How can the control of the vocal tract limit the speaker's capability to produce the ultimate perceptive objectives of speech ?

Christophe Savariaux, Louis-Jean Boë & Pascal Perrier
Institut de la Communication Parlée - UPRESA CNRS 5009
INPG & Université Stendhal
46 Avenue Félix Viallet - F - 38031 Grenoble Cédex 01 - France
savario; boe, perrier@cristal.icp.grenet.fr

Abstract

In this paper an extension of the lip-tube experiment proposed by Savariaux et al. (1990) is presented and analyzed. The question underlying the design of this experiment is whether speakers are able to produce an [u] with a large lip opening. Nine native speakers of French repeated the original experiment, and then were asked to produce the vowel [u] starting from [o] vocal tract configuration. It was shown that more subjects achieved the compensation when they shifted their articulation from [o] to [u]. The issue of a possible constraint imposed by a learned standard articulatory pattern is discussed in relation with the notion of the internal representation of the articulatory-to-acoustic relations. Proposals in favor of a standard pattern for [u] that would be velopalatal rather than velopharyngeal are discussed.

1. Introduction

The respective weight of the acoustic/auditory and articulatory domains in the definition of speech production objectives is still a matter of large debate. To contribute to the discussion, we present results of an experiment that was designed as an extension of the original lip-tube experiment proposed by Savariaux et al. (1995a).

2. Lip-tube perturbation of vowel [u] production
(Savariaux et al., 1995a)

2.a. Theoretical Background.

In an earlier work we presented an experiment, in which a labial perturbation was applied to speakers during the production of the French vowel [u], as a mean of testing the respective weights of the articulatory and acoustic levels in the control of vowel production. Indeed, using Fant's new model (Fant, 1992), it was demonstrated that it is theoretically possible to produce the vowel [u] from two very different vocal tract configurations giving the same F1/F2 pattern. One of these configurations is the one that was systematically observed in the articulatory typology of languages in the world (see Ladefoged & Maddieson, 1996): the articulation is located in the velo-palatal part of the vocal tract, and the lip area has to be small (around 0.3 cm^2). The other one corresponds to a velo-pharyngeal articulation and a large lip area is possible. To our knowledge there are no articulatory data, attesting that this latter configuration was ever observed for vowel [u] in any language in the world. However simulations made with Maeda's statistical model, derived from X-ray sagittal views of the vocal tract, suggested that it is anatomically possible for a human subject to produce such a vocal tract shape, if a perturbation prevents him to achieve the former one.

2.b. Experimental Setup.

In this context, a 4.9 mm^2 section lip-tube (25-mm diameter) was inserted between the lips of eleven French native speakers, as they were asked to produce an isolated French rounded vowel [u]. To prevent the jaw from moving down when the tube was inserted, the subjects had, in all conditions, a small bite-block between the teeth. Thus the jaw position was kept constant across repetitions and conditions. The subjects were recorded under three conditions: (i) without lip-tube, Normal (N) condition; (ii) with the tube, immediately after its insertion between the lips, Perturbed First (PF) condition; (iii) finally, with the tube, after a learning procedure of 19 trials, Perturbed Last (PL) condition. To evaluate the compensation capability of each speaker, the acoustic signal was recorded, and F1/F2 patterns were estimated by way of an LPC analysis. The articulatory strategies were analyzed from X-ray pictures in the mid-sagittal plane.

2.c Data analysis in the acoustic domain

A first analysis of the results, based on the observation of F1/F2 patterns, was presented in Savariaux et al. (1995a). It showed that only one out of eleven speakers (speaker OD) was able to compensate for the perturbation in the F1/F2 plane. For that purpose, as suggested from the theory, he moved his tongue backwards into the pharyngeal cavity, inducing a strong change in the constriction location. Remaining speakers showed a large variability: four speakers presented no relevant articulatory changes, and then no compensation effects, while six of them presented variable extents of tongue backward movements within the palatal region. Given the articulatory changes observed on speaker OD we concluded that the inability to compensate, which was observed for the large majority of subjects, was not the result of any physical or physiological limitation. We suggested that the explanation of this phenomenon should rather be found at the level of the control of the articulators. In addition, it was interesting to observe that for the whole set of speakers, the observed articulatory changes, if any, were all directed toward an enhancement of the sound in the F1/F2 plane. Hence we proposed that the subjects have a good sense of what the auditory effect should be, and that the goal of speech control be the production of this auditory objective. Finally it was noted that the large majority of subjects essentially kept producing a vocal tract configuration close to the normal one even with the tube.

