Mechanisms underlying human phoneme communication

Eduardo Marta
Institute for Telecommunications – Coimbra Pole

Even for simple tasks such as spelled letter communication, human performance vastly outdoes the best task-trained automatic recognizers.

< 1/10 the error scores of recognizers trained and tested on (different parts of) the same database

In this presentation…

1 – a review of human phoneme communication capabilities and "feats"

2 – a bundle of hypotheses, related to natural evolution concepts, aiming to explain these capabilities

3 – working out the hypotheses for the case of communication of stop consonant 3-way *place* of articulation

…always bearing in mind simple non-lexical communication tasks such as spelled letter communication

State of the art for spelled letter recognition

E-set (B C D E G P T V Z …*just 9 letters*) discrimination, over the telephone: approximately **30%** errors for intra-base testing

1) human phoneme communication capabilities

- **Capability#1 – "insensitivity" to inter-speaker variability** Some speakers produce extremely unusual acoustic forms and still elicit consistent recognition from listeners; even non-native speakers are tolerated (specially for stop consonants).

- (**Capability#2 – insensitivity to non-drastic filtering**) Independence relative to non-drastic changes in the frequency-gain curve (e.g. microphones, or listener head orientation); also, listeners suffer fluctuations in their audiograms. Lumping all these variabilities at the listener, we conclude that phoneme communication is robust against substantial parameter variation at the receiver.

- (**Capability#3 – insensitivity to articulatory imprecision**) Speakers are affected by articulatory imprecision, but only very rarely does this hinder phoneme communication (in most such instances the speaker himself acknowledges the error and repeats the utterance).

- (**Capability#4 – graceful degradation for drastic filtering**) Graceful degradation when going from full-band speech to band-pass (e.g., telephone-like) speech.

- (**Capability#5 - humans "know when they don't know"**) Adequate confidence scores: human classification errors are accompanied by low confidence.

- (**Capability#6 – recognition without previous training**) Human listeners don't need to be trained on the same speech database used for testing; there were never any reports of listeners needing "auditory training" to use the telephone for the first time in their lives (and they would be totally lacking category prototypes for telephone speech). Cross-database testing of automatic recognizers causes drastically reduced performance.

- (**Capability#7 – speaker flexibility**) Speakers want to accommodate articulatory confort, indulge articulatory variability induced by various reasons (the conveyance of a personal speaking style, emotional status, ...),... ...the speaker might even have some articulatory handicap (e.g., smoking a pipe). Speakers need flexibility: to have a "space" of options in realizing each phoneme.

Much of this builds up the well-known **paradox of constancy of (phonemic) perception in spite of acoustical diversity**.

This perceptual constancy has proved hard to explain; one (still) popular attempt at explanation – the *motor theory* – invokes extraordinary capabilities on the part of human listeners, such as that of being able to track the intended articulatory gestures of the speaker.

2 - A bundle of hypotheses linked to natural evolution concepts:

Hyp.1 ***(redundant information carriers)*** - Languages have tended to **select** phonemic contrasts that are rich in auditorily-salient features (or *information carriers* - ***IC***s). **For each phonemic contrast there exist several *IC*s that are redundant** if used concurrently. If some *IC*s are degraded the surviving *IC*s will ensure correct communication. Each *IC* is independently evaluated.

Hyp.2 ***(discriminatory role of ICs)*** - These auditory *IC*s are **discriminatory**, that is, they "register" away from between-categories boundaries. There exist extensive trade-offs - to the point of alternativity – between the "agreeing" *IC*s.

Hyp.3 ***(speakers unconsciously exploit redundancy/alternativity to "take liberties" in production)*** – During speech production acquisition by a new speaker (a child), once gross articulatory correctness (for a given phoneme) is achieved, **auditory feedback becomes the only significant "evolutionary force"** and this auditory feedback incorporates trading relations between *IC*s. The new speaker may "rest satisfied" when he/she achieves strong emission of one of the *IC*s, thereby allowing relaxation in the emission of the other *IC*s. **Different speakers may end up with very different mixtures of *IC*s**, each of them yielding successful communication.

> ***This is reminiscent of the evolutionists' classic example of the panda's thumb. The panda is a bear (and bears have 5 aligned fingers) yet still achieves a grasping function… …but through a solution with no morphological conformity to the prevalent solution for that function. Instead, the panda has evolved a <u>false</u> thumb, which is really a wrist bone that has over millennia grown extraordinarily and now operates as a (rigid) thumb (FUNCTIONAL SUCCESS WITHOUT MORPHOLOGICAL CONFORMITY)***

Hyp.4 ***(prevalence of hard-wired ICs)*** – The *IC*s are often quasi-direct expressions of the metrics computed by some **specialized cells** that evolved in lower animals to enhance survivability. The tens of thousands of years of existence of complex languages can not have changed these cells, and so they are effectively **hard-wired**.

Human listeners performing phonemic discriminations are forced to resort primarily to single-pass, distributed processing embodied in peripheral auditory cells, because of the few tens of milliseconds available for processing the shorter phones.

