INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Does auditory-motor learning of speech transfer from the CV syllable to the CVCV word?

Tiphaine Caudrelier [1,2], *Pascal Perrier* [1,2], *Jean-Luc Schwartz* [1,2] *Amélie Rochet-Capellan* [1,2]

1 Univ. Grenoble Alpes, Gipsa-lab, F-38000 Grenoble, France
2 CNRS, Gipsa-lab, F-38000 Grenoble, France

tiphaine.caudrelier@gipsa-lab.grenoble-inp.fr,
amelie.rochet-capellan@gipsa-lab.grenoble-inp.fr

Abstract

Speech is often described as a sequence of units associating linguistic, sensory and motor representations. Is the connection between these representations preferentially maintained at a specific level in terms of a linguistic unit? In the present study, we contrasted the possibility of a link at the level of the syllable (CV) and the word (CVCV). We modified the production of the syllable /be/ in French speakers using an auditory-motor adaptation paradigm that consists of altering the speakers' auditory feedback. After stopping the perturbation, we studied to what extent this modification would transfer to the production of the disyllabic word /bebe/ and compared it to the after-effect on /be/.
The results show that changes in /be/ transfer partially to /bebe/. The partial influence of the somatosensory and motor representations associated with the syllable on the production of the disyllabic word suggests that both units may contribute to the specification of the motor goals in speech sequences. In addition, the transfer occurs to a larger extent in the first syllable of /bebe/ than in the second one. It raises new questions about a possible interaction between the transfer of auditory-motor learning and serial control processes.

Index Terms: speech units, speech goals, speech production, sensorimotor learning, transfer, auditory feedback perturbation

1. Introduction

Speech can be described as a sequence of units, such as sentences, words, syllables and phonemes. These units have formal linguistic descriptions and their sequencing is structured along a number of linguistic rules involving in particular a hierarchical organization [1]. In the PACT theory [2] these units are proposed to be associated with perceptuo-motor units, i.e. perceptually shaped gestures, associating auditory somatosensory and motor representations. Does this association operate specifically at the level of one of these units? If yes, which unit is it?

The Consonant-Vowel (CV) syllable has often been considered to be the main unit of speech production. In Levelt's model of speech production [3] utterances are broken down into syllables in order to be translated into motor commands, using a mental syllabary. On the other hand, the exemplar-based theories proposes that the words are the smallest units stored in the brain as exemplars rather than prototypes [4]. An exemplar corresponds to an element perceived or produced in a given situation. This perspective accounts for the variability of speech production depending on the speaker and the context. The theory of developmental phonology proposed by Vihman and Croft [5] provides an intermediate view, where the word is the initial basic unit of speech production in infants. The syllable would then emerge from the observation of similarities in the production of the first words [6].

Our goal is to question the nature of the speech units underlying speech motor realization using a classical approach in movement sciences: the transfer of sensorimotor learning. In limb motor control literature, the study of sensorimotor learning transfer is regarded as a behavioral window into the neural mechanisms that underlie the control of movements [7]. The idea is to artificially modify the relation between sensory targets and motor commands for a specific task (e.g. reaching a target with the hand) and observe (1) how this modification is stored in the control system and is used for the same task, once the perturbation has been removed (the so-called after-effects), and (2) how it generalizes to other situations (e.g. reaching a target in another direction, the so-called transfer effect). Houde and Jordan [8] introduced a similar approach with speech using the auditory-motor learning paradigm. In their experiment, participants had to repeat the word "head" in whispered conditions while they heard their voice in headphones. The auditory feedback was altered in real time by shifting formants so that the vowel sounded more like another one (e.g. changing "head" towards "had). During a training phase, the speakers gradually compensated for the perturbation, by pronouncing something closer to "hid", suggesting sensorimotor recalibration that is designated as adaptation. The after-effect can then be assessed by comparing the formants after stopping the perturbation to their baseline values measured before the training. Purcell and Munhall [9] studied the influence of several experimental parameters (e.g. amplitude of the auditory feedback perturbation, duration of the training) on the adaptation and on the after-effect decay speed. Houde and Jordan [10] also showed that the learning performed on one utterance can transfer partially to other utterances, in particular to the same vowel in another CVC word. This was regarded as evidence for the existence of a phoneme representation in speech production.