Hence at this stage of the analysis it was suggested that, to produce a vowel, the speakers could have learned to systematically recruit a standard vocal tract configuration, that would be associated under normal speech condition with the intended perceptual objective. In perturbed speech, once stated the inadequacy of this standard configuration, the ability to find another, more appropriate strategy, would depend on the articulatory skill of each speaker.

2.d Perceptual analysis

In a second paper (Savariaux et al., 1995b), we refined this analysis by running a test to evaluate the perception effect produced by both normal and perturbed vowels. In this test, 17 listeners were asked to rate the quality of sounds [u] on a scale going from 1 (not an [u] at all) through 7 (a very good [u]). Results were quite surprising. Indeed, the sound produced by speaker OD in PL condition was not rated at the highest level. At the same time, the stimuli produced in the same condition by three other speakers were perceived as very good vowels [u]. Hence it appeared that these subjects were successful in producing the right perceptual effect, even if they provided only little changes in the geometry of their vocal tract, and even if, consequently, the compensation was not achieved in F1/F2 space.

Additional spectral analyses suggested that F1/(F2-F0) space would be the relevant acoustical domain to characterize the perceptual quality of sound [u]. The quality is good as long as these two spectral parameters are small enough. The perceptual objective of the vowel production should thus not only be described in terms of formants, the vocal-tract resonances. It would rather correspond to a combination of features partly related to control of the vocal tract and partly related to the control of the vocal source.

Consequently, using the degrees of freedom allowed by the complexity of the perceptual objective for the vowels, the speakers would have different ways to compensate for the perturbation. In addition the impact of the perturbation induced by the tube can be suspected not to be identical for all subjects, since, in particular, a high intrinsic F0 frequency could reduce the perceptual changes associated with the increase of F2 induced by the tube.

This statement also explains why subject OD produced the observed strong reorganization of its articulation. In natural condition, his (F2-F0) value is indeed fairly high; he could then have preferred to reduce the increase of F2, to keep F2-F0 low enough. Additional perceptual tests effectively confirmed that OD adopted a good compensation strategy, since the strong backward movement of his tongue was clearly associated with a perceptual enhancement, even if the final result was not as good as in natural condition.

2.e Perceptual objectives and articulatory limitation.

Thus, the perceptual analysis strengthens our first conclusion that speech production is oriented toward perceptual objectives.

At the same time, these data show that 4 speakers, and not only one as suggested at a first glance from a pure acoustical analysis, were skilled enough to achieve different kinds of compensation. Therefore being able to compensate for the lip-tube perturbation happens not to be a peculiarity of a specific subject. Hence it is worth questioning back our original hypothesis, stating that standard learned articulatory patterns would limit the ability of each speaker to deal with the perturbation (see 2.c). Indeed since more than a third of the subjects were able to overpass this potential limitation, is it a very loose constraint? Could the behavior of the 7 remaining subjects not be explained by other factors such as, for instance, a lacking internal representation (a forward model according to Jordan & Rumelhart, 1992, or Kawato et al., 1987) of the relations between articulatory and perceptual changes, that would impair the chances for the subject to find the appropriate compensation strategy.

To further study this issue, a new experimental protocol was setup, in which different initial articulatory configurations were imposed to the speakers when they tried to compensate for the lip perturbation. The basic idea underlying this new experiment can be summarized as follows: if subjects failed in compensating because of lacks of their forward model, inciting them to start from an articulatory configuration that is close to the desired one should help them in finding the compensation strategy.

3. Lip-tube compensation from faciliting initial articulatory configurations (Cocusse, 1996)

3.a. Experimental setup.