> ***This corresponds to another concept from natural evolution: exaption, that is, the "seizing" by a new function (phoneme communication) of biological mechanisms that evolved previously as adaptations to other tasks (such as a basic survival-enhancing acoustic detection ability).***

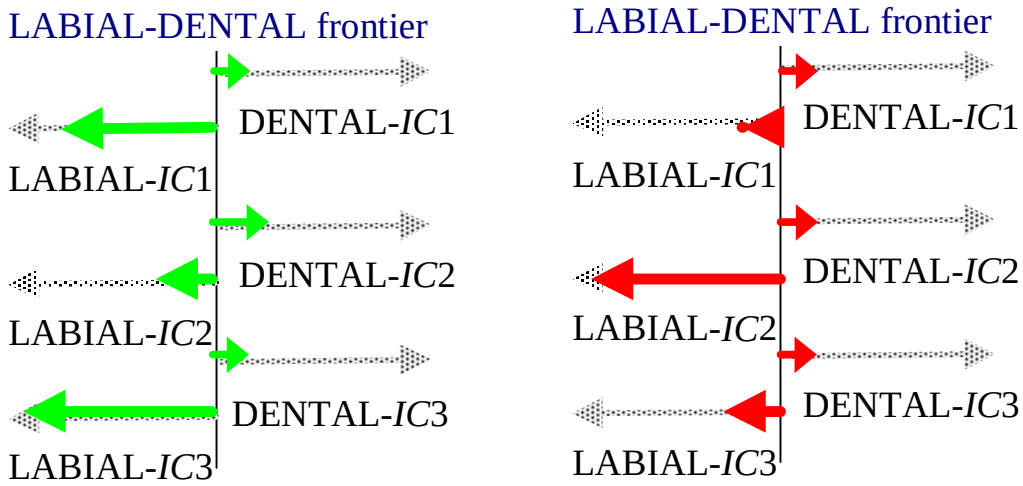
Hyp.5 ***(insensitivity to filtering)*** – Most of the auditory *IC*s are **inherently insensitive to non-drastic filtering** changes such as those caused by different listener head orientations. Indeed, this was already necessary for the hard-wired cells that provided crucial detection abilities for lower animals.

The explanatory power of this bundle of hypotheses

In the light of **Hyp.2 (discriminatory role of *ICs*, alternativity)** and **Hyp.3 (speakers exploiting this alternativity)**, the constancy of perception in spite of acoustical diversity starts to seem **much less of a paradox**.

Suppose for a moment that there exist 3 *IC*s working for the perception of DENTAL stop consonants, and also 3 for LABIAL stops (the number of *IC*s for competing categories is not necessarily equal). Labial stops from two speakers might appear as follows:

Labial stops from speaker GREEN... ...and from speaker RED

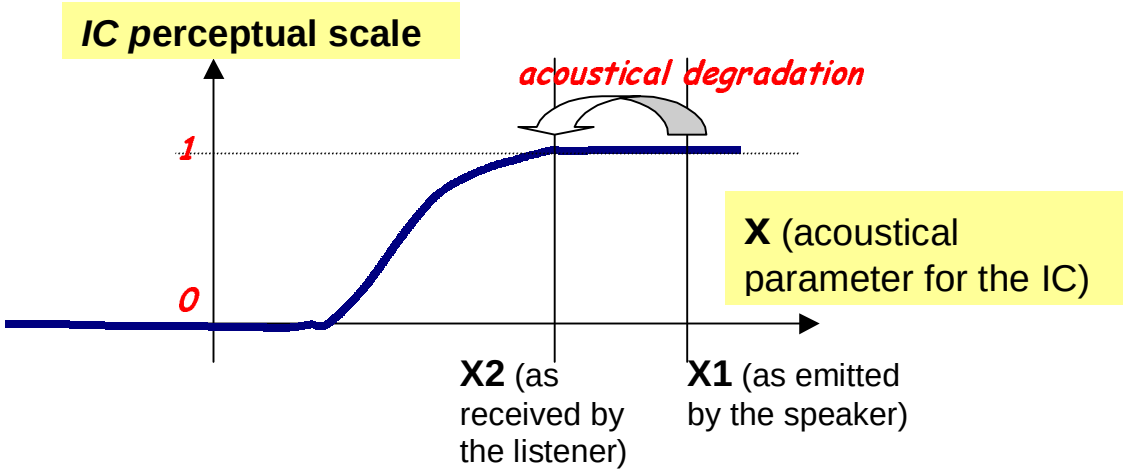


That is, given that there are **trade-offs between "agreeing" *IC*s**, **speakers have a space of options** (all of them perceptually convincing) for the production of a particular phoneme.

In fact, the ability to use different mixtures of "agreeing" *IC*s is just one level of the space of options available to the speaker. The acoustic formulation of the metric computed by each of the *IC*s also provides a space of acoustic forms that map into the same auditorily-computed score. This is so because these metrics always include one (or more) **sigmoid-like response curve**.