The paradigm of auditory-motor learning transfer was then used in several studies, mostly with CVC words. In Rochet-Capellan et al. (2012) [11], different groups of speakers were trained with different C-V-/n/ words (e.g. pen, ten). The transfer was then assessed to the testing word "pen" in all speakers. A significant transfer was observed in most groups, although its amplitude

http://dx.doi.org/10.21437/Interspeech.2016-262

was shown to vary according to the acoustic similarity between the training word and the testing word. The influence of similarity was also observed with Mandarin in the pattern of generalization from the triphthong /iau/ to other vowels [12]. Effects of linguistic factors were investigated in other studies. The lexical status (real word vs. pseudo-word) may influence the amplitude of the auditory-motor adaptation [13] while the word frequency seems to play a role in the transfer amplitude [14].

To our knowledge, the transfer of auditory-motor learning has not been used yet to question directly the nature of speech production units. Are motor commands associated with speech goals memorized at the level of the syllable or rather at the level of the word?

2. Method

2.1. Experimental design and hypotheses

The objective of this study was to contrast the transfer of sensorimotor learning at the level of the CV syllable and the corresponding disyllabic word. The training word was the syllable /be/ for all participants. At the end of the training phase, one group of participants (Group 1) pronounced /be/ while another group (Group 2) pronounced /bebe/ (which means "baby" in French), as testing words to assess the transfer. Our predictions were that if the transfer occurs at the syllable level we should observe a transfer equivalent in both groups and in both vowels in /bebe/. If it occurs at a word level only then no transfer to any of the /be/ syllables in /bebe/ should happen. Finally, if the transfer occurs at both levels, then a transfer smaller than in /be/ should occur in /bebe/.

The selection of the vowel /e/ was driven mainly by auditory perturbation constraints while the consonant /b/ was chosen to limit coarticulation. Linguistic factors were also taken into consideration in this choice. The syllable /be/ is as frequent at the beginning of French words as it is in the middle or at the end, as controlled with Lexique 2 database [15]. The frequency of auditory and articulatory neighbors (/bɛ/ and /bi/) had to be low since it may influence the amplitude of the adaptation [14].

2.2. Subjects, task and online perturbation

Thirty-six native speakers of French (15 females), from 18 to 35 years old, took part in the experiment. They had no reported language or audition impairment, and were naïve to the purpose of the experiment. They were seated in front of a monitor in a soundproof room, and wore headphones equipped with a microphone (Sennheiser HME 26-II-600). Words were displayed on a screen. The participants had to read them aloud in a natural way, without shouting or whispering. They were hearing their voice in the headphones at about 70 dB mixed with speech shaped noise at 50 dB.

A real-time perturbation of the first two formants (F1 and F2) was realized using the Audapter system [16]. The auditory feedback alteration consisted in transforming the vowel /e/ into /ɛ/ by increasing F1 by 27% and decreasing F2 by 10%. It triggered a 14ms delay which is not disruptive [17][8].

2.3. Experimental procedure

The experimental procedure is described in Figure 1. In a pre-test (block 1 to 3) speakers had to produce syllables in the /b/+V form (V among /a/, /ɛ/, /e/, /i/ and /u/) as well as minimal pairs (like real French words /epe/ and /epɛ/) to explore their vocalic triangle and the contrast /e/-/ɛ/ in their production, respectively.

The experiment then consisted of 14 blocks of 20 trials each (blocks 4 to 17). At the end of each block, the participants had to press a key to continue the experiment. In the baseline phase (blocks 4 to 6) speakers produced /be/ and then /bebe/ without any perturbation. During the training phase, the auditory feedback perturbation was gradually set up (block 7) and then maintained at its maximal amplitude (blocks 8 to 15) while both groups were pronouncing /be/. The perturbation was then stopped and speakers had to pronounce their testing word (/be/ in Group 1 and /bebe/ in Group 2, transfer phase, block 16). Finally both groups pronounced /be/ again to measure the after-effect (block 17).

