AUDITORY CELLS WITH FREQUENCY RESOLUTION SHARPER THAN CRITICAL BANDS PLAY A ROLE IN STOP CONSONANT PERCEPTION: EVIDENCE FROM CROSS-LANGUAGE RECOGNITION EXPERIMENTS

Eduardo Sá Marta, Luís V. Sá
Instituto de Telecomunicações - Pólo de Coimbra
Dept. Eng. Electrotécnica - Pólo II da Univ. Coimbra
3030 COIMBRA – PORTUGAL
emarta@it.uc.pt

ABSTRACT

A metric is presented which expresses the auditory detectability of spectral peaks with bandwidths smaller than one third of a critical band, extending for 8-20 milliseconds, prior to vowel onset in stop-consonant-vowel CVs. The metric incorporates characteristics, such as strong lateral inhibition and marked non-monotonicity in amplitude-response functions, which are evocative of cells with "pencil" and "spindle"-shaped frequency tuning curves, that have been studied mostly in the bat. It is found that medium to high values of the metric are found in 30% to 60% of DENTAL sounds, depending on the language. In contrast, almost no LABIAL sounds exhibit metric values in this range, which suggests that this is a significant auditory cue for the DENTAL versus LABIAL discrimination of such consonants.

1. INTRODUCTION. MOTIVATIONS

The findings reported in this paper are part of an undertaking to build a model of human perception of non-noisy CV's when C=stop consonant and V=high-F2 vowel, aiming to achieve human-like levels of performance in place discrimination. This circumscribed task was selected because: (i) the "high-F2" (>1.7K) restriction means that average-rate auditory representations are adequate, and interactions between the auditory representations of F1 and F2-F3 are simpler to estimate; (ii) for automatic alphabet recognition, the discrimination between consonants in the "E-set" has proved extremely difficult - thus, success in achieving human-like performance cannot be attributed to the low difficulty of the task; (iii) publicly available alphabet/spelling speech databases provide a very large number of exemplars pertaining to this task, and in many languages; (iv) success in this undertaking will be immediately useable in improving automatic recognition of spelled letters.
By (hopefully) achieving human-like performance in this circumscribed but difficult task, it is expected that some foothold will be gained in the general task of modeling human phonemic perception.
One aspect of human performance is the good, albeit not perfect, recognition of consonants (including stop consonants) uttered by speakers of a different native language. For instance, in [1] it is reported that native Korean listeners discriminate, with less than 1% errors, 3-way place (labial-dental-velar) of stop consonants uttered by American English speakers. Testing a perception model with the same consonants, uttered by speakers of different languages, clearly moves beyond the acoustic regularities that hold only within each particular language, and may provide convincing evidence for the model correctness.
Another, even more important, aspect of human performance, is the graceful degradation in correct recognition scores exhibited when going from full-band to band-pass filtered (e.g., telephone-like) speech. Clearly, any valid model of human perception must also exhibit such behavior.

In this paper, auditorily-formulated models for cues underlying the discrimination of DENTAL against LABIAL place will be tested with exemplars uttered by speakers of 3 different languages: Portuguese (33 speakers), American English (30 speakers) and German (50 speakers). Additionally, testing is done for full-band speech and for speech low-pass filtered at 3.5KHz. Models were refined using only Portuguese sounds and applied unaltered to other languages.

From psychophysical pilot experiments [2] it was hypothesized that one important cue discriminating DENTAL against LABIAL (and, to a slightly lesser degree, against VELAR/GLOTAL) place is the existence of initial "tone-burst-like" segments of very thin bandwidth, usually corresponding to the aspiration phase of the second formant (F2) or third formant (F3). Later, it was realized that auditory neurons with more than adequate capability for detecting such segments have been extensively studied in the bat, by Suga and other researchers [4,5]. Because of their high frequency selectivity, these neurons have been described as having "pencil or spindle-shaped" tuning curves, and because they maintain this selectivity even for high intensity levels they have been termed "level-tolerant". In the auditory cortex of bats, neurons with Q-50dB values in excess of 350 have been found. Such high Q values seem to exist only for the most ecologically relevant frequencies for bats, but Suga and colleagues point out [4] that:

...We believe that our data are important for understanding the neural basis of frequency difference limens in general because they demonstrate most dramatically that the mammalian auditory system has the capacity to produce sharply tuned and level-tolerant neurons. These neurons may be the neural basis for small frequency difference limens at high sensation levels...

