PERCEPTION OF COPRODUCED SPEECH GESTURES

René Carré
ENST, Dept. TSI, Unité Associée au CNRS, 46 rue Barrault, 75634 Paris cedex 13

ABSTRACT

In this paper, coarticulatory variations were first obtained with a speech gesture coproduction model. Then, we investigated the perceptual effects of the variabilities using stimuli synthesized with an acoustic tube model deduced from acoustic theory, the DRM model. Perceptual tests showed that the ears tolerate such variations to a considerable degree.

1. INTRODUCTION

Two main coarticulation perspectives have been proposed (see the discussion by Fowler [9]): one is feature spreading and the second one is coproduction. Kozhevnikov and Chistovich [12] proposed syllabic coproduction where the sequence CV is produced as a unit (for coproduction, see also Fowler [8, 9]). Indeed, Öhman [14] observed that, in V_1CV_2 production, the C-V_2 transitions are influenced by the first vowel V_1 and, reciprocally, V_1-C transitions are affected by the final vowel V_2. To explain these observations, he proposed a coarticulation model in which the observed formant transitions are determined by superimposing local perturbations, caused by consonantal articulation, on baseline V_1-V_2 transitions. Consonant and vowel are coproduced. If gestures, i.e., linguistic actions on the vocal tract [1], are used to describe vowel and consonant transitions "gesture coproduction" arises. Adopting a gestural syllabic coproduction approach, our first task was to build a speech production model that inherently integrates coarticulation phenomena and provides synthesized tokens. Second, using these tokens, perceptual tests were performed to study the behavior of our modeled coarticulation.

2. MODELING OF SPEECH GESTURE COPRODUCTION

The preceding gestural syllabic coproduction hypothesis was adopted to build a speech production model. Moreover, if we consider that speech has to be heard, then speech gestures are actions on the area function with acoustic goals.

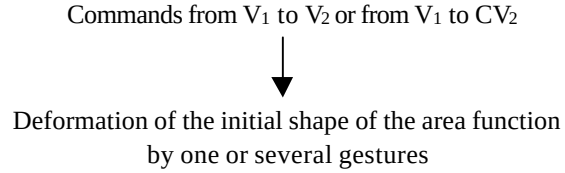


Figure 1. Gestural syllabic coproduction model

In figure 1, the deformation of the initial shape of the area function of the vocal tract is obtained via the articulatory tools by one or several gestures of the syllabic commands. It is hypothesized that the effects of the commands on the area function are not strictly in phase and of the same duration.

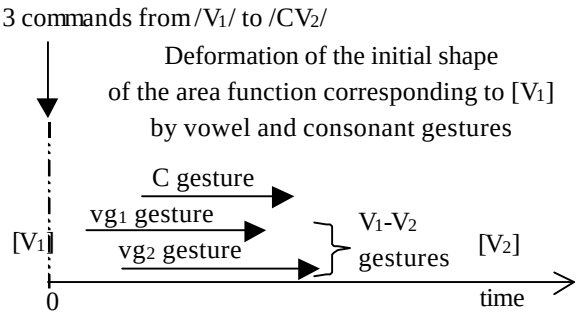


Figure 2. Gestural syllabic coproduction of /V_1-CV_2/

In the example shown in figure 2, there are two vowel gestures and one consonant gesture. According to our hypothesis, the same set of initial abstract commands planned (starting at time 0) will evoke gestures the effect of which on the area function can have a variety of different manifestations in the time domain, such as different temporal: durations, movement patterns and intergestural asynchrony. Such differences can be attributed either to specific articulatory constraints or speaker variabilities, or both. Indeed, the existence of any abstract representation means that the corresponding action must be detectable by way of invariant characteristics that are inherently variable.