Nine new male subjects were asked to go through the three stages of the original experiment. This time, however, they were asked to keep their pitch constant, F0, during each vowel production, and to adopt the same pitch over all conditions. To do so, in a preliminary session, the subject was asked to produce an [u] at his usual F0. Then during the experiment, a pure sound that had been tuned to this natural pitch, was emitted as the subject was just on the way to produce the vowel. The subject was then asked to adjust his F0 to this tone, and to maintain it during the whole production of the vowel.

It should be noted that in this experiment only the acoustic speech signal was recorded. No X-ray data were collected. Because of the pitch control imposed to the subjects, the assumption was made that observing the changes in F1/F2 patterns should be a reliable way to evaluate both the perceptual quality of the sound and the articulatory changes provided by the subjects. This was made to prevent unnecessary exposure of subjects to X-ray radiation.

Recordings have been made in a sound-treated room. Just like in the original lip-tube experiment, a small bite-block was put between the speaker’s teeth to keep the jaw position constant. Within a unique session, each subject was asked to produce the vowel [u] under five different conditions. The first three conditions were similar to the three ones of the original experiment: N, PF, and PL conditions. It should be noted that the training phase in PL condition was reduced from 19 to 10 trials. Three additional perturbed conditions were investigated. In two of them the speakers was asked to pronounce an [o] and to gradually shift toward an [u]. The first production (Perturbed First with Facilitation, PFF, condition) and the production following a 10 trials training session (Perturbed Last with Facilitation, PLF, condition) were analyzed. In the last condition (Perturbed Word, PW, condition) the speakers had to produce the nonsense word [ogu]. In each condition the speakers were asked to try to keep their pitch constant.

It should be noted that the contexts provided by vowel [o] in PFF and PLF conditions, and by syllable [og] in PW condition, has been selected because they induce the tongue to be set back in the mouth just before the vowel [u] is articulated. Thus under these conditions, the tongue shape at the beginning of the movement toward vowel [u] is close to the shape that is predicted, from the acoustical theory, to be associated with a good [u] with a lip-tube. Assuming that finding the right compensation strategy could depend upon the power of the forward model, the hypothesis underlying the specification of these three conditions is that being close to the optimal configuration could help finding it. Hence PFF, PLF and PW are considered to be perturbed conditions with a “facilitating phonetic context”.

The acoustical signal was sampled at 16 kHz. To extract the spectral properties of the sound, an LPC analysis has been carried out, to find the maxima of the spectrum. Then, the first three formants have been calculated as the mean values, over 500ms, of the first three spectrum maxima. When the relative difference between the formants measured in a perturbed condition and those obtained in N condition was smaller than 10%, the formant patterns are said to be “similar”, and the compensation is said to be “achieved”.

3.b Results

The results obtained in PF and PL conditions are conformable to those of the original experiment (Savariaux et al., 1995a):

- No compensation was achieved by any subject without a training session.
- At the end of this session, only one subject (LG) was able to produce an F1/F2 pattern similar to the pattern measured in N condition.

As concerns the perturbed conditions with facilitation it is very interesting to observe that 2 additional subjects (AR and TG) have been able to achieve the compensation after the training session, in PLF condition, while three other speakers have also provided a significant enhancement of the F1/F2 formant patterns in comparison to PL condition.

It should be noted that the improvement does not seem to be imputable to the fact that speakers were getting used to produce an [u] with the lip-tube. Indeed for 5 among 9 speakers, including AR and TG, the formant patterns measured in PFF condition were worse than in PL condition. Thus starting from an imposed articulatory configuration seems to require from the speakers to define a specific compensation articulatory strategy; the training phase preceding PL condition would then not have any noticeable influence on the way the speakers behave in the training session that precedes PLF condition.

3.c Discussion

Thus, for 5 among the 8 speakers who did not achieve the compensation in PL condition, the context of vowel [o] seems to have provided a valuable help, to enhance the quality of the vowel [u] pronounced with the lip-tube. This observation does not support our initial hypothesis that the articulation of an [u] would in priority recruit a standard learned velo-palatal articulatory pattern that would limit the ability of each speaker to compensate for the lip-tube. It rather supports the idea that an insufficient internal representation of the articulatory-to-acoustic relations would prevent the majority of speakers from finding the right compensation strategy. From this perspective, our observations can be explained as follows: because the articulation of [o] is close to the velopharyngeal articulation that is necessary to produce a good [u] with opened lips, the optimization problem consisting in finding the right compensation strategy, is more local than for any other context, and its resolution requires consequently a less sophisticated internal model.

