AUDITORY FEATURES UNDERLYING CROSS-LANGUAGE HUMAN CAPABILITIES IN STOP CONSONANT DISCRIMINATION

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

ABSTRACT

For some phonemic distinctions human listeners exhibit a marked cross-language capability, in that they are capable of highly correct classification in relation to sounds (like CVs or VCVs) uttered by speakers of another language. This is particularly true regarding distinctions that are perceived in a more categorical fashion, like that of 3-way PLACE discrimination in stop consonants. It is plausible that the reason for this is a mostly common (across languages) auditory basis for human communication of this discrimination. Also, human communication of this discrimination is notably impervious to non-drastic variations in the frequency-transfer curve, which suggests that the relevant auditory features must have some inherent insensibility to these variations.

Models for two specialized auditory cells (onset cells with wide receptive fields, which can detect weak onsets synchronized across frequency, and sequence cells which detect frequency-ascending sequences composed of two onsets) were refined for the discrimination of DENTAL vs LABIAL stop consonants and applied to large spelling databases in Portuguese, German, and U.S. English. Similar discriminatory capability was observed both for German and U.S. English. Integration with a 3rd auditory feature resulted in error scores of approximately 2% when exactly the same model is applied to either German or U.S. English sounds.

1 - INTRODUCTION

1.1 – Human cross-language capabilities in stop consonant PLACE discrimination

It is well established that stop consonant PLACE discrimination is very well carried across languages (contrary to the voiced/unvoiced distinction, which for instance carries very poorly from U.S. English speakers to Portuguese listeners).

In a recent study [1], it is shown that native Korean listeners are capable of discriminating stop consonant PLACE, as uttered by U.S. English speakers, with less than 1% errors. This result, however, was obtained with utterances previously selected to be consistently classified, as well as to be the highest rated in goodness judgements, by native listeners of U.S. English. Thus, the results that might be obtained with an unselected mix of speakers, comprising speakers of good to below average intelligibility, could be somewhat poorer. That is, speakers of less good intelligibility strain the classification capabilities of native listeners, but these listeners are still able to maintain a very high score of correct recognition. Non-native listeners, on the other hand, may incur in significant error rates when faced with these poorer speakers.

The above discussion is useful in that it suggests expectations for a wholly correct model of human recognition (for listeners of a particular language, and for the above mentioned task): error scores as low as 1% may not be attainable, when the model is faced with databases of a different language which includes a significant <u>proportion of speakers of less good intelligibility</u>. On the other hand, if this proportion is not high (say, less than 5%), the error rate should not be much higher than 1% (say, on the order of 2%) and should not suffer appreciably from mild variations of the frequency-transfer curve (say, on the order of ±2dB/octave in the range above 1KHz).

1.2 – General assumptions about human phoneme communication

These assumptions have been given elsewhere [2] but are recast here along with some additional considerations. We are bearing in mind communication tasks – such as spelling, and communication of nonsense words – in which humans exhibit a clear capability of speaker-independent phoneme communication. Nonetheless, the mechanisms crucial to this capability will obviously also be operative in word or sentence communication - though they may then be used for the communication of other speech units.

In the former tasks, there emerges – with very clear contours - the paradox of constancy of perception, in spite of variation of form. That is, the same CV, uttered by different speakers, presents very diverse acoustic forms (so diverse that extensively trained automatic recognizers incur – persistently - significant error rates) whereas human listeners correctly recognize the consonant, with apparent ease. The paradox is heightened when we consider that recording sounds through different microphones, or including speakers native of a different language (provided these are of "good quality") does not diminish appreciably the performance of human listeners, while wreaking havoc with automatic recognizers' performance.

To solve the paradox, we propose to consider the following visual communication analogy:

Suppose that a person is asked to draw pictures of a small set of fruits (pineapple, banana, orange, …) just good enough to be correctly recognized when briefly flashed on a screen.

One particular drawer might present the PINEAPPLE texture very markedly; this will allow him to relax, for instance, the contour of the pineapple… …which may even be rendered in a form ambiguous between PINEAPPLE and ORANGE Another drawer might "synthesize" a weakly marked texture, but then trace the contour in a very marked way.