There is a time delay between the syllabic command (at time 0) and the corresponding actions. At the action stage, the beginning of the syllable is at the beginning of the vg_1 gesture, as shown in figure 2. But at the signal stage, if the beginning of the syllable is considered to coincide with the onset of the consonantal closure, then there is an anticipation of V_2 because the acoustic effect of the gesture vg_1 is observable in the preceding syllable (in V_1). In our model, at the planning stage, the beginning of the syllable includes what is generally referred to as anticipation.

Assuming that the objective of speech gesture is to generate some acoustic perturbation, it is reasonable to suppose that the tools used for the deformation of the area function (i.e., the articulators) should be well adapted for the task. Here, for the sake of simplicity, we want to focus our attention to deformation gestures which are well adapted for generating specific acoustic perturbations, without taking into account articulatory constraints.

The deformation gestures of our interest are automatically obtained using a deductive approach. We apply two criteria to the deformation of an acoustic tube of 17.5 cm length: 1) maximum acoustic contrast, and 2) efficiency (or minimum effort: a

small gesture deformation should lead to a large acoustic effect). In fact, these minimal constraints have formed the point of departure for the derivation of the distinctive region model (DRM) [13], [6]. In this model, an asymmetrical behavior is observed: whereas a front constriction is automatically associated with a back cavity and vice-versa, a central constriction will be automatically associated with two lateral cavities. Acoustically efficient places of articulation also automatically follow from the model [7] (Figure 4): Although these model places are obtained without any knowledge about articulatory observations, they nevertheless coincide with the places used to produce vowels and consonants. Gesture deformations acting on back, front or central constriction (and front end) are used to produce vowels; whereas closing/opening one of the three front regions (as shown in figure 4) is used to produce labial, palatal, or velar consonants.

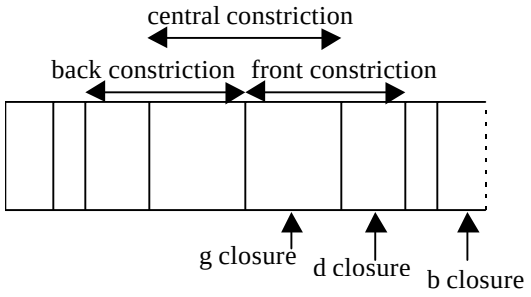


Figure 4. DRM with places of articulation

With the help of a similar model, we reproduced the precise V_1CV_2 formant patterns measured by Öhman [3] by superimposing a consonant gesture on a V_1V_2 transition.

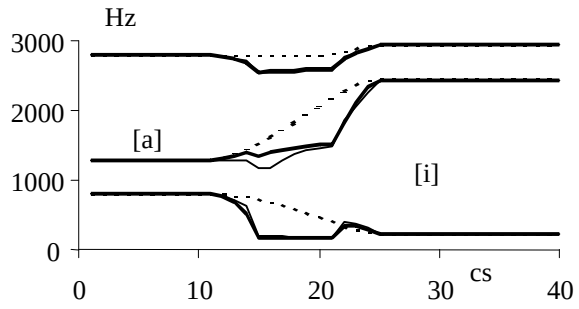


Figure 5. Formant transitions for [ai] (dotted line). For [abi] (thick line), the consonant gesture is superimposed in phase with the vowel gesture. For [abi] (solid line) the beginning of the vowel gesture correspond to the beginning of the full consonant closure.

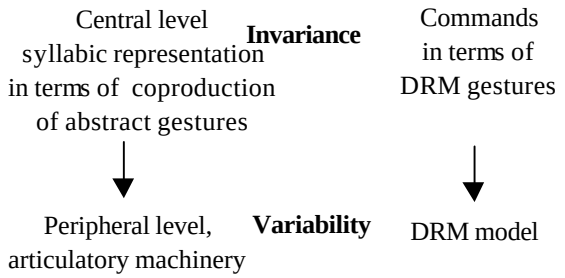


Figure 6. Direct link between phonological representation and phonetic level

Figure 5 shows an example of an [abi] production. The formant transitions for [ai] are shown in dotted line. If the vowel and consonant gestures are in synchrony, the token [abi] marked with the thick line is obtained. The C-V_2 transitions are influenced by the first vowel V_1 and, reciprocally, the V_1-C transitions are affected by the final vowel V_2 [14]. When the onset of the vowel gesture toward V_2 is synchronized with the onset of the full consonant closure, as observed by Gay [11], the transition V_1C in [abi] is not influenced by V_2 (solid line).