The present paper thus reports on a model of auditory detection of initial "tone-burst-like" segments of very thin bandwidth as presumably used for stop PLACE discrimination, assuming that in the auditory system of human listeners there exist neurons with "level-tolerant" characteristics, and "pencil/spindle" frequency tuning curves. But only much more modest values of Q are assumed, in the order of 10-20.

This acoustical cue is not sufficient to explain human discrimination performance, and integration of this cue with others is also discussed.

2. GENERAL ASSUMPTIONS ABOUT HUMAN COMMUNICATION OF PHONEMES

Although in word communication, or continuous speech communication, it is debatable whether phonemes are the most relevant speech communication unit, there are instances – namely spelling and communication of monosyllabic nonsense words – in which humans evidence the capability of speaker-independent phoneme communication. The mechanisms underlying this capability are, of course, also of importance in spoken word or sentence communication – although in those cases they may be marshaled to the communication of other units.

Bearing in mind those most simple communication tasks, our most general assumption about human phoneme communication may be expressed by 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 leads us to assume that

- acoustic/auditory cues have a role that is centrifugal relative to between-categories boundaries
- there are multiple cues for each phonemic distinction, and there exist, between them, trading relations that may extend to the point of alternativity

This view, while at least partially shared by other authors (see for instance [6]), is directly in confrontation with the seemingly more prevalent assumption that for each phoneme there is a "nativistic" articulatory program, and successful phoneme communication is hinged on the adherence of the speaker (albeit with some random imprecision) to that articulatory program, and on very powerful "decoding" or even "tracking" capabilities on the part of the listener [7].

Arguments in support of our view include:

(i) We are assuming that in speech production acquisition by a child, the drive for articulatory correctness is an "evolutionary force" (that acts through unacceptance by listeners interacting with the child) that decreases in importance when gross articulatory correctness (for a given phoneme) is achieved. Within the vicinity of gross articulatory correctness, auditory feedback may well be the only significant "evolutionary force" and this auditory feedback incorporates trading relations between cues. This means that the new speaker may "rest satisfied" when he/she learns to emit sufficient cues with substantial intensity. Different speakers may end up with different mixtures of cues.

(ii) About the question of how the cues are "seeded": Cues may be seeded by articulation; very forceful production of a phoneme may entail emission, in good intensity, of the complete, or nearly complete, set of cues. The evolution of languages may, in turn, have privileged phonemes (or rather, phonemic oppositions) that are rich in auditorily represented cues [8]. For instance, if a phoneme is most commonly perceived through some particular auditorily-represented signal, but if a slight articulatory mishap results in an acoustical form that is also well detected, but through other auditory signals, this phoneme will be adequate for robust communication (it is perceived in "either... or..." fashion).

(iii) It has been demonstrated that humans fitted (for experiments) with byte blocks or lip tubes that prevent them from using the normal articulation for some phonemes are often able to evolve alternative articulatory schemes that achieve correct recognition by listeners; the same goes for many of the persons that have parts of their tongues surgically removed [9]. In radiological studies, it is found that different speakers, for a given phoneme, may use very different "articulatory programs" [10].

3. A FUZZY-LOGICAL MODEL OF NEURONS WITH SHARP FREQUENCY TUNING

3.1 Refinement criteria

As mentioned, from pilot psychophysical experiments and from hypotheses inspired by extensive spectrogram reading, we are assuming that firing by such neurons will constitute an information carrier for DENTAL versus LABIAL. Since we are assuming the existence of multiple cues, or information carriers, for each phonemic distinction, we do not require that each and all DENTAL exemplars will evidence high levels of firing by these neurons. On the other hand, even without positing any particular form of integration with other information carriers, it may be expected:

- a fair proportion (some tens of percent) of the DENTAL exemplars will show significant to high values in any coarsely correct metric expressing firing by those neurons
- ideally, no LABIAL exemplars should show significant to high values in such a metric (that is, the significant to high range of values should be exclusively DENTAL); still, if a small number of LABIAL exemplars show high values, then it must be ascertained that these particular exemplars also possess very high values in a metric(s) expressing LABIAL information carriers (so that it may be assumed that sufficient correct information exists to base the listeners' correct identification).

