Anticipatory Coarticulation in Vowel-Consonant-Vowel sequences: A crosslinguistic study of French and Mandarin speakers

Liang Ma[1, 2], Pascal Perrier[2], Jianwu Dang[3]

Laboratoire Parole et Langage, UMR CNRS 6057, Univ. de Provence, Aix-en-Provence, France

Institut de la Communication Parlée, UMR CNRS 5009, INPG & Univ. Stendhal, Grenoble, France

Japanese Advanced Institute of Sciences and Technology, Ishikawa, Japan,
liang.ma@lpl.univ-aix.fr, perrier@icp.inpg.fr, jdang@jaist.ac.jp

Abstract. *Anticipatory coarticulation within V1CV2 sequences is studied for two different languages, French and Mandarin. EMMA data and acoustic signals were collected for 3 speakers of each language. The corpus was consistent with one another in both languages, V1 and V2 being one of the set /i,a,u/ and C being either /t/ or /k/. The influences of V2 on V1 and of V2 on C were more specifically analyzed in this paper. Our results suggest that anticipatory coarticulation takes into account the whole sequence V1CV2 for the speakers of French, while it is strictly limited to the syllable CV2 for the speakers of Mandarin.*

1. Introduction

It is now commonly accepted that the control of speech production sequences involves a planning process in the central nervous system, which uses internal representations (Jordan, 1990; Kawato *et al.*, 1990) of the speech production apparatus (Guenther, 1995; Guenther et al., 1998; Perkell et al., 2000; Perrier et al., 2005), in order to optimally achieve goals in an acoustic, perceptual and/or articulatory domain. Beyond this general agreement, crucial questions are still much debated, among which we can found the level of complexity of the internal representations (see Gomi & Kawato, 1996; Gribble et al., 1998; Perrier, 2006), the nature of the criterion to be optimized (distance, effort, jerk, force…, Nelson, 1983), the relative weights of perceptual and motor control constraints in the optimization process (Lindblom, 1990; Nguyen & Fagya, 2006), and the number and type of subsequent phonemes for which the criterion is optimized (CV, VC, VCV, words…).

This study is part of a larger cooperation project that aims at implementing and assessing models of speech production control. ICP and JAIST have developed in parallel models of speech production including physical models and motor control models (Dang & Honda, 2004; Dang et al., 2005; Perrier et al., 2003; Perrier et al., 2005). The general aim of the project is to compare the articulatory and acoustic speech signals that can be generated using these models with data collected from human speakers in different languages, in order to see the strengths and the limits of the models and to improve them. Based on studies carried out with a biomechanical tongue model, we found that anticipatory behaviour observed on articulatory movements seems not to be the result of physical influences, such as articulatory dynamics (Perrier et al., 2004, but see also Ostry *et al.*, 1996). Accordingly, coarticulation can be supposed to correspond to a fair image of the high level motor control strategies. Therefore, our focus is first put on anticipatory coarticulation. To start with, we collected consistent kinds of utterances of French and Mandarin and analyze anticipatory coarticulation using the VCV utterances in these data.

2. Speech material

2. 1. Corpus

Speech material consists of 15 VCV nonsense words where the vowel was /a/, /i/ or /u/ and the consonant was /k/ or /t/. The words were uttered at a normal speech rate by three native speakers of French and three native speakers of Mandarin. Each target word was embedded in a carrier sentence: "C'est VCV ça?" in French and "这是 VCV 吗?" in Mandarin. Each carrier sentence was repeated 10 times, except for the Chinese subject JW, who only produced 4 repetitions.

The corpus was elaborated, in order to have consistent coarticulation environment in both languages, in spite of well-known differences in their respective phonemes inventory and in their respective linguistic structures. In particular, in Mandarin we avoided sequences that could be ambiguous due to uncertainty about the tonal structure. We also did not consider sequences such as [V1ki] that do not exist in Mandarin.

Acoustic and articulatory data were recorded simultaneously. The articulatory data were collected with an electromagnetic midsagittal articulograph (EMMA; AG100 Carstens Electronics). Four sensors were placed on the tongue from around 1cm to 5 cm from the tongue tip. One sensor was also glued on the upper lip, on the lower lip, and on the lower incisor. Reference sensors were located on the upper incisor and one on the bridge of the nose. All the sensors were carefully located in the midsagittal plane, in order to ensure the best measurement accuracy. The sensors glued on the tongue are called T1, T2, T3 and T4, from tongue tip to the tongue back. The articulatory signals were sampled at 200Hz for two Chinese subjects (SK and JW) and at 500HZ for the three French subjects (AV, PB, CV) and the third Chinese subject WS. Articulatory data were smoothed by a 20 Hz bandwidth low-pass filtering, before they were analyzed.