With this simple model, phonological representation and phonetic level are directly linked (figure 6) without any translation required [10]. Moreover, invariant gestural phonetic commands lead to variable corresponding area function actions.

3. PERCEPTION TESTS

Our speech production model was used for the synthesis of tokens with different gestural characteristics. With these tokens, we then examined the perceptual effect of variations on the characteristics of gestural actions for two cases: 1) [ai] transitions represented by only one gesture for different action duration, and 2) [ay] and [abi] transitions represented by two gestures with different action asynchronies (for more details on the experiments see [2, 4, 5]).

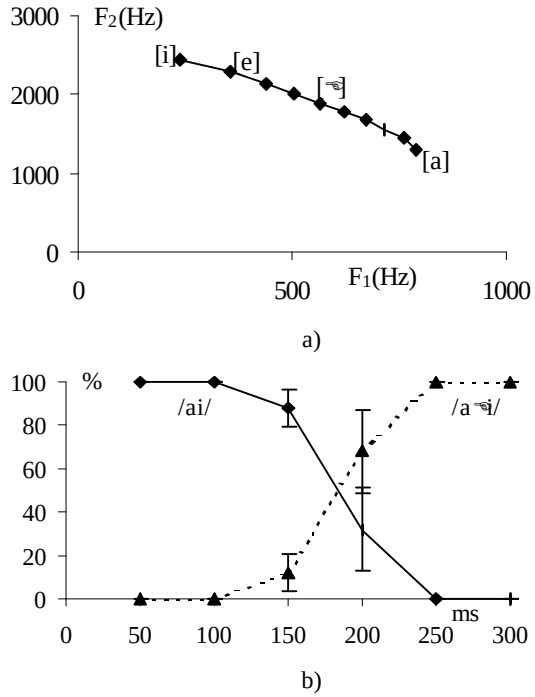


Figure 7. a) [ai] formant trajectory, b) percept (in % with standard deviation) of [ai] tokens as a function of gestural duration (from 50 to 300 ms by steps of 50 ms)

3.1. Perceptual effects of gesture duration

When the model produces an [ai] token, the formant trajectory crosses the regions of the French vowels [ɛ] and [e] (see figure 7a), although the presence of these vowels is not heard in normal speech. In an experiment, we examined the percept generated by [ai], with gesture duration ranging from 50 to 300 ms in 50 ms steps. The duration of the first vowel was 100 ms and that of the second vowel 150 ms. Figure 7b illustrates the results of this experiment, with French listeners as subjects. The vowel complex /ai/ was perceived with gesture durations between 50 but around 180 ms, then an /aɛi/ percept was reported. That is, at these longer durations an additional interme-

diate vowel, /ə/, was heard, despite the absence of a segmental marker. The 180-ms duration is in the syllabic range and, if phonetic memory had such a span, it could account for the results. A vowel encountered when the transition reaches the end of the time window that the 180 ms phonetic memory span represents could automatically trigger a label even without the presence of a segmental marker. Thus, our syllabic coproduction model seems to be well-adapted for the French language, which is classified as syllabic.