3.2 The model

Knowledge acquired from pilot experiments showed that the thin-bandwidth segments seemed to be phonemically valid only before the onset of the vowel in the CV. More specifically, there seemed to be a window of phonemic validity, which was closed by the first acoustical event that substantially excited onset-type cells (this acoustical event may be defined as a synchronous onset of excitation over a wide range of frequencies). Such an event might in some sounds occur much before the onset of the vowel; in some rare sounds, the onset of the vowel might be so gradual, with excitation at first limited only to the F1-region, that it might be surmised that onset cells would not be substantially excited by the vowel onset.

These dependencies have not yet been given a detailed implementation in the present model; phonemic validity of the thin-bandwidth segments is simply taken as uniformly maximal before the vowel onset and nil after that.

One very important characteristic of the sharp-frequency-tuned, level-tolerant neurons reported in [5] is the drastic non-monotonicity of their amplitude-response function. For one such neuron reported in [5], maximal firing rate is elicited at a stimulus amplitude of 50dB SPL, but at 75dB SPL the firing rate is reduced to little higher than the spontaneous rate.

This non-monotonicity implies that normalization should aim to recover the level of the sound at a normal dialogue distance, as estimated by the speaker – the "autophonic scale". Even for simple CV sounds, there seems to be no sufficient information in the literature to base this normalization. We used the expeditious step of using as the normalizable measure the maximum of the average (over 40-ms) of the energy integrated over the 250-5500Hz range; a subsequent correction was made in the rare cases where the energy in the 1500-3500Hz range was smaller by more than a 25dB difference (this is to ensure that F2-F4-F5 always have a minimum impact on the normalization process). This measure was normalized to the arbitrary value of 72dB (which was simply the average value found over the first set of sounds so processed).

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. For each such point where F∈(1600Hz,4000Hz), a (local) **Raw Bilateral Prominence Index** is calculated:

$$\mathbf{BiPr}(f,t) = \mathbf{Sat}(E(f,t)) -$$
$$- Max\Big(\underset{f'=f-2}{\overset{f-4}{Max}}(E(f',t)),\ \underset{f'=f+2}{\overset{f+4}{Max}}(E(f',t)) - 4dB\Big)$$

Sat is a saturation function; onset of saturation was assumed to occur at 51dB, and hard saturation to take effect at 54dB. These parameters remain to be optimized; first shots were estimated from pilot psychophysical experiments. The **BiPr**(f,t) values were zeroed when occurring after the onset of vowel and subject to fuzzy intersection with five μ factors expressing:

1- the local energy **E**(f,t); a locally prominent peak, but one of negligible energy is not perceivable
2- the degree to which Backward Masking may obliterate the perceivability of the prominent peak
3- the degree to which lower-frequency energy within the same critical band, occurring simultaneously or briefly before, may suppress/inhibit/adapt excitation of the modeled neurons
4- the degree to lower-frequency energy within the 2 lower critical bands, occurring simultaneously or briefly before, may suppress/inhibit/adapt excitation of the modeled neurons
5- the degree to which input to the modeled neurons may be adapted by energy in the same critical band occurring briefly before

The fuzzy intersection operator selected was the product. That is, the final intersective factor is

$$\mu = \mu 1 \times \mu 2 \times \mu 3 \times \mu 4 \times \mu 5$$

For each of these μ factors a lifted (starting at 0.4 instead of 0.0) simple Γ function was used, which depends only on two parameters L1 and L2. The physical parameter was always an energy measure or a difference between two energies. As an example, $\mu 3$(f,t) took the form

$$\mu 3(f,t) = \left| \begin{array}{ll} \min(1; 0.4 + 0.6 \times \dfrac{[Sat(E(f,t)) - ELFP(f,t)] - (L1)}{L2 - L1}) & \\ \quad , \text{IF} \quad Sat(E(f,t)) - ELFP(f,t) \geq L1 & \\ 0 \quad , \text{ELSE} & \end{array} \right.$$

ELFP(f,t) is the **E**nergy **L**ower in **F**requency, (temporally and frequencially) **P**roximal.