In this "thought experiment", it may also be expected that *to draw a well perceived PINEAPPLE, a drawer may produce a texture that is much more marked than in any real pineapple (thus getting away from any conceivable category centroid), and by that he will still be aiding correct recognition*

This analogy suggests that for each phonemic distinction there are multiple *information carriers (IC*s) - or *cues*, or *features* - evaluated independently of each other, all being orthogonal to between-categories boundaries and that there exist, among the cues, trade-off relations that may extend to the point of alternativity. It is even conceivable that two different speakers may successfully communicate the same CV using entirely disjoint *ICs* ; this might be the case if a new speaker undergoing speech acquisition finds especially easy to emit a particular *IC* with high "intensity": this speaker may then "rest satisfied" and relax the emission of other *ICs* to the intended category. This plausible process is reminiscent of natural selection [3]: the well-known case of the panda's thumb, which achieves functional success (grasping action) with no morphological conformity (no real thumb) is particularly enlightening with respect to the paradox.

Another concept from natural selection which may be relevant to the phoneme communication problem is that of *exaption*, that is, the "seizing" by a new function (phoneme communication) of biological mechanisms that evolved previously as adaptations to other tasks (such as recognizing species calls in some distant animal ancestor, or as an even more basic survival-enhancing acoustic detection ability). This makes it likely that some of the *ICs* are mostly direct expressions of the acoustical metrics computed by some "hard-wired" (that is, not substantially modified in response to speech use) neural assemblies.

Use of several *ICs* pointing to the same category achieves redundancy and robustness to signal degradation: when degradation is not drastic, some *ICs* may be obliterated, but if some others survive, correct recognition by listeners will still be obtained.

The speakers also want to accommodate articulatory ease, indulge articulatory variability induced by various motivations (the conveyance of a personal speaking style, emotional status, etc.) – all of this is made possible by the extensive trade-offs between the several *ICs* for the same category.

1.3 – The set of *Information Carriers* (*ICs*) for human communication of the PLACE distinction in stop consonants

Characterization of this set is the subject of our on-going research, but the following *ICs* are thought to be well stabilized; further additions may have to be made, but their importance will be of a secondary degree.

Introduction of the *ICs* was driven by the need to explain the perception of PLACE in natural or edited sounds when no explanation could be found in terms of the *IC*s known at a particular time in the research undertaking. It order to provide a substantial number of such "driving sounds", the need for considering several languages was recognized early on; for a single language, most speakers conform to acoustic regularities particular to that language and the number of sounds that provide a clear-cut challenge for explanation of their perception is very limited.

The current characterization of each *IC* is the result of an hypothization endeavor, followed by satisfactory results in the application of a model of the *IC* to large number of sounds.

Acceptable *ICs* must be biologically plausible and must exhibit some degree of independence to non-drastic variations in the frequency-transfer curve. Some *ICs* may correspond closely to metrics computed by some specialized auditory cells – in this case, the neural algorithms computed by these cells may yield a high selectivity in frequency and/or in time. Other *ICs* may correspond to *speech schemas* [4] and thus must be expressible in terms of plausibly auditory-salient representations such as gross integration of energy.

The (current) set of *ICs* for PLACE discrimination into the three categories LABIAL, DENTAL and GLOTAL/VELAR is then:

LABIAL-IC1: <u>ascending sequence in the F2/F3 zone</u>. This is assumed to be evaluated by ascending sequence cells such as those that have been found in the primary auditory cortex of primates [5]. Since the abruptness of onset is the most important characteristic of each of the two components of the sequence, insensitivity to non-drastic variations of the frequency-transfer curve is assured. There are many references in the perception literature to an "ascending" quality being a cue for LABIAL (see for instance [6]).

LABIAL-IC2: <u>ascending trajectory of the dominant low-frequency skirt in the F2-F3 zone</u>. For this *IC* there is also a two-times comparison but the "after" term is to be evaluated through temporally-gross integration, and the onset of the vowel functions as a temporal marker signaling this "after" term.