Two sets of simulation involving two different kinds of internal model argue in favor of this latter hypothesis. Indeed, using an internal model describing the local sensitivities of the articulatory-to-acoustic relations, that were inferred from Maeda’s articulatory model of the vocal tract, we have run an inversion procedure based on a Gradient-optimization method. The aim of the procedure was to find the tongue position that allows to produce a standard [u] formant pattern with a large lip area. Different initial articulatory configurations were tested. It was found (Bentoumi, 1996) that, when the initial vocal tract shape corresponded to a front palatal articulation (as for instance for [i], [e] [ɛ]), the algorithm did not find the compensation strategy. When the initial shape corresponded to a back pharyngeal articulation ([ɔ] or [a]), the velo-pharyngeal compensation was found. This tends to confirm that a simple local forward model could not be powerful enough to systematically permit the speaker to find the way to compensate. In parallel Guenther and his colleagues (Guenther et al., 1997) have also run an inversion based on the same principles. However their internal model, built through an extensive “babbling phase” (see Guenther et al., 1995), provided a larger description of the articulatory-to-acoustic relations. Their controller has systematically converged toward the optimal velo-pharyngeal compensation strategy.

Thus, it seems that the behaviors of the subjects observed during the different lip-tube experiments can be explained by a combination of two factors: (1) the quality of the forward model, and (2) chance that induces variability in the initial articulatory configuration. Hence, it is no longer useful to assume that a standard learned articulatory pattern would in general limit the capability of the speakers to compensate.

However we still think that such a learned pattern exists, and can influence the way a speaker is looking for the compensation strategy. Our opinion is justified by three observations, that raise three questions:

- In the original lip-tube experiment (Savariaux et al., 1995b) three speakers did not compensate at all in the perceptive space for the lip perturbation, although their vocal tract shapes were very close to the standard velopalatal one. Since this configuration is not optimal at all, why should an optimization strategy converge toward such a solution?
- In the current experiment, in spite of the facilitating [o] context, 3 subjects did not find the compensation strategy either. How is it possible to explain that, even with a very rudimentary forward model, they did not find the right solution, since it was a very local optimization problem?
- The velo-pharyngeal [u] was never observed in French, whatever the phonetic context. How is it possible to explain, that in the context of a back vowel the coarticulation effects do not result in a velopharyngeal articulation of vowel [u]?

Our hypothesis is that, at one level of the representation that a speaker has of the speaking task, there is a standard learned vocal tract shape associated with each vowel. This would not be a constraint of the production, but rather an attractor in the articulatory space that would help finding the vocal tract shape appropriate for each vowel. If it is possible to reach this configuration, the speaker will produce it preferably to any other one. If it is not reachable, then the speaker will try to find another solution using the internal model; if no other solution can be found, the standard shape would be produced.

4. How can we explain that the learned articulatory pattern for [u] is the velopalatal one?

In this perspective, a question arises: Why is, for vowel [u], the learned articulatory pattern the velopalatal one and not the velopharyngeal one? Perkell (1996) would suggest that a minimization of an articulatory cost could be at the origin of the choice. Indeed using a biomechanical tongue model, he suggested that it would be harder for a speaker to produce a constriction in the pharynx than in the palatal part of the vocal tract. However the velopharyngeal configuration of vowel [u] is too similar to vowels [o] and [ɔ], that are very common in French, to agree with such an explanation.

However the preference could be due to perceptual constraints. Of course, both [u] have the same F1/F2 formant pattern.

However they don't have necessarily the same formant-to-cavity affiliations.

Bentoumi (1996) has carried out a set of simulations with four tube approximations of two different vocal tract area functions that give a standard F1/F2 [u] pattern. The one was velopalatal, and the second velopharyngeal. It should be noted that both area functions had small lip area. However the results suggest a possible explanation in favor of the velopalatal shape, which is based on the concept of focalization that was introduced by Badin et al. (1990). A focalization was defined as a convergence of two formants, and a permutation of the formant-to-cavity affiliation as the tongue moves across the point of convergence. Boë et al. (1994) have suggested that the focalization concept would be helpful for the prediction of vowel sytem. In particular it permitted to predict vowel systems featuring a single internal /y/ vowel. Badin et al. (1990) have also proposed that the focalization would correspond to formant patterns that are less sensitive to articulatory changes, and then require less articulatory accuracy.