This brings us to partially-explaining **Capability#2 (insensitivity to filtering)** and **Capability#3 (insensitivity to articulatory imprecision)**. Speakers can **oversaturate the neural evaluators** for the *IC*s they are using. If degradation occurs, because of channel imperfections or articulatory imprecision, the degraded form will still project into a perceptual scale value of **1** (or into a **0**).

*Suppose that the speaker has emitted parameter **X** (for a given IC) at the value of **X1**, but degradation has caused the received value to be **X2**.*
***X2** still maps into a perceptual score of **1**.*

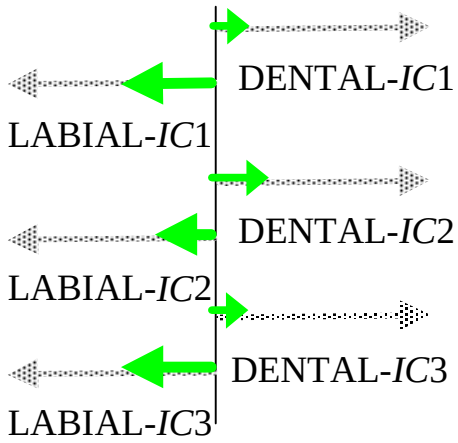


Most of the rest of explanation of **Capability#2** lies in **Hyp.5 (the *IC*s are inherently insensitive to non-drastic filtering)**. Specific details on some instances of **Hyp.5** will be seen later, in relation to most of the *IC*s relevant for stop PLACE discrimination.

As for **Capability#4 (graceful degradation for band-pass speech)**, its explanation lies in that **some (but not all) of the *IC*s are almost as well "excited" by band-pass speech** as by full-band speech (again, we will see instances of this).

Capability#5 ("knowing when you don't know" at the phoneme level) is crucial in continuous speech communication, where it happens frequently that the speech segment corresponding to a given phoneme does not actually carry enough information to permit its "stand-alone" recognition… …this has been termed as hypo articulation (Bjorn Lindblom's H&H theory). This raises the need for **phoneme recognition mechanisms that can also handle hypo-articulated instances… …yielding "hypo-recognitions" as the really useful result**.

According to **Hyp.2**, a hypo-articulated phone would be transcribed by low scores in all *IC*s (slightly larger in some of the "correct" *IC*s). An hypo-articulated labial stop might look like:



Adding some channel degradation… … the listener will have very little evidence to base discrimination… …he may misrecognize. But the listener will be able to evaluate this situation as **"all *IC*s have low scores"**.

The worst-case outcome will then be a "weak error" (one which will be easily bridged over by lexical or semantical information).

Conventional, class-prototype-based recognition would often yield confident (and damaging) errors.

Capability#6 ("no need for training with the same database") might at first thought not seem relevant to human phonemic communication. It just might be assumed that human listeners have had access, by the time they are adults, to an extremely huge database. But, at least for some phonetic distinctions (such as stop consonant PLACE), **human listeners recognize well acoustic forms that they are not used to**, namely those from non-native speakers. Also, imagine a person first using the telephone only in his adult years… …how could he dare to, with no class-prototypes for telephone speech ?

It would seem instead that **there is a "snappy" quality to the acquisition of phoneme recognition ability** (by a child), at least for some phonemic discriminations (such as stop PLACE). Stops are much briefer sounds than continuants, and thus are more likely to use specialized peripheral cells to capture their discriminatory features. Thus, perception of stops is likely to be much more hard-wired than that of vowels, and **perception acquisition involves mostly the cognitive "snapping" onto the signals from the hard-wired structures**.

3) - Working out the explanation for communication of stop consonant 3-way "place". Models for (neural) auditory features underlying human communication of this discrimination

Stop "PLACE" perception *(wrong problem ! …the right problem is that of "PLACE" communication)* has been a popular research problem.

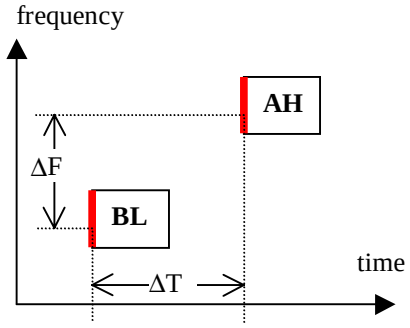
It represents the epitome of the paradox of constancy of perception.

Seemingly, none of the existing proposed accounts has given origin to a computational model that can be compared to human performance.

Additionally, the human capability of insensitivity to variations in the frequency-response curve has not been addressed. As will be seen shortly, a robust explanation for this capability can be construed in terms of acoustic metrics computed by several types of auditory cells that have been researched (by neurophysiologists) in animals. **It is remarkable that this insensitivity is attained without any sort of compensation for channel characteristics**.