LABIAL-IC3: <u>complete or near-complete absence of unvoiced energy prior to vowel onset</u>. This is similar to the "burstless" quality referred in [7] as a cue for LABIAL. There are a number of studies in the literature that also concur in this finding, many of which are also cited in [7] . However, we added the detail that high-frequency energy occurring just at the vowel onset may proved a non-burstless quality. This

IC is thought to be evaluated through temporally-gross integration.

LABIAL-IC4: initial brief (<3-5ms) "vertical bar" in the spectrogram, followed by no significant high-frequency energy. This is thought to be evaluated with the help of wide-receptive field onset cells, in conjunction with temporally-gross integration of energy.

DENTAL-IC1: initial tone-burst-like segments (>6-8ms) (usually corresponding to the initial aspirated or voiced segments of F2 or F3) of very thin bandwidth. This *IC* is primarily based on the output of hypothetical cells similar to *level-tolerant* neurons [8], although there seems to be also a temporal windowing mechanism involving onset cells.

DENTAL-IC2: existence of upward inflections in the spectrum, occurring at "high" (>3.5KHz) frequencies during the *burst+aspiration* segment. This is – albeit distantly - related with [9]. The metric for this *IC* is assumed to be dependent on auditory cells exhibiting marked lateral inhibition from the lower sideband.

DENTAL-IC3: segment prior to vowel release (that is, the *burst+aspiration* segment) having a considerably stronger high-frequency content then the ensuing vowel. This is likely to depend on temporally-gross energy integration, and on the use of the vowel onset as a temporal marker to distinguish the 2 terms of the comparison.

DENTAL-IC4: grossly equifrequencial sequence in the F2/F3 zone. Assumed to be based also on sequence cells of the primary auditory cortex.

GLOTTAL-IC1: descending sequence in the F2/F3 zone. Also based also on sequence cells of the primary auditory cortex

GLOTTAL-IC2: Strong onset of "compact energy" in the F2-F4 zone, followed briefly (<10-20ms) by abrupt offset. No specific auditory cells have been found in the literature to account for the evaluation of such a metric, but their existence has some biological plausibility.

GLOTTAL-IC3: descending trajectory of the dominant low-frequency skirt in the F2-F3 zone. The fact that such a trajectory will continually meet unadapted cells in the auditory nerve provides a basis for its auditory evaluation.

The above characterizations, and the present state of development of models for some of the *ICs* has been the result of extensive studies with natural and edited sounds. As a first step, we tried to predict the perception of sounds (of unknown PLACE) based on inspection of several spectral displays, and the estimation of how the relevant auditory structures would react to the sound. This in turn led to the development of fuzzy-logical, auditorily-plausible, models for some of the *ICs*.

We develop/refine the models through inspection of their results in 5 sets of sounds: an in-house research database of /ti/ and /pi/ sounds from 33 Portuguese speakers (representative of Portuguese unvoiced stops), the letters "T" and "P" from the first set (30 speakers, 120 sounds) of the Oregon Graduate Institute ISOLET Database (representative of U.S. English unvoiced stops), the letters "D" and "B" from the first set of ISOLET (representative of U.S. English voiced stops), the letters "T" and "P" from the first 50 speakers of the Bavarian Archive for Speech Signals PHONDATA1 Database (representative of German unvoiced stops), and the letters "D" and "B" from the same 50 speakers (representative of German voiced stops). It is to be emphasized that for each *IC* the same model is used throughout, with no adaptation whatsoever, and that the different sets have obviously used different microphones, as well as recording conditions.

2 – AN *INFORMATION CARRIER* FOR THE LABIAL CATEGORY, BASED ON ONSET CELLS

In this section, a model for **LABIAL-IC4** is discussed, along with its motivation.

In our research towards being able to predict the perception of sounds of unknown PLACE, we came across some sounds ("P" and "B" in spelling databases in German and U.S. English) whose LABIAL perception seemed more robust than could be explained in terms of the other three *ICs* for LABIAL (which were uncovered, and characterized, first). More definite conclusions could be extracted from some particular sounds which lent themselves to filtering or editing operations that clearly removed (or greatly diminished) the other *ICs* for LABIAL; many of these sounds maintained a clear LABIAL perception, raising the need for another LABIAL *IC*.