2. 2. Labelling

The aim of labelling was to detect both acoustic and kinematic events. Acoustic events were related to formant patterns for vowels and to bursts for stops. Kinematic events were related with velocity maxima and zero crossings in order to get information about articulatory positioning and articulatory movements. In this paper, only the results about articulatory positioning are presented.

The labelling was carried out manually. For the consonants, the onset of the burst was measured to characterize the most canonical articulatory positioning. For the vowels, the point with maximum stability of the first three formants on the spectrogram was labelled in a first step. Then, in a second step, in order to achieve a more accurate detection of the most canonical vocal tract configuration, the label was automatically moved towards the extreme position of the tongue back sensor T4 in the vicinity of the position detected on the spectrogram. The extreme position of T4 was defined for vowel [a] as the lowest point, for vowel [i] as the most anterior point, and for vowel [u] as the most posterior point. T4 was taken into account here, because we considered that it is the best index about the global back/front and low/high positioning of the tongue. Figure 1 shows an example of this labelling. It depicts the trajectory of the four tongue sensors for the first vowel [a] in the [aka] sequence within a 100 ms interval around the first label extracted from the spectrogram (subject AV). The palate shape is plotted on top of the figure. Front is on the left, back is on the right. The circles represent the tongue position at the time specified by this first label. The star points represent the tongue shape at the time specified by the second label. It can be seen that a difference as big as 2mm exists between the sensor positions observed for these two labels.

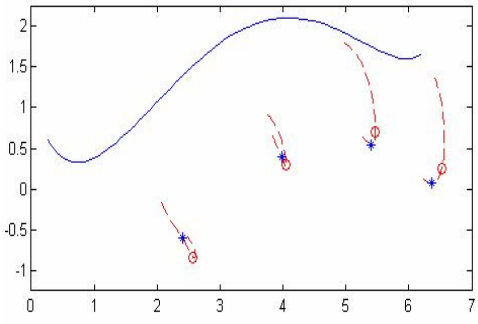

Figure 1: Articulatory labelling (see text for details)

2. 3. Data analysis

As mentioned above, data were analyzed in the aim to study anticipatory strategies in articulatory positioning. To do so, for each sequence Vowel1-Consonnant-Vowel2 (V1CV2 henceforth) and for each subject, we statistically computed the influence of V2 on the articulation of the preceding phonemes.
The tongue position for the vowel was characterized with the three sensors T2, T3 and T4 since the tongue tip is less constrained in vowel production. For the consonant, the 4 tongue sensors T1, T2, T3 and T4 were taken into account when they were available.

In this aim, a variance analysis ANOVA (Repeated Measures) was carried out for vowel V1 and consonant C separately. The independent variables were the horizontal and

vertical positions of the sensors for V1 and for C, and the independent factor was V2. SPSS™ for Windows was used for this analysis.

It is important to mention here that the accuracy of our EMMA system was estimated to be around 0.5mm. For this reason, we considered differences in tongue positioning to be significant only if they were statistically significant at $p<0.05$ and if they were larger than 0.5mm.

3. Results

Because of hardware problems that happened during the experiment, a number of data is missing for some subjects. As a result, only the data of sensor T1 and T4 were available for subject WS, while sensor T3 was missing for subject AV; sensor T1 was not recorded for subject CV and SK.

3. 1. Effects of V2 on V1

The careful observation of the dispersion ellipses that were calculated, for each vowel V1 separately, from the measurements of all repetitions of [V1tV2] and [V1kV2] respectively suggests two general conclusions. First, the token-to-token variability is very much speaker dependent (small for speakers PB and JW, and large for speakers CV and SK), but similar variability distributions exist for both languages. This observation suggests that for vowels /i/, /a/ and /u/ the impact of motor and perceptual constraints on token-to-token variability is similar for both languages. Second, the variability of the average tongue position associated with change in V2 is larger for the native speakers of French (left column) than for the native speakers of Mandarin (right column).