IC1 for the LABIAL category: ascending-sequence cells

Ascending sequence cells have been shown (in animal studies) to exist in the auditory cortex of mammals; they react to sequences such as

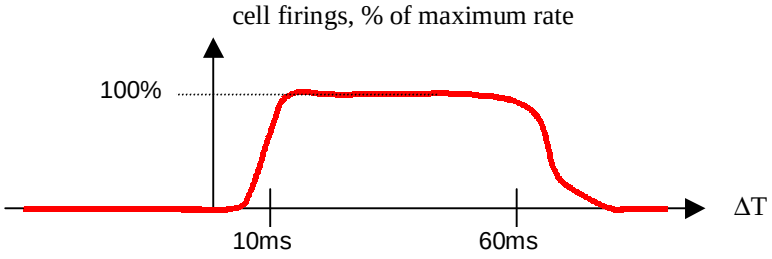


…but they do not react to either of the 2 events (**BL** = **B**efore-**L**ow **;** **AH** = **A**fter-**H**igh) presented alone.

The animal studies are very sparse (one reason being that animals must not be deeply anesthetized) and have given only very sketchy details.

We have ventured some assumptions, mainly reasoning that the detection of the two components is done by more peripheral cells adequate for onset detection, and that these react mostly to temporally-abrupt and frequencially-wide onsets (thin-bandwidth components might be *stealthy* relative to sequence cells).

The most important part of each of the two events is its onset. The **dependence on ΔT and on ΔF is broadly trapezoidal**: for example, ΔT values anywhere between 10ms and 60 milliseconds would "work" nearly the same.

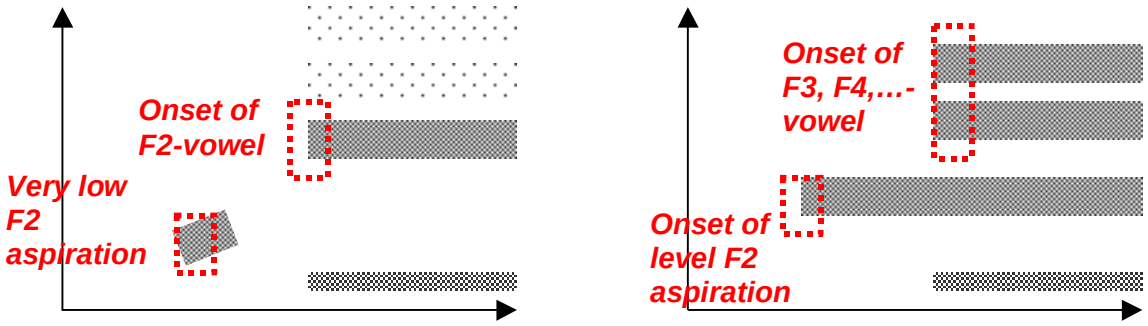


We can already note at this point:
1) insensitivity to filtering (e.g., different microphones) arises from the fact that as long as none of the 2 components is totally obliterated the sequence will still exist
2) little articulatory precision is required because of the broad trapezoidal dependencies on ΔT and on ΔF.

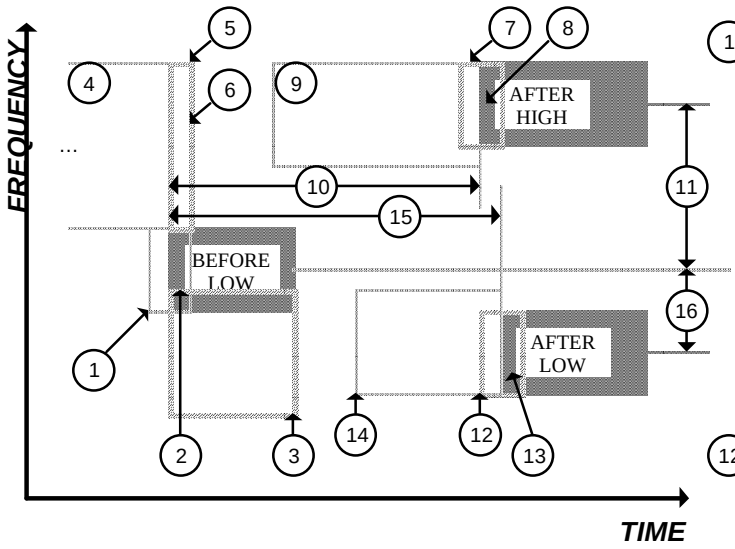
We found that **ascending-sequence cells have a primary role in discriminating the LABIAL category** against the confusable ALVEOLAR/DENTAL and GLOTTAL/ALVEOLAR categories):

1) ascending-sequence patterns occur in most (but not all) exemplars of LABIAL stops; they do not occur (with very rare, and explainable, exceptions) in ALVEOLAR/DENTAL or GLOTTAL/ALVEOLAR stops
2) in most instances, editing the sounds of non-LABIAL stops to force an ascending pattern will cause perceptual migration towards LABIAL

Acoustically, the ascending sequence patterns may have different descriptions in terms of the formants; all these forms are equivalent from the point of view of the cells:



We developed a simple fuzzy-logic model for the firing of these ascending-sequence cells and applied it to the discrimination of LABIAL against ALVEOLAR/DENTAL and GLOTTAL/ALVEOLAR stops in spelling databases.