The common acoustical trait among these sounds was the presence of an initial "vertical bar" in the spectrogram followed by (at least) a few milliseconds with little energy across higher frequencies. One difficulty in the way of making this observation was that most often the "vertical bar" seemed to be of such low energy (relative to the rest of the speech signal) that at first it seemed improbable that it would play a significant part in perception.

But it was realized that some onset cells in the cochlear nucleus could exhibit an extremely wide receptive field, measured using the concept of *two-tone facilitation* [10] and that this could result in measurable responses even with the "best-frequency" tone being as low as 30dB below threshold. So, if it turns out that a fair proportion of LABIAL stops are capable of exciting these cells, while non-LABIAL stops are not, it is clearly conceivable that this came (during the evolution of languages) to constitute a valuable *IC* for LABIAL.

Since there are varied types of onset cells (with differently wide receptive fields), and members of each type may be found with central frequencies all along the audible range, there remains the question of establishing the characteristics of those onset cells that are mobilized for LABIAL-IC4. Cells sensitive to very low frequencies (say, below 2KHz) would tend to give unreliable information, since acoustical accidents due to non-speech noise are apt to cause excitation of such cells; we considered only cells with receptive fields extending upwards from 3.5KHz, up to 7.0KHz. Another issue is the frequencial width of the receptive field; we considered a fixed width of 1400Hz (a point which is to be refined in the future).

The essence of the neural algorithm for onset cells is the summation of the outputs from a large number of auditory nerve cells (spanning a wide frequency range), occurring simultaneously. It is possible that for some cells contributions emanating from a restricted frequency range will not suffice to excite the cell, however strong these contributions (it is even possible that a very strong frequency-local contribution will turn off the cell, through the hypothetical mechanism of *shunting inhibition*).

We implemented a fuzzy-logical model to account for these dependencies. For direct comparison with commercial spectrographic displays and sound editing software, we use simple FT spectra as the input representation. The speech signal is represented by FFT spectra calculated, with a Hamming window, over frames of 11.6ms, with a 3-ms frame advance. Thus the input matrix is composed of points E(F,T) where F=fx86Hz and T=tx3ms. At each such point, we computed the ***Unadapted-Increment***(F,T) considering the energy at point (F,T) and energy previously occurring at frequencies proximal to F. ***Synch-Increment***(F,T) was computed with a metric similar to summation applied to ***Unadapted-Increment***(F',T) with F' spanning from F to F+1400Hz. In this summation-like metric, the contribution of outstanding peaks is subject to limitations. The most adequate form of these limitations is still being studied; for instance, the intriguing possibility that an extremely outstanding peak might actually decrease the response of the cell, through *shunting inhibition*, is for the time kept open.

It is interesting to note that this metric is unavailable to conventional automatic recognizers, since their input representation has, as a rule, much less time resolution than used here.

The model was refined (in the version reported here, only about 10 parameters were explored) primarily using U.S. English "P" and "T" sounds and was applied unaltered to German and Portuguese "P" and "T" sounds. The histograms for U.S. English are presented below:

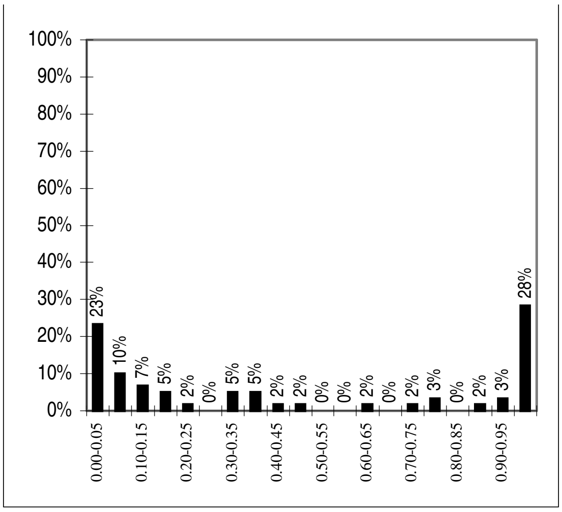

Figure 1 – Histogram for the fuzzy variable expressing LABIAL IC 4 for 60 U.S. English "P" sounds (Isolet, 1st set)