L1 and L2 values were refined only for the Portuguese sounds. The aims for refinement were not so much to obtain a minimum classification error, but to be "perceptually fair". That is, sounds which yielded metric results not in the best accordance with their PLACE were visually examined or subject to a few editing/audition experiments. For instance, a DENTAL exemplar might have a zero value in this metric, and this might still be judged as perceptually fair if it was ascertained that this sound possessed other information carriers for DENTAL and that initial thin-bandwidth segments seemed to be absent from this sound.

Finally, measures S3, S4, S5 and S6 were obtained to express the maximum average activity of neurones at the same frequency f or 2 contiguous frequencies f and f+1 over 3, 4, 5 and 6 signal frames. For each of these measures an exclusively-DENTAL range was estimated, and this in turn allowed fuzzyfication of these measures.

4. CLASSIFICATION RESULTS

4.1 **Exclusively-DENTAL ranges for full-band stimuli**

The histograms for the metric values elicited by 66 /ti/ and /pi/ sounds from 33 Portuguese speakers (from an in-house research database) are shown in *Figure1*. The abscissa is the fuzzy union of the fuzzy variables corresponding to S3, S4, S5 and S6.

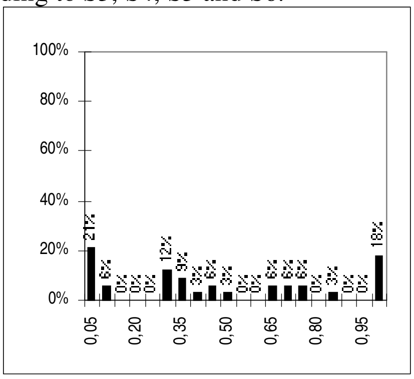

Figure 1A – Histogram for 33 Portuguese /ti/ sounds

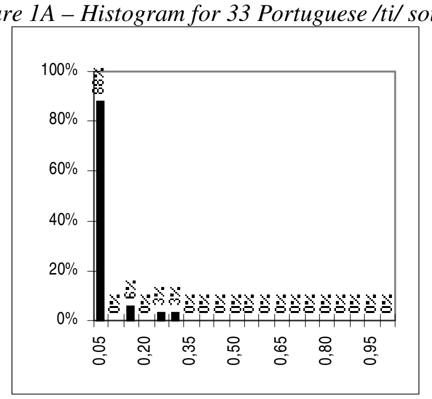

Figure 1B – Histogram for 33 Portuguese /pi/ sounds

It can be observed an exclusively DENTAL range for the metric: 60.6% of DENTALs had a metric value higher than the highest NON-DENTAL (LABIAL). 86% of the NON-DENTALs have very low values (<0.05).

In *Figure 2A* and *Figure 2B* are shown histograms for the metric values elicited by 120 /di/ and /bi/ sounds from 30 American English speakers (first set of the ISOLET Spoken Letter Database – Oregon Graduate Institute). /di/ and /bi/ sounds were chosen instead of /pi/ and /ti/ because American English /pi/ and /ti/ typically show very long (often extending for more than 100ms) and intense *burst+aspiration* segments; for such segments more complex phenomena such as local normalization and adaptation with a temporally non-proximal adaptor must be brought into the model.