(1) to (6) *are factors affecting the detectability and "goodness" of the BEFORE-LOW component.*

1 - onset abruptness **2 -** energy at onset

3 - existence of a low-frequency energy inflection

4 - non-existence of preceding higher-frequency energy

5 - non-existence of simultaneous higher-frequency energy (approx. 5ms integration)

6 - non-existence of simultaneous higher-frequency energy (approx. 8ms integration)

(7) to (11) *are factors affecting the detectability and "goodness" of the AFTER-HIGH component.*

7 - onset abruptness **8 -** energy at onset

9 - intensification in relation to preceding equifrequencial zone

10 - temporal separation from the onset of the BEFORE-LOW component

11 - ascending ΔF from the BEFORE-LOW component

(12) to (16) *are factors affecting the detectability and "badness" (capability to prevent firing of the "ascending sequence cells") of the eventual AFTER-LOW component.*

12 - onset abruptness **13 -** energy at onset

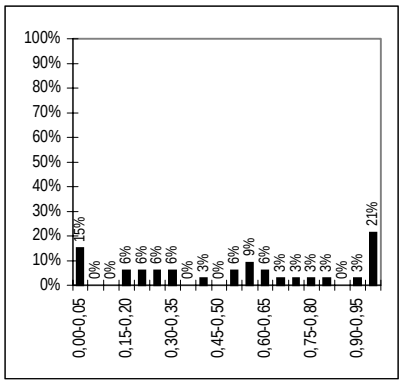
14 - intensification in relation to preceding equifrequencial zone

15 - temporal separation from the onset of the BEFORE-LOW component

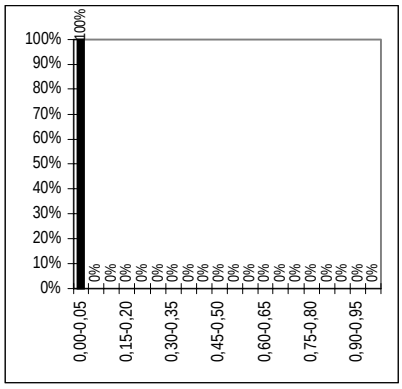
16 - descending ΔF from the BEFORE-LOW component

The goodness of the **Before Low** event is expressed by the intersection of several fuzzy intersective factors. Candidates for the **Before Low** event are searched all over the frequency * time matrix, except that no candidates are accepted after the vowel onset in the CVs.

The following histograms (the horizontal axis is the final fuzzy variable expressing the degree of existence of an ascending sequence) were obtained by applying the **same model** (no adaptation whatsoever; that is, we are trying to emulate the known human listener capability of perceiving this distinction even from non-native speakers) to LABIAL and ALVEOLAR/DENTAL stop sounds from spelling databases in various languages:

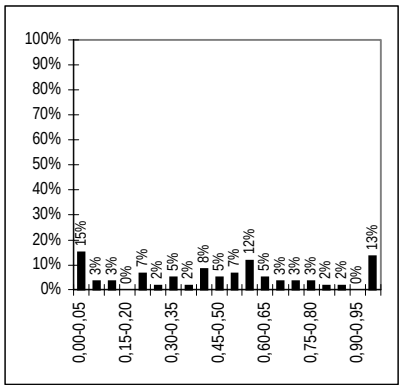


LABIAL stops (/pi/)

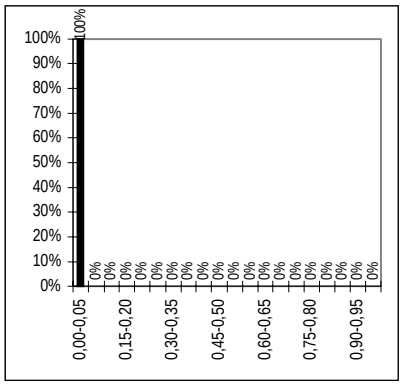


DENTAL stops (/ti/)

66 sounds from 33 speakers, **Portuguese** *in-house* database

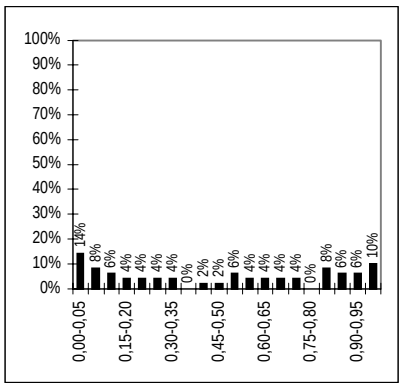


LABIAL stops (letter B)

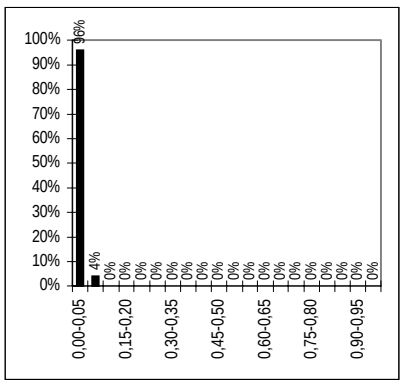


DENTAL stops (letter D)

120 sounds from 30 speakers, **U.S. English - Isolet** database**, CSLU/OGI**



LABIAL stops (letter P)

