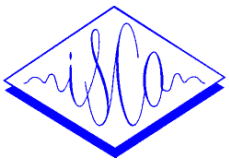


4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

RESEARCHING THE PROCESSING STRUCTURES OF HUMAN PHONEME RECOGNITION BY ANALYSIS OF NATURAL STOP-CONSONANT-VOWEL UTTERANCES THAT ELICIT CORRECT RECOGNITION THROUGH UNUSUAL ACOUSTIC PATTERNS.

Eduardo Sá Marta, Fernando Perdigão, Luis Vieira de Sá
Dep. Engenharia Electrotécnica, FCTUC (Universidade de Coimbra)
Instituto de Telecomunicações - Pólo de Coimbra
Largo Marquês de Pombal, 3000 COIMBRA - Portugal

ABSTRACT

One of the main ultimate objectives of speech perception research is the description, on a neurophysiological basis, of the mechanisms involved in human phoneme recognition. Direct probing is obviously out of the question, but some novel inspiration for hypotheses may be gleaned from human phonemic productions that elicit correct and robust identification in spite of acoustical patterns that blatantly infringe known (production and/or perception) "rules".

To find a number of productions as such, the detailed analysis of a multinational database of 64 non-professional speakers was undertaken. Some productions showed to grossly violate the previously known "approximate rule" that states that in a stop-high vowel CV, a "slightly ascending F2 transition cues dental place, whereas a markedly ascending transition cues labial place". A biologically-plausible model of a processing structure ancillary to formant-transition perception is proposed that is compatible not only with these productions, but also with the more common productions that conform to the "approximate rule" mentioned. Furthermore, the model appears to explain why a slightly ascending transition (prevalent in dental-stop, high-F2 vowel CV's) may be perceptually similar to a markedly descending transition (prevalent in dental-stop, low-F2 vowel CV's).

10.21437/Eurospeech.1995-520

1. INTRODUCTION

It is highly plausible that neuronal processing structures, located in auditory nuclei, play a crucial part in phoneme recognition by humans. Obviously, it cannot be considered to physically probe the actual neuronal signals in human auditory nuclei, which has resulted in an almost complete lack of knowledge in this area. But knowledge about any natural world domain can in fact be acquired without the direct probing of signals. Biologically-plausible hypotheses can be launched and perceptual experiments be designed to validate, refine or discard any hypothesis. Experimental facts that appear "surprising" (not easily explainable), but are repeatable and well isolable, constitute a welcome source of inspiration for hypotheses.

In human phoneme recognition there is available a profusion of such facts: utterances of CV syllables by some (unusual but not rare) human speakers, that do not appear to conform to any known acoustically-defined "norm" or "approximate rule" for the consonant being spoken and, in spite of this, elicit correct and robust identification by all listeners. It has been the practice of speech production - as well as automatic recognition - researchers, no doubt struggling to find some order in the overwhelming variability of speech productions, to summarily dismiss these more "surprising" utterances as "outliers". But it should be recognized that these utterances, being the result of a long period of auditory-based refinement during speech acquisition by the speaker, are in fact repositories of (tacit) knowledge about human phonemic perception. Among all the natural non-lexical utterances robustly recognized by humans, the most unusual forms are precisely those that can supply the most novel information about human phonemic perception.

One approximate (that is, allowing unexplained exceptions) rule about place of articulation perception in CV's with front vowels is that markedly ascending F2 transitions cue labial, while slightly ascending cue dental [2, 3]. The consensus on this approximate rule, interpreted as a production constraint, can be appreciated from the fact that automatic recognizers designed for stop consonant discrimination have often incorporated explicit rules about F2 gradients [6] or used frequency gradient operators [4] or included time-delay structures that propitiated the learning of formant gradient information [9]; in all of these recognizers, this yielded higher correct recognition scores, which shows that this rule has a strong statistical importance. The explicit rules used by DeMichelis et al [6] state that, for front vowels, the maximum ascending $\Delta F2$ that can be spanned by F2(vowel nucleus)-F2(locus) is 200Hz.

But (as will be shown in this paper) this rule, statistically significant as it may be, is grossly violated by some human CV productions that nevertheless achieve robustly correct recognition by human listeners. It may be noted in passing that these gross violations are not at all evident from either the automatic speech recognition (ASR) or the speech perception literatures, but this is probably due to the fact that the ASR researchers are apt to treat these gross violations as "outliers" (as pointed above) while the generality of speech perception studies have mobilized a very small number of speakers (with many studies being single-speaker-based), presumably of a professional or near-professional speaker quality.