Results of ANOVA confirm the second observation as will be shown below.

a. [V1tV2] sequences,

For three possible vowels V1, speaker AV has a significantly higher T2 position if V2= /i/ than if V2=/a/ or V2=/u/ *(V1=/a/, average distance(i-a)=2.4mm and average distance(i-u)=1.6mm; V1=/i/, average distance(i-a)=0.6mm and average distance(i-u)=0.6mm; V1=/u/, average distance(i-a)=1.3mm and average distance(i-u)=1.5mm).* For sensor T4 some significant differences are also observed in some cases, but not systematically and they are not consistent across the three possible vowels V1.

Speaker PB shows significant differences in T2 or T3 vertical positions, which are not all observed for the three possible vowels V1, but which all go in the same direction. For V1=/a/, T2 is lower if V2=/a/ than if V2=/i/ or V2=/u/ *(average distance (i-a) =1.6mm and average distance (u-a) =1.9mm).* For V1=/i/, T3 is higher if V2= /u/ than if V2=/a/ *(average distance (u-a) =0.8mm).* And For V1=/u/, T2 is higher if V2=/i/ than if V2= /u/ *(average distance (i-u) =1.8mm),* while T3 is higher if V2=/i/ than if V2=/a/ *(average distance (i-a) =0.7mm).*

Speaker CV shows significant differences in T2 and in T3 vertical positions in case of V1=/a/ or /u/. These sensors are higher if V2=/i/ than if V2=/a/ or V2=/u/ *(For T2: V1=/a/, average distance(i-a)=2.1mm and average distance(i-u)=1.7mm; V1=/u/, average distance(i-a)=3.3mm*

and average distance(i-u)=3.4mm) (For T3: V1=/a/, average distance(i-a)=2.6mm and average distance(i-u)=2.0mm; V1=/u/, average distance(i-a)=2.6mm and average distance(i-u)=3mm). Other differences were also observed for sensor T4, or in the x-direction, but not systematically and not consistently across the three possible vowels V1. For V1=/i/, no significant differences were observed.

For speakers SK and WS, no significant difference was observed, that was consistent across the three possible vowels V1.

For speaker JW, we did not observe any significant difference that was shared by two different vowels V1. For V1=/i/, there is no significant difference. For V1=/a/ differences exist in the horizontal position of T2, T3 and T4. These sensors are further back if V2=/i/ than if V2=/a/ or V2=/u/ *(For T2: average distance(i-a)=3.2mm and average distance(i-u)=2.3mm) (For T3: average distance(i-a)=2.4mm and average distance(i-u)=2mm) (For T4: average distance(i-a)=1.7mm and average distance(i-u)=1.5mm).* For V1=/u/, T2 and T3 are significantly higher if V2=/i/ than if V2=/a/ or V2=/u/ *(For T2: average distance(i-a)=2.6mm and average distance(i-u)=1.8mm) (For T3: average distance(i-a)=2.1mm and average distance(i-u)=1.7mm).*

b. [V1kV2] sequences

For speaker AV no significant difference was observed for V1=/a/ and V1=/i/. For V1=/u/, T2 and T4 are more anterior if V2=/u/ *(average distance for T2=2.2mm; average distance for T4=2.7mm)*, and T2 is higher if V2=/a/ *(average distance=2.4mm).*

For PB, the only significant difference that was observed consistently across at least two vowels V1, is the fact that for V1=/a/ and for V1=/i/ T4 is higher if V2=/u/ *(for V1=/a/, average distance=0.9mm; for V1=/i/, average distance=0.8mm).*

For CV, the only significant difference that was observed consistently across at least two vowels V1, is the fact that for V1=/a/ and for V1=/u/ T2 is higher if V2=/u/ *(for V1=/a/, average distance=1.5mm; for V1=/u/, average distance=1.3mm).*

For speakers SK and WS, no significant difference was observed.