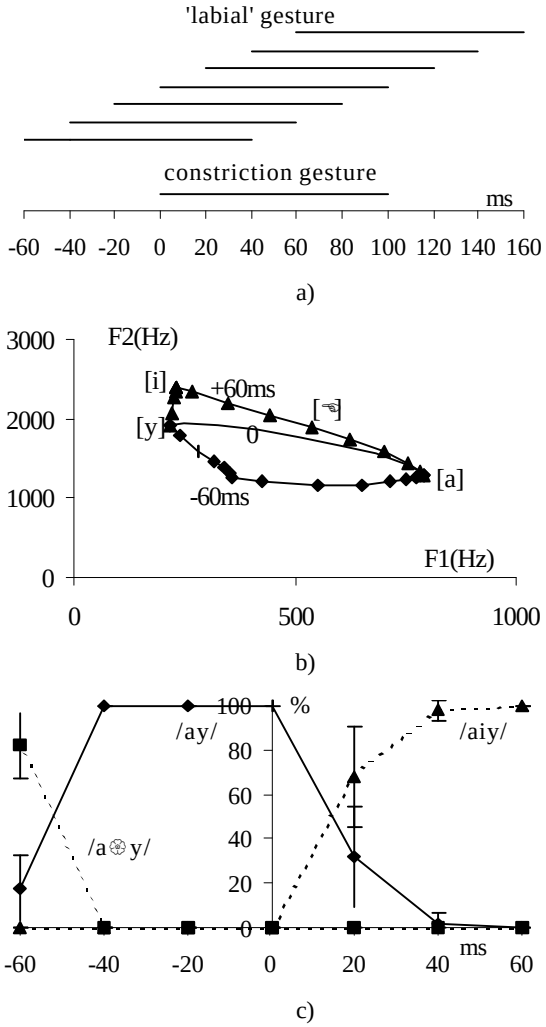


Figure 8. a) Constriction and labial gesture asynchrony; b) Corresponding formant trajectories for the three cases of labial lags of –60, 0, and +60 ms; c) Percent [ay] percept as a function of gestural asynchrony (from -60 to +60 ms in steps of 20 ms)

3.2. Perceptual effects of gesture asynchrony

To study the perceptual effect of gesture asynchrony, tokens with transitions generated by at least two gestures are required. In French, the [ay] transition is obtained by the coproduction of constriction gesture and labial gesture, the [ɶ͡ʊ] transition by the production of a single gesture (the constriction gesture because [ɶ] and [y] are both labial vowels), [ɔ͡ɸ͡ʜ] by two gestures (the constriction and the b-consonant gesture), [ɔ͡ɸ͡ʊ] by three gestures (the constriction, the labial gesture, and the b–consonant gesture), and [ɶ͡by] by two gestures (the constriction and the b-gesture, because [ɶ] and [y] are both labial vowels). In the following section, we describe two experiments that examined the perceptual effect of asynchrony of two gestures.

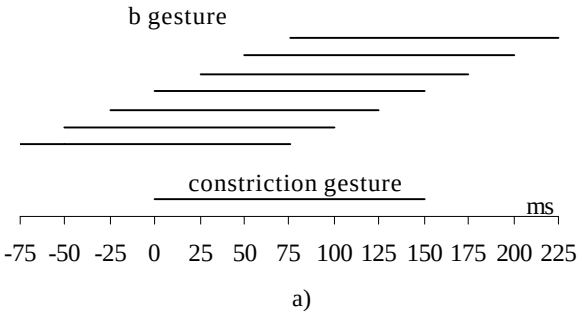
3.2.1. [ay] tokens. As indicated above, [ay] is a coproduction of two gestures: the constriction and labialization. In the following experiment, the duration of the two gestures were both fixed at 100 ms and the asynchrony of their relative onsets (i.e., relative phasing) varied from –60 to +60 ms in steps of 20 ms, where a negative asynchrony refers to anticipated labialization (figure 8a). The time domain function of both gestures was cosine.

Figure 8b shows the formant trajectories for labial gesture onset lags of –60, 0, and +60 ms with respect to the onset of the constriction gesture. In the –60 ms case, it is the labial gesture that is activated first (leading to a decrease of F_1, with F_2 remaining approximately stable), then the constriction gesture becomes dominant, in order to reach [y]. In the +60 ms case, it is the constriction gesture that is activated first but it is aimed at [i] before turning toward [y] due to the labial gesture delay. If the two gestures are precisely synchronous, the 0 ms case, the formant trajectory is first aimed at [i] before swerving to aim at [y]. This means that, in fact, when the onset of the two gestures is simultaneous, the acoustic effect of the labial gesture is delayed with respect to that of the constriction gesture.