Also, a long-standing question has been that while for CV's with high-F2 vowels dental place information seems to be expressed by slightly ascending F2 transitions, for CV's with low-F2 vowels this information seems to translate into markedly descending F2 transitions. Accounting for the commonality of category identification, in spite of acoustic diversity (even opposition, one might say) has been difficult. One theory [5] claims that human listeners are able to track intended gestures of speakers, and thus interpret a slightly ascending F2 transition into a high-F2 vowel as revealing a dental consonant, while a markedly descending F2 transition into a low-F2 vowel reveals also a dental consonant; but the proponents of this theory offer no backtracking mechanisms (biologically plausible or otherwise).

In the cadre of all these difficulties, it seems that at the moment there is no neurophysiologically expressed theory of formant transition perception in stop-vowel CV's.

In *Sec.2* of this paper, a search is made of human speaker productions that most grossly violate the known approximate rules for F2 behavior according to place.

In *Sec.3*, the above productions are taken as inspiration for sketching a model of a neuronal structure plausibly involved in the detection of dental-related information in F2 transitions.

2. SEARCHING FOR PREVIOUSLY UNREPORTED WAYS OF ELICITING DENTAL IDENTIFICATION THROUGH FORMANT TRANSITIONS

The approach taken here is that any valid theory of phonemic perception must account for all robustly perceived (natural) utterances. An approximate rule that allows 10 to 30% of infractions (some of those being gross infractions) may turn out to simply express a high degree of statistical correlation between the rule criteria and what actually dictates perception or even what constitutes the most common way of articulating the phoneme in question. On the other hand, no strong claim is made of contextual invariance. That is, it is conceivable that the mechanisms involved in the perception of place of stop consonants in CV's may be somewhat different in the case of high-F2 vowels and in the case of low-F2 vowels, the commonality existing only at the cognitive level. We thus focused on stop-vowels in which the vowel is either /i/ or /ɛ/, since from attempts at automatic recognition of alphabet letters in several languages it is well known that the /i/-sets and the /ɛ/-sets are the most difficult (for example, Cosi *et al* [1] report a 35% error rate for the /i/-set and a 12% error rate for the /ɛ/-set in Italian alphabet speaker-independent recognition), while (for comparison purposes) at the same time also acquiring exemplars of the more easily classified stops before low-F2 vowels.

The search for previously unreported acoustic forms for natural (and robustly perceived) stop-consonant productions was developed through a large-scale undertaking: the analysis of over 30 utterances of each of 64 (non-professional) speakers, 54 being Portuguese, 4 German, 3 Spanish , 2 French and 1 Belgian (the rationale for including speakers of other mother tongues was that these would provide even more diverse acoustic forms that are nevertheless well identified by Portuguese listeners; also, it should be expected that the relevant processing structures are largely language-independent, at least among European languages with 3-valued place of articulation in stops).

In our research corpus some utterances - albeit robustly perceived by listeners - very markedly infringed the above mentioned approximate rule for dental vs labial place discrimination, providing opportunities for refining the rule (with the ultimate aim of building a model possessing a plausible neurophysiological formulation). One such example (/tɛ/ from a female speaker) displaying a ΔF2 of ≈600Hz (which trebles the De Michelis limit) - is shown below in ***Fig1.a*** and ***Fig.1.b***.

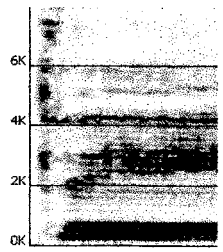


Fig.1.a

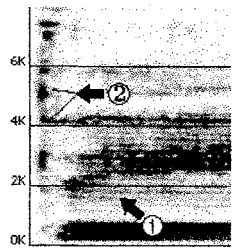


Fig.1.b

The (tentative) perceptual analysis of this utterance provides some knowledge about the perception of ascending F2 transitions and its part in stop place recognition. In ①, it can be observed that some harmonics well below F2 draw a non-ascending trajectory starting from the onset of voiced F2. These harmonics provide the lowest energy inflection in the F2 zone. Since it is plausible that lateral inhibition plays a part in the perceptual appreciation of F2 and its trajectory, it may be suspected that the auditory pathways representation of this F2 transition is much less ascending than that of the F2 top (as seen in the spectrogram). In ②, it can be observed that there is an energy inflection slightly above 4KHz. It is known from the early studies of Delattre *et al* [2] that high frequency (>3.5KHz) energy in the burst cues dental consonants. Again invoking lateral inhibition (as well as 2-tone suppression), it is to be expected that the most relevant measure should not be "absolute energy at high-frequency", but that energy with its perceptual impact modulated by inflection in the spectral energy density function.