For speaker JW, we observed that for V1=/a/ and for V1=/i/, T3 is significantly higher if V2=/u/ *(For V1=/a/, average distance=1.2mm; for V1=/i/, average distance=0.7mm)* and that for V1=/i/ and for V1=/u/, T4 is significantly higher if V2=/u/ *(For V1=/i/, average distance=1.2mm; for V1=/u/, average distance=0.8mm).*

3. 2. Effects of V2 on C

For the six analyzed speakers, we observed significant differences in tongue positioning for consonant /t/ when V2 changes. The amount of difference associated with V2 changes varies with vowel V1 but not the direction. For AV, PB, and CV, the main and most consistent difference is in the vertical position of T2 and/or T3. This position is higher if V2=/i/, than if V2= /u/, while the lowest position is reached if V2=/a/. Differences are large and their maximal ranges vary, according to vowel V1, between 5 and 10mm for speaker AV, between 6 and 7mm for speaker PB, and between 6 and 8mm for speaker AV. Noticeable differences are also observed in the vertical position of T4, with a trend for it to be higher if V2=/u/, but this is less consistent across speakers and conditions than difference in T2 and T3 positions. For all Mandarin

speakers, differences are related to T4 vertical position. It is systematically higher if V2=/i/ than if V2=/u/ and then if V2=/a/. Here also the range of variation is large: between 5 and 11mm for SK, around 7mm for WS and around 6mm for JW.

For consonant /k/, significant differences were observed in all cases for the three French speakers, but nothing was consistently observed across conditions for any Mandarin speakers. For all French speakers, noticeable differences exist in both in horizontal and vertical positions for T2, T3 and T4 and for the three vowels. Tongue position of /k/ is further back and lower if V2=/u/ than if V2=/a/. For AV the range of variation is between 1.3 and 4.5mm; for PB it is between 1.3 and 3.4mm and for CV 0.5 and 2.5mm.

4. Discussion and conclusion

The native speakers of French often show a significant articulatory variability for V1, when V2 varies. The amount of variability is depending on V1. Indeed, vowel /i/ was found to be in general less sensible to the variation in V2. This is in agreement with the classical view that /i/ is more constrained in the articulatory domain than the other vowels. The main trend of the measured variability suggests the hypothesis that in French the articulation of V1 within V1CV2 sequences anticipates the articulation of V2. Indeed, in the majority of cases V2=/i/ influences specifically the positions of T2 and T3 of V1, which are located in the constriction region of /i/. Therefore, higher and/or a more anterior positions of T2 and T3 are consistent with the fact that the forthcoming articulation of /i/. Similarly in some cases V2=/a/ differs from V2=/u/ and V2=/i/ because V1 shows lower position for T2 and T3. Here again it is consistent with the fact that the tongue for /a/ is flat and low in its anterior part.

For the speakers of Mandarin, the only case of significant V1 variability due to V2 variation was observed for JW in [atV2] sequences. The tongue was higher and further back when V2=/i/. This is contradictory with the fact that /i/ is articulated more front than the other two vowels. Therefore, it cannot be considered to be a direct anticipation of the next vowel. We assume that it is the consequence of the influence of the articulatory configuration associated with the consonant /t/, when it is pronounced before vowel /i/ by this speaker. Indeed, in this case, the tongue of speaker JW for /t/ is located further back and the tongue dorsum is higher and much closer to the palate, than before vowel/a/ or vowel /u/. Possible explanations for this phenomenon could be found either at a phonological level (in JW's Mandarin, /ti/ could be more affricate than /tu/ or /ta/) or at the level of the vocal tract morphology (JW has namely a palate sensibly more arched in the sagittal plane than SK or WS). However, further investigations are necessary to assess these hypotheses. In any case, it seems reasonable to consider that V1 variability associated with V2 changes is not the direct result of an anticipatory strategy toward the coming articulation of /i/.

Our results suggest that an anticipation of the articulation of V2 exists during C for both groups of speakers. Indeed, the observed variation of C associated with the variation of V2 is compatible with the articulatory characteristics of V2. The absence of anticipation for /k/ in Mandarin does not discard this conclusion, since clear evidence for anticipation exist for /t/. It rather suggests that the amount of acceptable anticipatory variability is determined by perceptual requirements, and that these requirements vary

across consonants and across languages (see for example Manuel, 1990, for a similar hypothesis related to V1-V2 coarticulatory variability).

In conclusion, our data suggest for French that the planning of V1CV2 sequences takes into account the whole sequence. This finding is compatible with models of coarticulation like Öhman's model, the MEM model (Abry & Lallouache, 1996) or optimal models of planning such as those proposed by Jordan (1990), Kawato et al (1990), Perkell et al. (2000) or Perrier et al. (2005), which all take into account sequences longer than the syllable. Further work using different models of control applied to a biomechanical model of the tongue will aim at testing these different hypotheses. For Mandarin, it seems that the planning is limited to the syllable CV, which would be in agreement with Kozhevnikov & Chistovitch (1965) about the major role of the syllable.