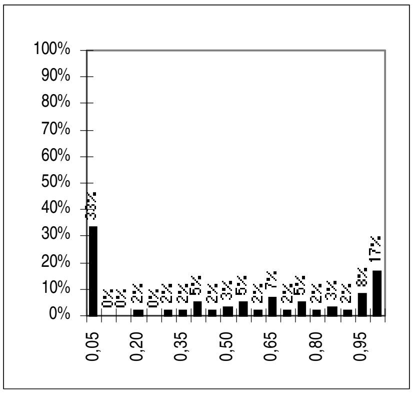

Figure 2A – Histogram for 60 American English /di/ sounds

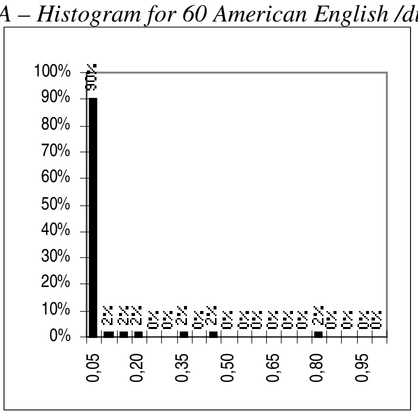

Figure 2B – Histogram for 60 American English /bi/ sounds

If we exclude a lone /bi/ sound showing a value near 0.8, there is also a satisfactory exclusively DENTAL range, starting at approximately μ=0.4, which contains 57% of the DENTAL exemplars. The lone /bi/ sound referred is mteb0B1 (CSLU/OGI labels).

Finally, *Figure 3A* and *Figure 3B* show the histograms for the metric values elicited by 100 /bi/ and /di/ sounds from the first 50 speakers, in alphabetical order, of the PhonData1 database - Bavarian Archive for Speech Signals. For these German sounds there is also a satisfactory exclusively DENTAL range, starting at approximately μ=0.5, but which contains only 30% of the DENTAL exemplars

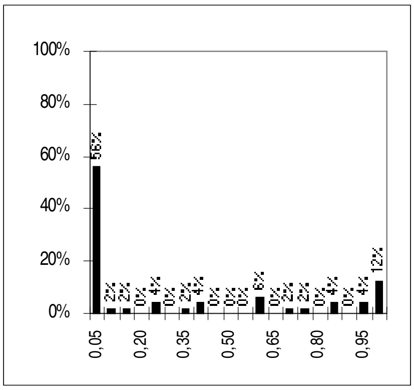

Figure 3A – Histogram for 50 German /di/ sounds

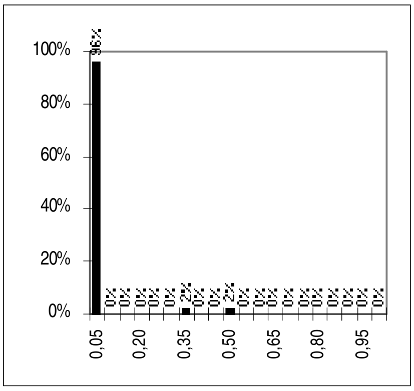

Figure 3B – Histogram for 50 German /bi/ sounds

4.2 **Integration with other auditorily-represented information carriers (cues) for full-band speech.**

In [3], a fuzzy-logical model of ascending sequence cells is developed which was shown to yield a very satisfactory exclusively LABIAL range. This model has since been improved, and histograms for the same American English sounds as in *Figure 2* are shown in *Figure 4*:

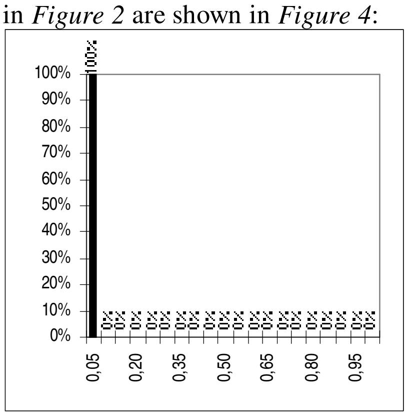

Figure 4A – Ascending-sequence histogram for 60 American English /di/ sounds

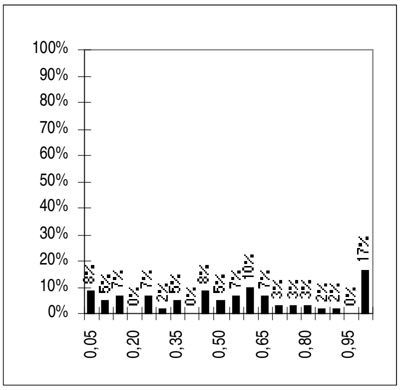

Figure 4B – Ascending-sequence histogram for 60 American English /bi/ sounds

In these histograms, a clear <u>exclusively-LABIAL</u> range can be seen. Other two information carriers with plausible auditory representations were considered:

- Information Carrier LABIAL-IC2 : absence of unvoiced energy prior to vowel onset
- Information Carrier DENTAL-IC2: high-frequency (>3.5KHz) inflection in the spectrum

Fuzzy-logical expressions for these 4 cues were integrated by simple fuzzy union and intersection operators (other cues, already identified, remain to be implemented) The final fuzzy value that expressed LABIAL (intersected by DENTAL) was compared to various thresholds to obtain error scores. These are shown, for Portuguese and American sounds, in *Table 1*.

	Error score	
Threshold	**Portuguese**	**American**
0,05	1,5%	8,3%
0,1	1,5%	6,7%
0,2	3,0%	5,0%
0,3	4,5%	5,0%
0,4	9,1%	4,2%
0,5	9,1%	5,0%
0,6	15,2%	6,7%

Table 1 – Error scores for various thresholds(full-band speech)

Error scores are seen to be low, especially considering that they were obtained using only a part of the auditorily-represented cues that were identified. Moreover, the implemented cues have been very coarsely modelled.

4.3 **Integration with cues for pass-band speech.**

The error scores for the same sounds as in *Table1*, but now low-pass filtered at 3.5KHz (abrupt cut-off) are presented in *Table2*. It should be stressed that the model used is still the same model as refined for Portuguese sounds in full-band conditions. No parameters were changed, which is somewhat unrealistic: it is quite plausible that listeners, when faced with distinctly low-pass sounds, change their expectations about the relevant auditory “quantities”.

	Error score	
Threshold	**Portuguese**	**American**
0,05	13,6%	8,3%
0,1	12,1%	7,5%
0,2	12,1%	5,8%
0,3	10,6%	6,7%
0,4	10,6%	9,2%
0,5	16,7%	11,7%
0,6	18,2%	15,8%

Table 2 –Error scores for various thresholds(pass-band speech)

It can be observed that American sounds are still classified very well. Portuguese sounds seem to require more extensive modeling, including a gating role for onset-type cells. This is currently under way.

ACKNOWLEDGMENTS

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. Part 1. Korean perception of English”, J.Acoust. Soc. Am. 99 (5), pp.3201-3211, 1996

[2] 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, Madrid, Spain, 1995

[3] Eduardo Sá Marta, Luis Vieira de Sá, “Impact of ascending sequence AI (auditory primary cortex) cells on stop consonant perception”, 5th European Conference on Speech Communication and Technology, Rhodes, Greece, 1997

[4] Suga,N. and Tsuzuki,K., “Inhibition and Level-Tolerant Frequency Tuning in the Auditory Cortex of the Mustached Bat”, Journal of Neurophysiology,Vol.53, pp.1109-1145, 1985

[5] 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

[6] Smits,R., Bosch,L., Collier,R., “Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. II. Modeling and evaluations” J.Acoust. Soc. Am. 100 (6), pp.3865-3881, 1996

[7] A.M. Liberman and J.G.Mattingly, “The motor theory of speech perception revised”, Cognition, 21, pp.1-36, 1985

[8] Ohala,J., “Speech perception is hearing sounds, not tongues”, J.Acoust. Soc. Am. 99 (3), pp.1718-1725, 1996

[9] C. Savariaux, P. Perrier, J. Orliaguet, “Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production” – J.Acoust. Soc. Am. 98, pp.2428-2442, 1995

[10] K. Johnson, P. Ladefoged, M. Lindau, “Individual differences in vowel production” - J.Acoust. Soc. Am. 94, pp.701-714, 1993