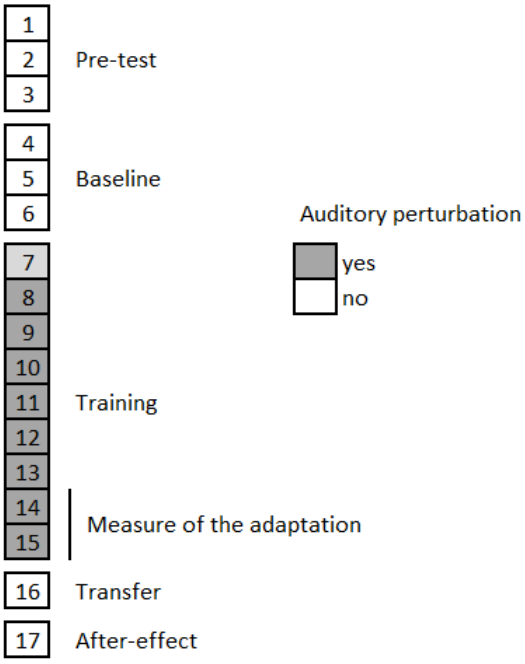

Figure 1: *The experiment consisted of 5 phases, each number corresponding to one block of 20 trials. The auditory perturbation (in grey) was applied during the training phase.*

2.4. Data analysis

F1 and F2 were assessed in a 30ms window in the stable part of each produced vowel. Formants were then expressed as a percentage of change compared to the corresponding vowel measured in the baseline phase in the same speaker. For example, the first /e/ of each repetition of /bebe/ was compared to the average first /e/ of /bebe/ productions in the baseline. The main measure chosen to study transfer was F2-F1, with F1 and F2 expressed in Barks.

$$fBark = 7 * argsinh\left(\frac{fHertz}{650}\right) \text{ [18]} \quad (1)$$

The adaptation in each speaker was estimated by comparing F2-F1 in the last 40 trials of the training phase to the baseline using a one-tailed paired t-test. A one-way between-subjects ANOVA was carried out to assess any difference in the adaptation between the two groups. An ANOVA could not be conducted on transfer data since the position of the syllable in the word is a within-subject factor that only applies to Group 2 (the group

being a between-subject factor). The transfer was therefore assessed using planned comparisons. In block 16, the F2-F1 change in /be/ relative to the baseline was compared to the change in the first vowel of /bebe/ using a one-tailed t-test. The transfer in the first vs. the second vowel of /bebe/ was contrasted by carrying out a two-tailed paired t-test. The intensity, pitch and duration of the first vs. the second vowel were compared using two-tailed paired t-tests. A one-way between-subjects ANOVA was conducted to assess any difference in the after-effect between the two groups. In particular this could show any possible impact of the transfer phase on the after-effect decay.

3. Results

3.1. Adaptation

We were expecting participants to compensate for the perturbation of their auditory feedback by increasing their F2-F1 in /be/ during the training phase. Since this study is focused on transfer of adaptation, only participants who showed a significant adaptation were selected in the analysis. There were 27 such participants, of which 13 speakers (3 women) were in Group 1 and 14 speakers (3 women) were in Group 2.

The evolutions of F2-F1 for the 3 main phases of interest are represented in Figure 2. During the adaptation phase, F2-F1 increased on average by 8.0% (±0.6%) relative to the baseline values. The adaptation amplitude was significant. It did not depend on the group ($F(1,25) = 0.026$, $p = 0.872$). It represented a compensation of about 25% of the perturbation, which roughly matches results obtained in previous studies [19].

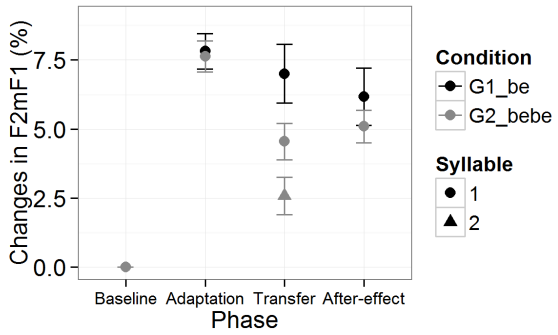

Figure 2: *Evolution of F2-F1 in percentage of change compared to baseline, by phase, by group and by syllable. Each point represents an average of the trials of the phase for the group. The bar represents the standard error.*