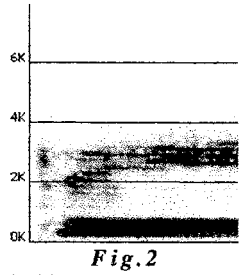


Fig.2

Removing, through filtering, both the low harmonics and high-frequency energy resulted in the sound shown in ***Fig.2*** which is perceived as labial (obtaining a clear identification of /p/ required both modifications). In summary, the additional knowledge provided by this analysis and associated experiments consisted in: (a) the lower-frequency skirt of F2, and/or any significant harmonics that may exist well below F2, seem to condition the perception of the F2 transition *(this has been consistently verified in other utterances that have significant harmonics well below F2, or a non-steep F2 low-frequency skirt)* ; and (b) neural information expressing a rising F2 transition *(which "points" to labial place)*, will be weighted against neural information expressing high-frequency energy inflections in the sound onset *(which points to dental place)*. This last point (b) is in fact well known [8, 4] and even consensual. The first point (a) has not been given attention in the literature, perhaps because the consideration of the F2 skirts' characteristics would make it difficult to enunciate a rule-formulated description.

A second utterance, which commands the introduction of a radically novel interpretation, is shown in ***Fig.3*** (another /tɛ/ from a female speaker, showing a ΔF2 of ≈600Hz). This time, there are no significant harmonics below F2, and removal of high-frequency energy does not budge the listeners' identifications away from T.

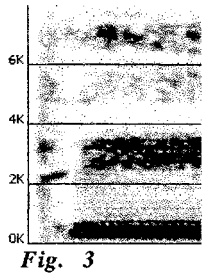


Fig. 3

Insight on what caused dental perception was gained when the first 6ms of the burst, low-pass filtered and amplified, were placed *in lieu* of the 6-ms burst of a natural /pɛ/ (***Fig.4.a***) from the same speaker, which operation did not change the ΔF2 (from F2 onset to the F2 at the vowel nucleus) of the /pɛ/ sound. Listeners' identification of the consonant in this edited sound (***Fig.4.b***) clearly migrated

towards /t/ *(it may be noted that the perception of this edited sound is also in violation of other approximate rules for perception that do not mention F2 transitions, such as the one in [4])*. It may observed that the /tɛ/ burst has a segment of F2 aspiration with tone-burst-like characteristics (reasonably stationary frequency, sharp spectral definition). Actually, the same migration was obtained with a 6-ms tone burst at the same frequency (this tone burst having been somewhat jittered, and prefixed by a short, 1-ms impulsive sound, in order to increase its naturality, and ramped-up to avoid spectral dispersion at the onset). Identification of any of the hybrid sounds reverted to /p/ when the prefixed burst was subjected to amplitude envelope modification of the form shown on ***Fig. 4.c***, or when its frequency was jittered +/-1/4 octave in rapid alternation (with a 1.5-ms period, 0.75ms + 0.25 octave, 0.75ms -0.25 octave)