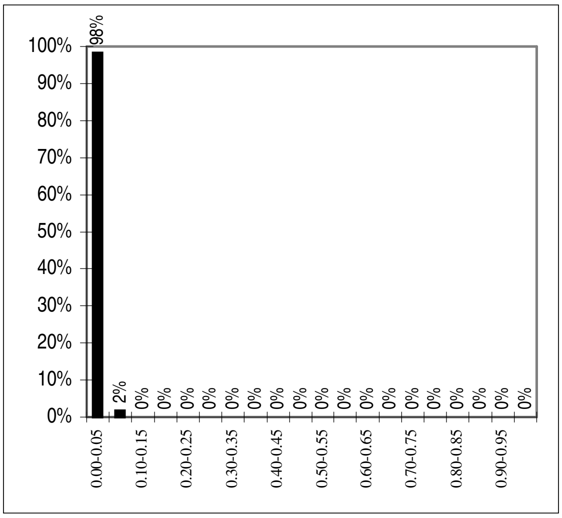

Figure 2 – Histogram for the fuzzy variable expressing LABIAL IC 4 for 60 U.S. English "T" sounds (Isolet, 1st set)

From these histograms, it is apparent that significant to high values in this metric only occur for LABIAL sounds, and not at all for DENTAL sounds, making it an obviously useful *information carrier* for the discrimination between these two categories. It is evident that the metric exhibits a generous "exclusively LABIAL" range of medium to high values, which range is only attained by LABIAL sounds.

The results obtained applying the same model to German sounds are given below:

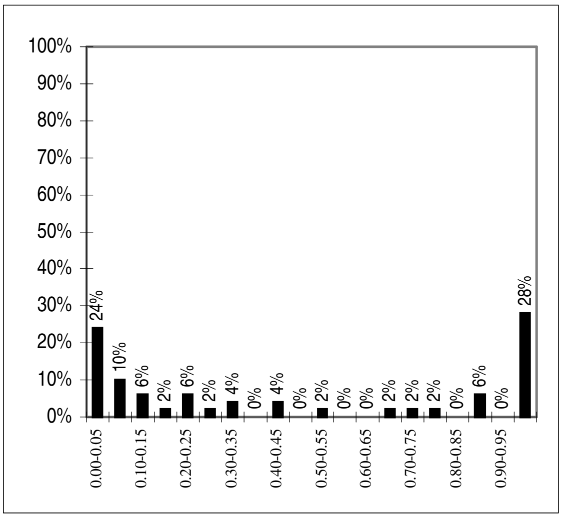

Figure 3 – Histogram for the fuzzy variable expressing LABIAL IC 4 for 50 German "P" sounds (PhonData1, first 50 speakers))

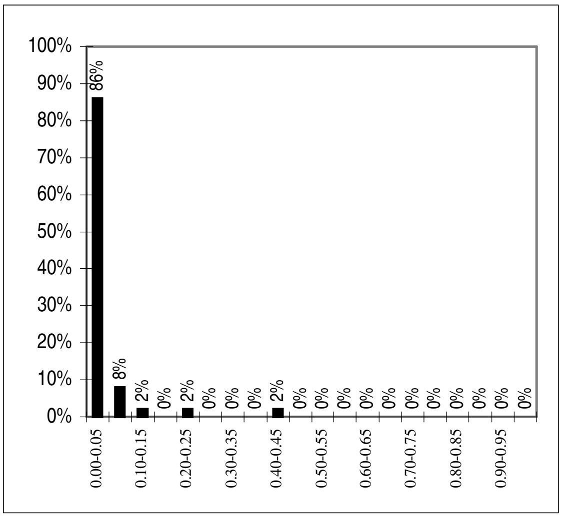

Figure 4 – Histogram for the fuzzy variable expressing LABIAL IC 4 for 50 German "T" sounds (PhonData1, first 50 speakers))

The results for German are similar to those for U.S.English. The "exclusively LABIAL" range is here somewhat spoiled by a single sound with a mark at 0.42 (sound "hdbdT" – PhonData1 labels) but that likely is the result of imperfect refinement of the model. The results for Portuguese, however, are very poor: "P" sounds almost never elicit significant values in the metric. But this is not surprising, since "P" sounds in Portuguese have very weak and short *burst+aspiration* segments; in fact, it is uncommon in Portuguese for these segments exceeding 15ms in duration, whereas for U.S. English durations in excess of 100ms are frequent.