Figure 8c illustrates the perceptual results and shows that an /ɔ͡ʊ/ percept is reported for labial lags between –50 and +10 ms. Because the acoustic trace onset of the constriction gesture occurs before that of the labial gesture, the identification function is not centered. When the asynchrony is less than –50 ms, the percept is /aʏy/ and when it is more than +10 ms, it is /ɔ͡ʜʊ/.

3.2.2. [abi] tokens. A similar experiment of coproduction with two gestures was designed to examine the [abi] production case. This consonant was chosen to avoid lingual coarticulation between vowel and consonant gestures. The duration of V_1 was fixed at 100 ms and that of V_2 at 150 ms. The area function during the V_1-V_2 transition was determined by a cosine interpolation of the two vowel targets. The vowel transition and the consonant gesture both had an identical 150 ms duration. The consonant gesture was simulated by a closure occurring during the transition between the two vowels. The consonantal closing, full closure, and opening durations were 50 ms each. The V_1CV_2 signals were produced without a burst. The lag between the two gestures varied from –75 to +75 ms in 25 ms steps, where the 0 value corresponds to a precise synchrony of the two gestures (fig. 9a). At some asynchrony values an acoustic segment marker (a transient) at the [ə] boundary can be observed (fig. 9b).

The effect of gesture asynchrony on perceptual boundary is shown in figure 9c. The percept /abi/ was reported for lags between approximately –50 and +35 ms. There is also a wide range of lags at which the percept was constant. The listeners reported hearing /abei/ for lags under –50 ms and /aəbi/ for lags above +35 ms.



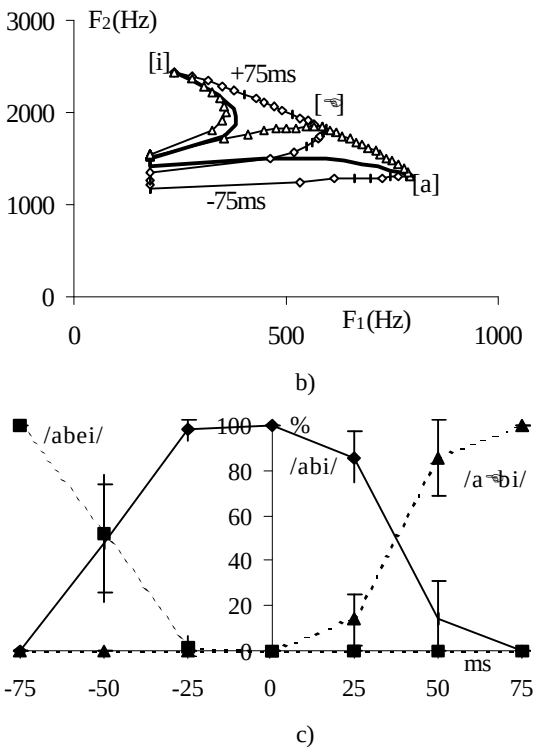


Figure 9. a) Constriction and consonant gesture asynchrony; b) Corresponding formant trajectories in case of asynchrony equal to –75, 0, +75 ms; c) Percept (in % with standard deviation) of [abi] tokens as a function of gestural asynchrony (from -75 to +75 ms in steps of 25 ms)