Acknowledgments:

We express our gratitude to Noël Nguyen for his advices and comments. This work is partly supported by the "Programme de Recherche en Réseaux Franco-Allemand" ("POPAART" project) funded by the CNRS and the French Foreign Office.

5. References

Abry C. & Lallouache T.M. (1996). Le MEN: un modèle d'anticipation paramétrable par le locuteur. Données sur l'arrondissement du Français. *Bulletin de la Communication Parlée*, 3, Grenoble, France: Institut de la Communication Parlée.

Dang, J. and Honda, K. (2004). Construction and control of a physiological articulatory model, *Journal of Acoustical Society of America, , 115(2)*, 853-870

Dang, J., Wei, J., Suzuki, T. and Perrier, P. (2005). Investigation and Modeling of Coarticulation during Speech. *Proceedings of Interspeech 2005* (pp. 1025-1028). , International Speech Communication Association..

Guenther, F. H. (1995). Speech sound acquisition, coarticulation and rate effects in a neural network model of speech production. *Psychological Review*, *102*, 594–62.

Guenther, F. H., Hampson, M. & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, *105*, 611–633.

Gomi, H. & Kawato, M. (1996). Equilibrium-point control hypothesis examined by measured arm stiffness during multijoint movement. *Science*, *272*,117-120.

Gribble, P.L., Ostry, D.J., Sanguineti, V. & Laboissière, R. (1998). Are complex control signals required for human arm movement? *Journal of Neurophysiology*, *79*, 1409-1424.

Jordan, M.I. (1990). Motor Learning and the Degrees of Freedom Problem. In M. Jeannerod (Ed.), *Attention and Performance* (pp. 796-836). Hillsdale, NJ: Erlbaum.

Kawato, M., Maeda, Y., Uno, Y. & Suzuki, R. (1990). Trajectory formation of arm movement by cascade neural network model based on minimum torque-change criterion. *Biological Cybernetics*, *62*, 275-288.

Kozhevnikov, B. & Chistovitch, L. (1965). Speech: Articulation and Perception (Washington DC: Joint Publication Research Service), 104-118

Lindblom B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In W.J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling*, 403-439. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Manuel S. (1990), The role of contrast in limiting vowel-to-vowel coarticulation in different languages," *Journal of the Acoustical Society of America* 88, 1286-1298.

Nelson W.L. (1983). Physical principles for economies of skilled movements. *Biological Cybernetics*, *46*, 135-147.

Nguyen, N. & Fagyal, Z. (2006). Acoustic aspects of vowel harmony in French, *Journal of Phonetics* (In Press)

Ostry D.J., Gribble P.L. and Gracco V.L. (1996). Coarticulation of jaw movements in speech production: is context sensitivity in speech kinematics centrally planned? *Journal of Neuroscience*, 16: 1570-1579.

Perkell, J.S., Guenther, F.H., Lane, H., Matthies, M.L., Perrier, P., Vick, J., Wilhelms-Tricarico, R. & Zandipour, M. (2000). A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. *Journal of Phonetics*, *28*, 233-272.

Perrier P., Payan Y., Zandipour M. & Perkell J. (2003) Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study. *Journal of the Acoustical Society of America*, *114(3),* 1582-1599.

Perrier, P., Payan, Y., & Marret, R. (2004). Modéliser le physique pour comprendre le contrôle: le cas de l'anticipation en production de parole. In R. Sock & B. Vaxelaire (Eds.), *L'anticipation à l'horizon du Présent* (pp. 159-177). Pierre Margala Editeur, Sprimont, Belgique.

Perrier, P., Ma, L. & Payan, Y. (2005). Modeling the production of VCV sequences via the inversion of a biomechanical model of the tongue. *Proceedings of Interspeech 2005* (pp. 1041 – 1044), International Speech Communication Association.

Perrier, P. (2006). About speech motor control complexity. In J. Harrington & M. Tabain (eds), *Speech Production: Models, Phonetic Processes, and Techniques* (pp. 13-26). Psychology Press: New-York, USA