3 – INTEGRATION OF DIFFERENT *INFORMATION CARRIERS*

Even granting success in modeling the different *ICs* , the problem of their integration adds another layer of complexity. We will simply show – using the simplest possible fuzzy-union operator (the maximum) - how two different LABIAL *ICs* yield discrimination superior to that of the better of those *ICs* .

The LABIAL *IC* which has the better discriminatory power is LABIAL-IC1: ascending sequence. Histograms for Isolet 1 "P" and "T" are shown below:

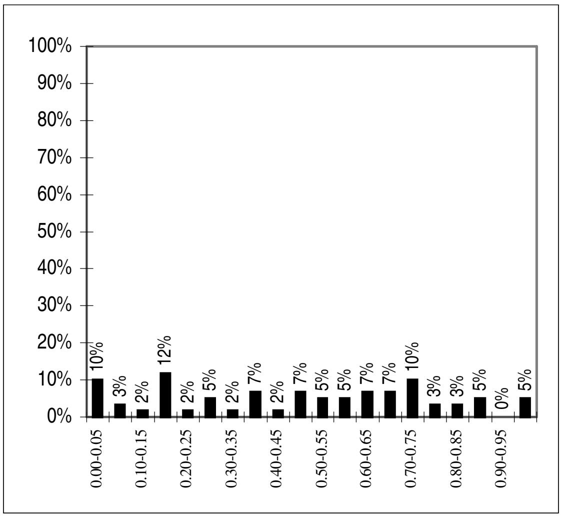

Figure 5 – Histogram for the fuzzy variable expressing LABIAL IC 1 for 60 U.S. English "P" sounds (Isolet, 1st set)

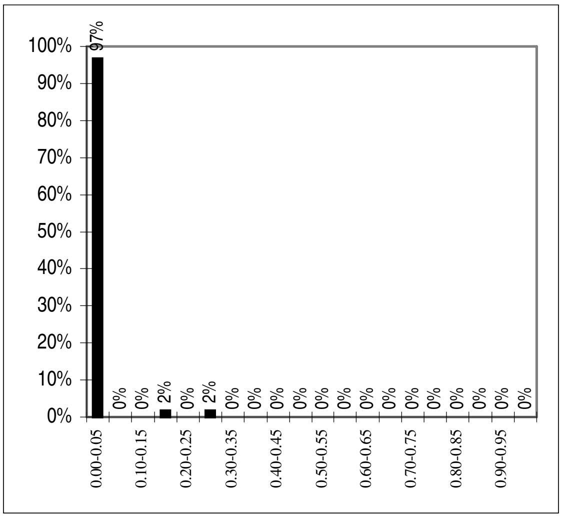

Figure 6 – Histogram for the fuzzy variable expressing LABIAL IC 1 for 60 U.S. English "T" sounds (Isolet, 1st set)

Simply taking the maximum of the two fuzzy variables expressing LABIAL IC 1 and LABIAL IC 4 yields the following histograms:

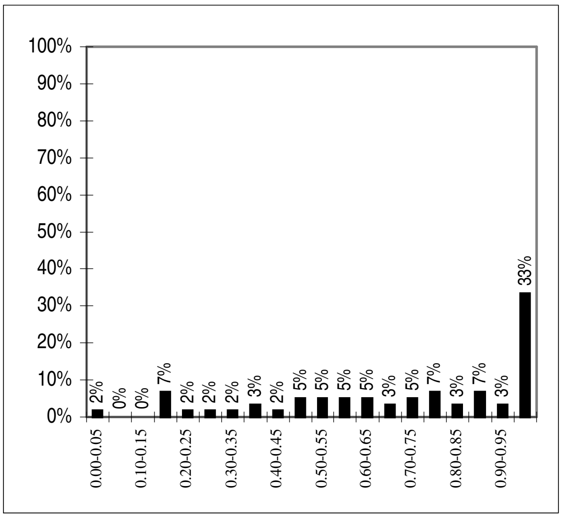

Figure 7 – Histogram for the maximum of LABIAL IC's 1 and 4 for 60 U.S. English "P" sounds (Isolet, 1st set)