4. CONCLUSIONS

Our study indicates that, in VV and VCV utterances, a surprisingly high degree of perceptual invariance can be achieved despite relatively large variations of gesture characteristics, such as gesture asynchrony, duration, and movement trajectory (as reported in [4]). While this invariance is consistent with the view that the perceptual representation of these utterances is gestural, we recognize that it provides only a necessary but not a sufficient proof for a full-fledged gesture perception theory. We note, nevertheless, that a large range of variability may be regarded as one necessary condition for speech communication between two speakers and for infant phonetic repertory learning. It may also be indispensable for the process of phonological sound changes as, at precise gesture asynchrony conditions, new acoustic segment markers are likely to arise and generate new phonetic categories. The tokens used in our perception experiments were synthesized by means of a deductive, area function model performing deformations that produce acoustic effects much in a manner the human vocal tract does. Since the strategies for using articulatory tools can vary with a high degree of freedom, the mismatch between linguistic competence and vocal tract performance is minimized. Then, if any carryover coarticulation is detected, it can be assigned to speaker-specific strategies rather than to limitations of the physical characteristics of the articulatory system.

But is strict coproduction of two gestures possible when, theoretically at least, the two correspond to different and conflicting tasks, each directed at modifying the area function in its own idiosyncratic way, and each having a different acoustic targets? This question cannot be answered without further study of what is generally called "lingual coarticulation," not in solely articulatory terms but, above all, in terms of an acoustic purpose.

ACKNOWLEDGMENTS

Thanks to Pierre Divenyi for many stimulating discussions.

REFERENCES

[1] Browman, C. and Goldstein, L. 1986. Towards an articulatory phonology. In C. Ewan and J. Anderson (ed.), *Phonology yearbook*. Cambridge, Cambridge University Press.

[2] Carré, R., Ainsworth, W.A., Jospa, P., Maeda, S. and Pasdeloup, V. Submitted. Perception of vowel-to-vowel transitions in terms of "gestures" as defined by the DRM model.

[3] Carré, R. and Chennoukh, S. 1995. Vowel-consonant-vowel modeling by superposition of consonant closure on vowel-to-vowel gesture. *J. of Phonetics*, 23, 231-241.

[4] Carré, R., Chennoukh, S., Divenyi, P. and Lindblom, B. 1994. On perceptual characteristics of "speech gestures". *J. Acoust. Soc. Am.*, 96, S3326.

[5] Carré, R., Chennoukh, S., Jospa, P. and Maeda, S. 1996. The ears are not sensitive to certain coarticulatory variations: Results from VCV synthesis/perceptual experiments. *Proc. of the First ESCA T.R.W. on Speech Production and Modeling*. Autrans.

[6] Carré, R., Lindblom, B. and MacNeilage, P. 1994. Acoustic contrast and the origin of the human vowel space. *J. Acoust. Soc. Am.*, 95, S2924.

[7] Carré, R. and Mody, M. 1997. Prediction of Vowel and Consonant Place of Articulation. *Proc. of the Third Meeting of the ACL Special Interest Group in Computational Phonology, SIGPHON 97*. Madrid.

[8] Fowler, C. 1977. *Timing control in speech production.* Bloomington, Indiana University Linguistics Club.

[9] Fowler, C.A. 1992. Phonological and articulatory characteristics of spoken language. *Haskins Laboratory Status Report on Speech Research*, SR-109/110, 1-12.

[10] Fowler, C.A., Rubin, P., Remez, R. and Turvey, M.T. 1980. Implications for speech production of the general theory of action. In B. Butterworth (ed.), *Speech Production I: Speech and talk*. London, Academic Press.

[11] Gay, T. 1977. Articulatory movements in VCV sequences. *J. Acoust. Soc. Am.*, 62, 183-193.

[12] Kozhevnikov, V.A. and Chistovich, L.A. (1965) *Speech, articulation, and perception*. NTIS, US Dept. of Commerce.

[13] Mrayati, M., Carré, R. and Guérin, B. 1988. Distinctive region and modes: A new theory of speech production. *Speech Communication*, 7, 257-286.

[14] Öhman, S. 1966. Coarticulation in VCV utterances: spectrographic measurements. *J. Acoust. Soc. Am.*, 39, 151-168.