered in the present partial implementation.

(iii) The refined model was applied again to the 33 Portuguese speakers -set. Interestingly, the error rate halved to 1.01%

It is enlightening to examine histograms for the final fuzzy membership function (μ) for LABIAL IC1:

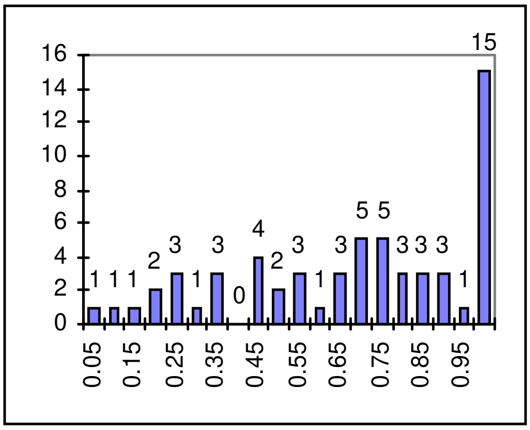

Figure4 Histogram of μ-LABIAL IC1 for ***/bi/*** *in ISOLET-1*

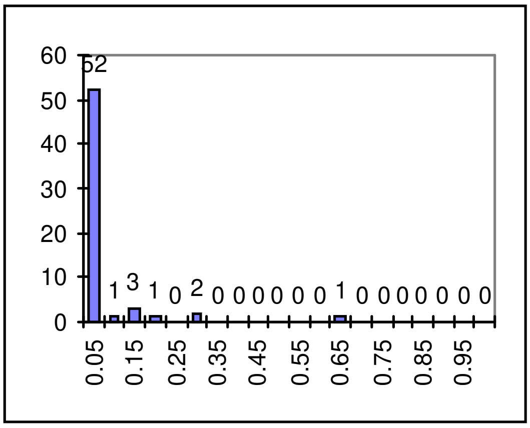

Figure 5- Histogram of μ-LABIAL IC1 for ***/di/*** *in ISOLET-1*

It may be observed that for **/bi/**, while a good number of the exemplars show μ approaching 1, many are sparsely distributed over the whole range. This, in our view, reflects the fact that speakers are not forced to "build the perception" of their LABIAL stops exclusively in terms of LABIAL IC1; they may resort to LABIAL IC2, 3 and 4. In contrast, the histogram for **/di/** shows almost exclusively near zero values (the one single exemplar with a substantial value, in the 0.65 bin, might raise questions; but this sound was visually inspected and found to contain no AFTER-HIGH event …which, as previously explained, was not checked for). This, in our view, reflects the fact that when producing a given phoneme, speakers have to ensure that they do not excite, to any significant degree, the cells whose output would signal a confusable phoneme.

A final mention: the Portuguese set of utterances was recorded with 4 different microphones; the ISOLET set was obviously recorded with another different microphone.

5 - CONCLUSIONS

The ultimate scientific description of human phoneme perception should be made in terms of biologically plausible processing structures and should be able to replicate such human capacities as

(a) - correctly discriminating across some phonemic contrasts in other languages than the listener's native one

(b) - being largely insensitive to such filtering as caused by different recording microphones

(c) - correctly estimate the "distinctiveness" of phonemic segments (this ensures that the listeners almost never make hard errors in phoneme identification)

The results reported in this paper supply evidence for the adequacy of a novel approach in achieving **(a)** and **(b)**. Although not supported by these results, this approach is also thought to be able to achieve **(c)**.

REFERENCES

[1] - Chapter 4. "Physiology of speech processing" (Shihab Shamma and Alan Palmer). In a forthcoming volume in the Springer Handbook of Auditory Research series, entitled "Speech Processing in the Auditory System"

[2] - Björn Lindblom, "Role of articulation in speech perception: clues from production", J. Acoust. Soc. Am. 99 (3), March 1996, 1683-1692

[3] - Eduardo Sá Marta, Fernando Perdigão, Luis Vieira de Sá - "Psychophysical evidence for a sampling process, related to properties of onset cells, in stop consonant perception".- Proceedings of the ESCA Workshop on the Auditory Basis of Speech Perception, Keele University (UK) - 1996

[4] - Eduardo Sá Marta, Fernando Perdigão, Luis Vieira de Sá - "Researching the processing structures of human phoneme recognition by analysis of natural stop-consonant-vowel utterances that elicit correct recognition through unusual acoustic patterns" - *4th European Conference on Speech Communication and Technology Eurospeech 95* - Madrid, Spain, 1995

[5] - Anna Marie Schmidt, "Cross-language identification of consonants. Pat 1. Korean perception of English, J. Acoust. Soc. Am. 99 (5), May 1996, 3201-3211

[6] - Stevens, K.N., and Blumstein, S. E. - "Invariant cues for place of articulation in stop consonants", J. Acoust. Soc. Am. 64, 1978, 1358-1368

[7] - H. Riquimaroux, "Processing of sound sequence in the auditory cortex" - Proceedings of the ESCA Workshop on the Auditory Basis of Speech Perception, Keele University (UK) - 1996