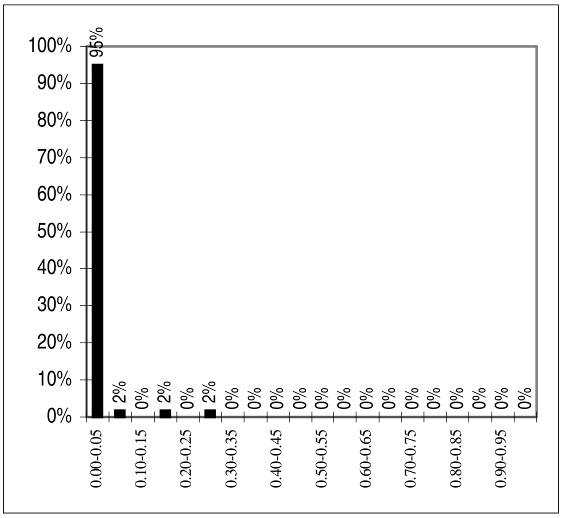

Figure 8 – Histogram for the maximum of LABIAL IC's 1 and 4 for 60 U.S. English "T" sounds (Isolet, 1st set)

Bringing in more *ICs* further improves discrimination. Performing fuzzy intersection of the above maximum (of LABIAL IC1 and LABIAL IC4) with the fuzzy variable expressing DENTAL IC-3 results in a fuzzy variable which, thresholded at 0.15 yields 1.7% "P" vs. "T" discrimination errors for U.S. English and 2% for German.

4 – CONCLUSIONS

A small number of *Information Carriers*, each with a reasonably simple characterization in terms of known auditory processes, is shown to be able to approach human capabilities in the cross-language communication of stop consonant PLACE. This was shown through modeling of some of the *Information Carriers* relevant to the LABIAL vs. DENTAL distinction.

Low error scores were maintained not only across languages, but also in spite of differences in recording settings that exist between databases. This suggests that the proposed metrics are substantially insensitive to non-drastic variations in the frequency-transfer curve.

Further work is going on in connection to *Information Carriers* relevant to the discrimination of GLOTAL/VELAR PLACE.

ACKNOWLEDGEMENTS

The research reported in this paper has been conducted under the Research and Development Contract Praxis 2/2.1/TIT/1558/95 of the PRAXIS XXI Program of the Junta Nacional de Investigação Científica e Tecnológica.

REFERENCES

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

[2] - Eduardo Sá Marta, Luis Vieira de Sá - "Auditory cells with frequency resolution sharper than critical bands play a role in stop consonant perception: evidence from cross-language recognition experiments" - Proceedings of the *NATO Advanced Study Institute on Computational Hearing,* Il Ciocco, Italy, 1998

[3] – Gary Cziko, "Chapter 11 – The Evolution, Acquisition, and Use of Language" in "Universal Selection Theory and the Second Darwinian Revolution", MIT Press, 1995

[4] – Bregman, A.S., "Auditory scene analysis: the perceptual organization of sound", MIT Press, 1990

[5] - 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

[6] A. Lahiri et al - A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: evidence from a cross-language study - J. Acoust. Soc.Am. 76, pp.391-404, 1984

[7] Smits,R., Bosch,L., Collier,R., "Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. I. Perception experiment" J.Acoust. Soc. Am. 100 (6), pp.3852-3864, 1996

[8] Suga,N., Zhang,Y. and Yan,J., "Sharpening of Frequency Tuning by Inhibition in the Thalamic Auditory Nucleus of the Mustached Bat", Journal of Neurophysiology,Vol.77, pp.2098-2114, 1997

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

[10] - Dan Jiang, Alan R. Palmer, Ian Winter – "Frequency extent of two-tone facilitation in onset units in the ventral cochlear nucleus" – Journal of Neurophysiology, vol.75, pp.380-395, 1996