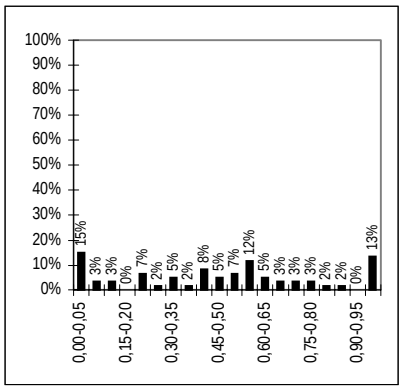


DENTAL stops (letter T)

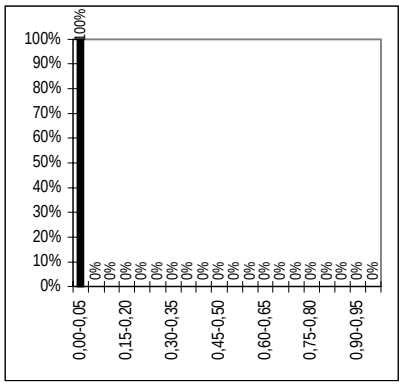
100 sounds from 50 speakers, **German – PhonData1** database**, BAS**

Usefulness for **discrimination across various languages** is clearly evident.

The above histograms are for full-band sounds.
Insensitivity to non-drastic filtering and graceful degradation for drastic filtering (low-pass) is demonstrated by the following histograms (U.S. English sounds)

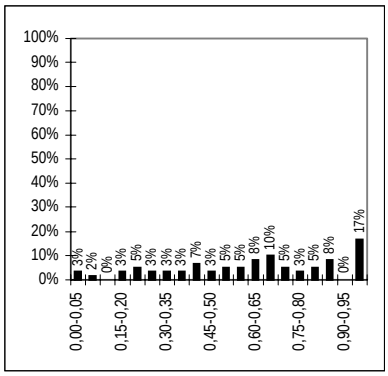


LABIAL stops (letter B)

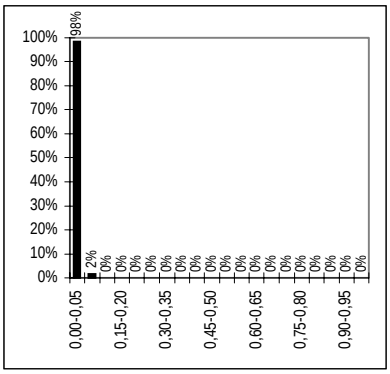


DENTAL stops (letter D)

120 sounds from 30 speakers, **U.S. English** - ***full-band***

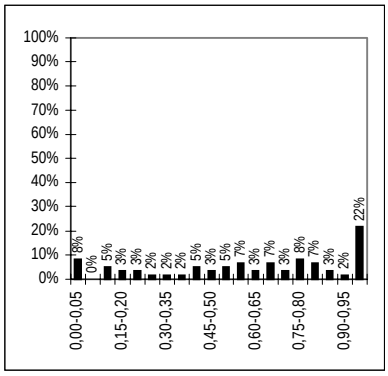


LABIAL stops (letter B)

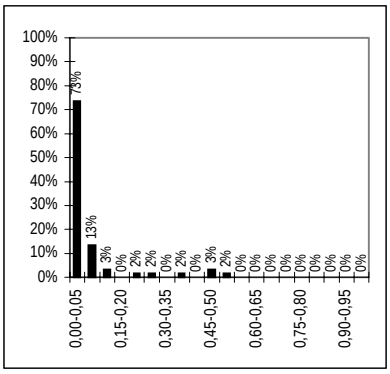


DENTAL stops (letter D)

U.S. English - ***sloping low-pass filter, -2dB/octave above 2KHz***



LABIAL stops (letter B)