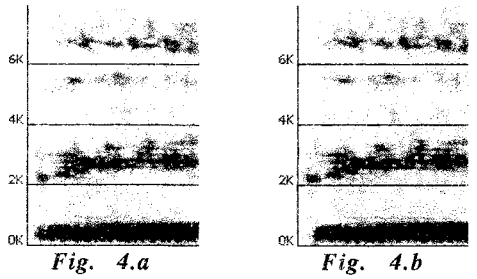


Fig. 4.a *Fig. 4.b*

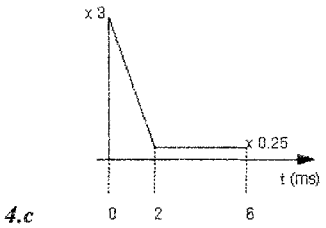


Fig. 4.c

This seemingly disproportionate perceptual impact of such short tone bursts, which do not even change the ΔF2 of the sound, is in fact rather unsurprising if we consider known properties of some "onset responders" and "primary-like" cells in the cochlear nucleus. These cells respond to sudden-onset, constant amplitude tone bursts with very high firing rates during a few milliseconds, and then fall down to a much lower rate. Tone bursts with a rapidly-decaying envelope (these are usual for the F2 aspiration in labials as well as velars) have a more diffuse spectrum and it would seem that this would result in excitation of inhibition bands of cochlear nucleus cells. Since stop consonants are the only speech sounds in which it can be assumed that *fast adaptation* is completely reset at the onset of the sound, it does not seem strange that this property would have been exploited for perceptual place discrimination.

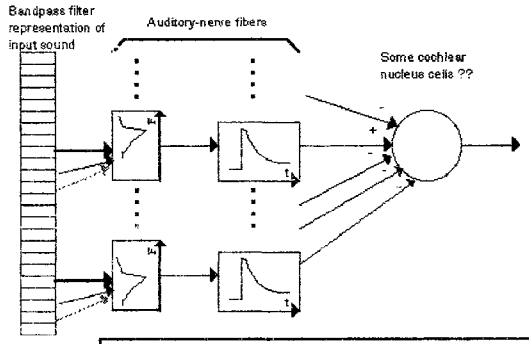
Further investigation (with other dental-stop utterances showing markedly rising F2 transitions) revealed that:

- F2 onsets that seemed to cause or contribute to dental perception where not only those that had F2 aspiration or voiced onset possessing a "tone-burst-like" spectrum with steep lower and higher skirts and very thin bandwidth (100-200Hz between -10dB points); rather, what seemed to be crucial was a steep lower skirt (that we tentatively defined as >15-20dB from the lowermost harmonic in F2 to its lower frequency neighbor), although it also seems that too wide a F2 bandwidth will contribute energy to upper inhibitory bands, decreasing the dental-related neural information
- in some cases, non-initial segments of F2 (exhibiting the steep lower skirt characteristic) seemed to contribute to dental perception; this happened when there was an hiatus in F2 excitation (in speakers of very high F0) - this would have the effect of resetting the fast adaptation, enabling a non-initial segment to elicit high firing rates again.

The above conclusions have sufficed to provide a rough perceptual analysis in all cases of dental-stop/front-vowel CVs with markedly ascending F2 transitions. One interesting spin-off of these conclusions regards instances of labial-stop/front-vowel CV's in which the F2 transition is non-ascending. In these cases, it has been always found that the F2 onset does not have the steep-lower-skirt property.

3 - A NEURONAL STRUCTURE ANCILLARY IN THE PERCEPTION OF DENTAL PLACE-RELATED INFORMATION IN F2 TRANSITIONS

In summary, it appears that there is F2-transition-induced neural information for dental when the following processing structure yields high (typical of pre-fast-adaptation behavior) firing rates at hypothetical cochlear nucleus cells.



The neurophysiological processing primitives that are invoked are the following:
- the shape of the receptive fields of auditory-nerve fibers, widening towards lower frequencies
- fast adaptation of the auditory-nerve fibers
- the lateral inhibition that seems to exist in the afferent connections to at least some cochlear nucleus cells

Fig. 5

This structure is being proposed solely for the case of front, high-F2 vowels; in spite of this being a difficult problem for automatic speech recognition, a neurophysiological interpretation is relatively more straightforward in this case. In fact, F1 and F2 are far apart (masking of F2 by F1 is minimal), and it seems that synchrony information does not play a considerable role in the frequency range typical of F2 in these vowels (so that interpretations solely in terms of average rate may more confidently be proposed). Also, the ascribing of the cells in the final column to the cochlear nucleus is made solely on plausibility arguments.