Bentoumi has shown that only the velopalatal [u] is focal: F_1 being associated with the second Helmholtz resonance (constriction + back cavity) and F_2 with the first one (lips + front cavity). For the velopharyngeal area function proposed for [u], F_1 and F_2 are both associated with the front cavity. So the velopharyngeal [u] is much more sensitive to front/back variations of the tongue body than the prototypical [u] which is very stable for such modifications.

4. Conclusion

Lip-tube experiments with facilitating context have shown that the speakers' capability to compensate for the lip perturbation can be explained by the quality of the internal representation of the articulatory-to-acoustic relations. Limitation would thus not be induced by the existence of a learned standard articulatory pattern.

However some aspects of our observations still incite us to believe that such learned patterns exist and are part of the representation that speakers have of the speaking tasks. It is suggested that for [u] a velopalatal pattern would be favored because of focal properties that would provide more stability and less articulatory accuracy.

Acknowledgements

Special thanks to Christian Abry, Pierre Badin, Frank Guenther, Joe Perkell and Jean-Luc Schwartz for stimulating discussions on the case of [u].

Many results analyzed in the paper are due to Nouredine Bentoumi, Marc Blanc, Sébastien Cleyet-Merle, Martial Cocusse, and Didier Godat

References

Cocusse M. (1996). *Stratégie de production et de compensation pour un robot parlant : Le cas du [u] avec "Lip-Tube"*. Diplôme d'Études Approfondies (Sciences Cognitives). Institut National Polytechnique de Grenoble.

Badin P., Perrier P., Boë L.J. & Abry C. (1990). Vocalic nomograms: Acoustic and articulatory considerations upon formant convergences. *Journal of the Acoustical Society of America, 87 (3)*, 1290-1300.

Bentoumi N. (1996). *Inversion acoustico-articuatoire et anisotropie de l'espace acoustique*. Diplôme d'Études Approfondies (Sciences Cognitives). Institut National Polytechnique de Grenoble.

Boë L.J., Schwartz JL. & Vallée N. (1994). The prediction of vowel systems: Perceptual contrast and stability. In E. Keller (Ed.), *Fundamentals of Speech Synthesis and Speech Recognition* (pp. 185-213). Chichestter, UK: John Wiley and Son.

Fant, G. (1992). Vocal tract area function of Swedish vowels and a new three-parameter model. *Proceedings of ICSLP 92* (Vol. 1, pp. 807-810). Edmonton, University of Alberta.

Guenther F. (1995). "Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production". *Psychological Review*, Vol. 102(3), 594-621.

Guenther F., Hampson M. & Johnson D. (1997). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review* (In press).

Jordan M.I. & Rumelhart D.E. (1992). Forward Models: Supervised Learning with a Distal Teacher. Cognitive Science, 16, 316-354.

Kawato M., Furukawa K. & Suzuki R. (1987). A hierarchical neural network model for control and learning of voluntary movement. *Biol. Cybern.*, 57, 169-185.

Ladefoged P. & Maddieson I. (1996). *The Sounds of the World's Languages*. Cambridge, MA: Blackwell Publishers.

Perkell J.S. (1996). Properties of the tongue help to define vowels categories: hypotheses based on physiologically-oriented modeling, *Journal of Phonetics*, Vol. 24, pp. 3-22.

Savariaux C., Perrier P. & Orliaguet J.P.(1995a) Compensation Strategies for a Lip-tube Perturbation of the Rounded Vowel [u]. *Journal of the Acoustical Society of America, 98 (5)*, 2428–2442.

Savariaux C., Perrier P. & Schwartz J.L. (1995b) Perceptual analysis of compensatory strategies in the production of the French rounded [u] perturbed by a lip-tube. *Proceedings of the XIIIth International Congress of Phonetic Sciences., Vol. 3, 584-588*. Stockholm, Sweden.