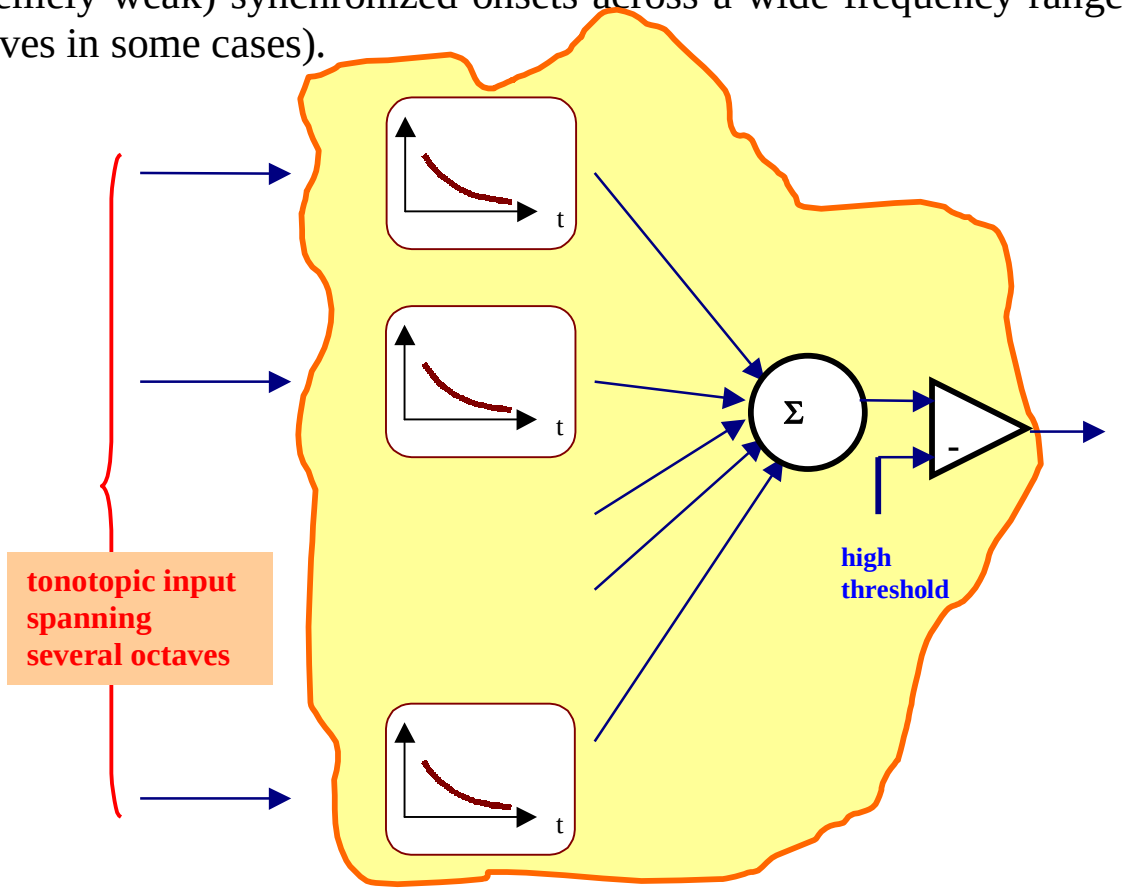


DENTAL stops (letter D)

U.S. English - ***abrupt cut-off (low-pass) above 3.5KHz***

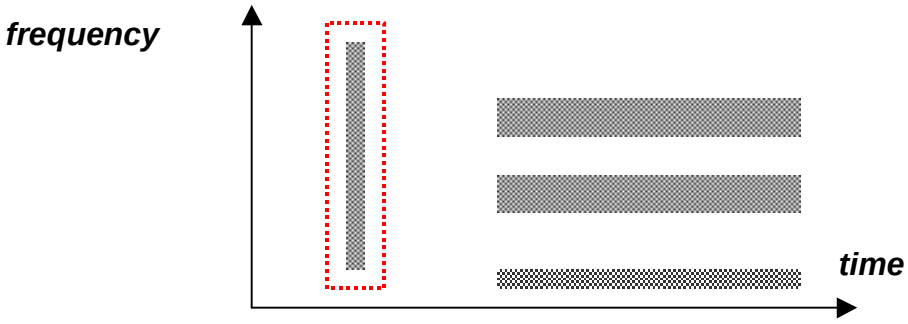
IC2 for the LABIAL category: duration-tuned cells responding to frequentially-wide onsets

ONSET cells in the *Cochlear Nucleus* respond to (possibly extremely weak) synchronized onsets across a wide frequency range (>2 octaves in some cases).



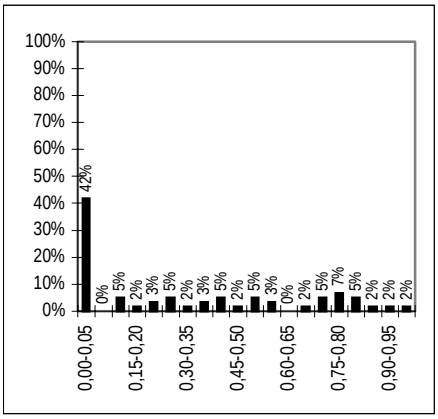
Input energy may be very weak (because many contributions are summed) but synchronization is crucial (because of the fast decay in gain).

Signals from *ONSET* cells, fed into **duration-tuned cells** (located more centrally) tuned to durations as short as 3-6 milliseconds, seem to provide another ***IC*** for the LABIAL category. In fact, patterns such as the following one are relatively common in LABIAL stops (a ***brief initial "vertical bar"*** – sometimes extremely weak - in the spectrogram)

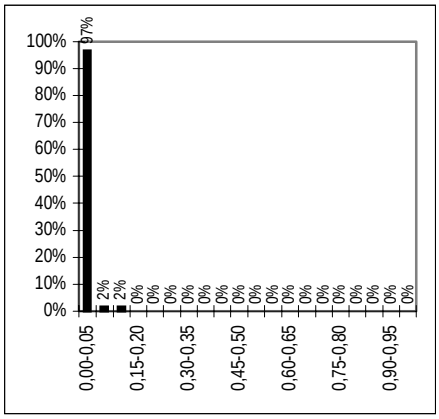


We have developed a *fuzzy logic* model for such an assembly.