The responses of this structure to various possible situations are roughly described below:

a) A rising F2 transition which does not have an initial F2 (aspirated or voiced) segment with the steep lower skirt property will not cause unadapted firing during the ascending transition, because the non-steep-skirt will provide input to the inhibitory bands of the cells, and also because the continuously ascending F2 will provide a limited time of stimulation near any one CF. During the transition,

below-CF energy will gradually enter the receptive field of the AN fibers located at the stabilized F2 frequency, so that when F2 stabilizes these cannot fire at unadapted rates, and so there is no unadapted input to the (hypothetical) cochlear nucleus cells.

b) A level F2 transition, but one in which the F2 spectral profile includes much energy in the lower skirt (and particularly one which as a gradual onset of F2 energy) will not be able to elicit high cell firing rates because of the excitation of inhibition bands

c) Initial (not too steeply ascending) F2 segments with the steep lower skirt property will provoke firing at unadapted rates in the (hypothetical) cochlear nucleus cells. This is independent of the possible marked rise of F2 afterwards of this initial segment

d) A level or slightly-ascending (and short, preferably) F2 transition, in which the F2 spectral profile has the steep lower skirt property, provides the best combination to elicit unadapted rates. Note that with a slightly ascending transition the F2-energy will be able, from the outset, to supply ample input to the AN-fibers with CF at the stabilized-F2 frequency, because the receptive field of these fibers will collect well this energy.

e) A markedly descending F2 transition will not supply input to the receptive fields (because these have a steep higher-frequency slope) of the AN-fibers with CF at the stabilized-F2 frequency until F2 is nearly stabilized, and then this supply will occur abruptly, allowing the fibers to fire at unadapted rates. Also, the fibers whose CF is visited by F2 during its descent will be able to fire at unadapted rates, since once F2 enters their receptive field it stays there for a considerable time.

f) A long and slowly descending F2 transition seems to be worse, in terms of eliciting unadapted firing, than either a markedly descending or a slightly ascending transitions, because the slowly descending F2 will progressively inject energy into the receptive field of the fibers whose CFs are at the stabilized F2 frequency, degrading the chances of unadapted firing there. Only the fibers visited during the descent stand to fire at unadapted rates.

It is to be emphasized that we do not claim that this processing structure is the only one involved in creating dental-related neural information. It was mentioned, for instance, that upward inflections in the spectral density function located in high-frequency (>4KHz) regions at the onset of the sound are a (well-known) contribution towards dental perception. However, even this contribution might be explained by a neuronal structure similar to the one sketched above (one interesting difference is that at higher frequencies the importance of higher inhibitory bands appears to be smaller).

As a final note, it has not been explained how neural information for P is created. Other experiments conducted on our research corpus revealed that P-information is created by a simple ascending sequence or contrast: Low BEFORE phase , followed by: High AFTER phase. This contrast may be created without any continuously ascending movement in the formants. For instance, if one takes a /ti/ with a level F2 transition, but without the steep lower skirt property, it is fairly easy to obtain a sound that is identified as /pi/ by erasing the initial part of F3 until it shows a lag of about 40ms relative to F2 onset. Then, the late onset of F3 appears to be taken as "new higher-frequency event in the AFTER phase", and it dictates perception of /p/. When it is the lower edge of F2 that rises in the AFTER phase, P-information is much stronger.

REFERENCES

[1] P. Cosi *et al* (1994) - *Speaker independent phonetic recognition using auditory modeling and recurrent neural networks* - ICANN 94 - International Conference on Artificial Neural Networks, Sorrento (Italy), 1994

[2] P. Delattre *et al* (1955) *Acoustic loci and transitional cues for consonants*, J. Acoust. Soc.Am. **2 7**, 769-773

[3] M. Dorman *et al* (1977) *Stop-consonant recognition: release bursts and formant transitions as functionally equivalent, context-dependent cues*, Perc. & Psychophys., **2 2**, 109-122

[4] A. Lahiri *et al* (1984) - *A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: evidence from a cross-language study* - J. Acoust. Soc.Am. **7 6**, 391-404

[5] A. Liberman, I. Mattingly (1985)- *The motor theory of speech perception revisited* - Cognition 21, 1-36

[6] P. De Michelis *et al* (1983) *Computer recognition of plosive sounds using contextual information*, IEEE Transactions Acoust., Speech, and Signal Proc. , **3 1**, 359-377

[7] R. De Mori, G. Flammia (1993) *Speaker-independent consonant classification in continuous speech with distinctive features and neural networks*, J. Acoust. Soc.Am. **9 4**, 3091-3103

[8] S. Blumstein and K. Stevens (1980) - *Perceptual invariance and onset spectra for stop consonants in different vowel environments* J. Acoust. Soc.Am. **6 7**, 648-662

[9] A. Waibel *et al, Phoneme recognition using time-delay neutral networks* (1987), ATR - Interpreting Telephony Research Laboratories Int. Pub. 1987.10