Results in discriminating the LABIAL category against the DENTAL category, for U.S. English, are expressed by the following histograms:



LABIAL stops (letter B)



DENTAL stops (letter D)

120 sounds from 30 speakers, **U.S. English** - ***full-band***

It is seem that over half of the LABIAL stops attain scores higher that the highest attained by the DENTAL stops.

The temporal resolution to evaluate this *IC* must be on the order of 3ms or smaller.

...totally out of the range for conventional recognizers

Integration of several ICs working for the same category

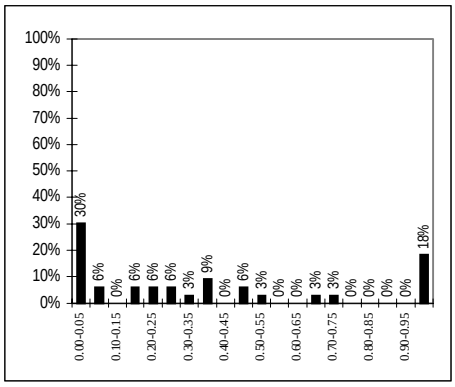
For any ***IC***,

- the histogram for the **implied category** is **spread out**
- the histogram for any **"contrary" category** is **squashed at zero**

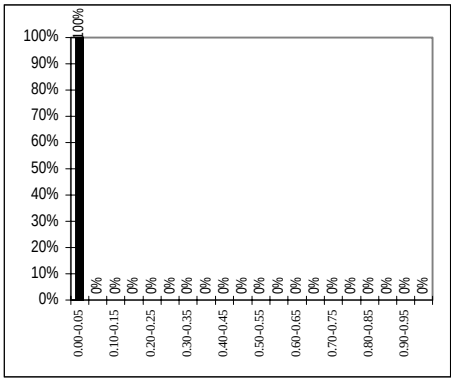
Performing integration through the most simplistic of *fuzzy logic* union operators (the maximum):

- the histogram for the **implied category** becomes **more right-heavy**
- the histogram for any **"contrary" category** stays **squashed at zero**

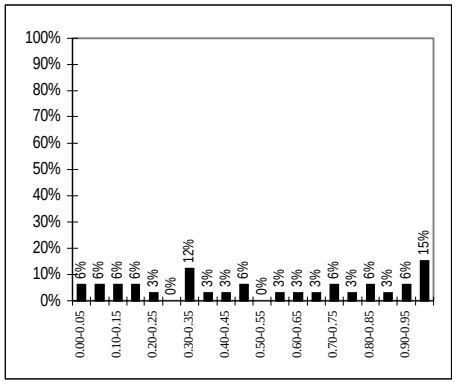
Example: integration of two ***ICs*** for the DENTAL category (all histograms: Portuguese full-band sounds; Dental ***IC1a*** and ***IC1b*** evaluate the high-frequency energy content of the sound segment prior to the vowel)



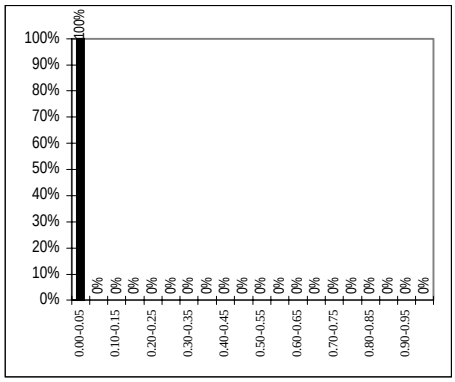
/ti/ - Dental *IC 1a*



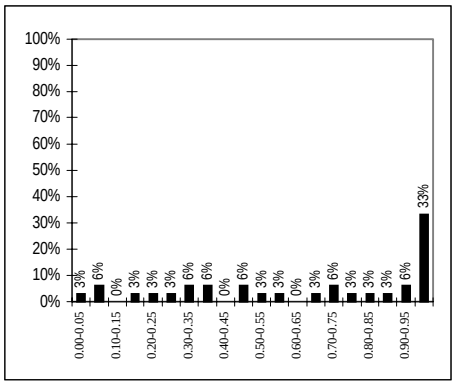
/pi/ - Dental *IC 1a*



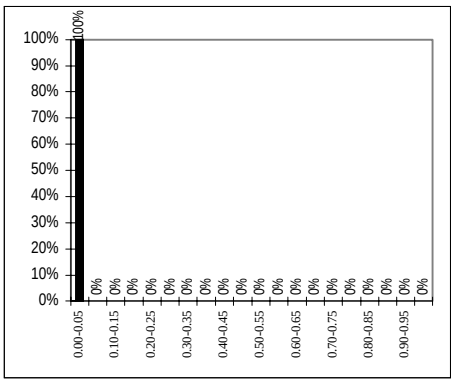
/ti/ - Dental *IC 1b*



/pi/ - Dental *IC 1b*



/ti/ - Max of Dental *IC 1a & 1b*



/pi/ Max of Dental *IC 1a & 1b*