3.2. Transfer and after-effect

During the transfer phase, F2-F1 increased by 7.0% (±1.0%) for /be/ (Group 1) relatively to the baseline. In Group 2, the amplitude of the transfer was 4.6% (±0.7%) in the first syllable and 2.6% (±0.7%) in the second syllable. The change was significantly higher in /be/ (Group 1) than in the first syllable of /bebe/ (Group 2) as revealed by a t-test ($t(25) = 1.97$, $p = 0.03$). The transfer in the first syllable was significantly higher than in the second syllable ($t(13) = 3.25$, $p = 0.006$).The amplitude of the after-effect, always tested with /be/, was 6.1% (±1.0%) in Group 1 and 5.0% (±0.6%) in Group 2. There was no significant difference between groups ($F(1,25) = 0.844$, $p = 0.367$) suggesting that the after-effect decay was not sensitive to the transfer phase.

Additional analyses related to prosodic cues show that the second syllable lasted longer (133ms ±10) than the first one (99ms ±5; $t(13) = -3.66$, $p = 0.003$). No significant difference between the two syllables was found regarding the pitch ($t(13) = -0.18$, $p = 0.8$) or the intensity ($t(13) = 0.29$, $p = 0.8$).

4. Discussion

The objective of this study was to contrast the transfer of sensorimotor learning at the level of the CV syllable and the word. In this section we will first comment on the two main effects related to the transfer, in order to provide a response to our research question. Then we will focus on the adaptation and discuss some limitations observed in this experiment.

4.1. CV vs CVCV

Our results indicate that some transfer occurs at the level of the syllable since a significant transfer occurred from /be/ to /bebe/. However the difference of amplitude in the transfer observed in /be/ (Group 1) and in the first and second syllables of /bebe/ (Group 2) suggests that transfer is also influenced by a larger context, in this case the produced word. This result questions the assumption that a "mental syllabary" [3] would represent the only link between a targeted speech sequence and the articulatory gestures to produce it. Overall, the results suggest that several units, including both the CV syllable and the word, may contribute to define the motor goals associated with an utterance. This observation in adults is consistent with the developmental theory suggested by Vihman and Croft [5] according to which several speech production units may develop gradually and therefore coexist in adults.

It may be noticed that the speakers who produced /bebe/ in the transfer block pronounced twice as many syllables as the other group. The higher number of repetitions of the syllable /be/ could be expected to trigger a faster after-effect decay. If this was the only factor driving the higher transfer in /be/ than in /bebe/ then we should observe a difference between groups in the after-effect amplitude measured in the block that followed the transfer. However, there was no significant difference between groups in the after-effect amplitude. Thus the number of repetitions could not be regarded as the main factor explaining the difference of transfer amplitude between groups.

4.2. The position of the syllable

The significant difference of transfer amplitude observed in the first and the second syllable of /bebe/ will be referred to as a position effect. This is another observation that questions the concept of a mental syllabary as it suggests an influence of the serial order of speech on the transfer. Several explanations could account for this effect.

First, this position effect could be linked to an influence of prosody on sensorimotor learning transfer. A stressed syllable could be produced in a more precise way than an unstressed syllable. In most participants, the second syllable lasted longer than the first one, which is consistent with French stress pattern [20]. However, no correlation among speakers was observed between the stress pattern and the difference of transfer amplitude between the first and the second syllable. The potential link between prosodic patterns and transfer amplitude could be further investigated by evaluating transfer while

asking speakers to produce more contrasted stress patterns in CVCV words.

The observed position effect may also depend on the auditory perturbation. The training was realized on a monosyllabic word. The perturbed syllable was thus the first one (and the only one) of the utterance. In this context, the speakers may have learned to modify more specifically the first syllable of the sequence. The auditory-motor paradigm has not yet been applied in humans with words consisting of several syllables. Some research in zebra finches [21] focused on the transfer of auditory-motor learning in song sequences. The birds were wearing small headphones providing them with their auditory feedback. An alteration of their pitch was applied on a specific syllable in a given position within their vocalization. It led the birds to adapt by compensating partially for the perturbation as observed in humans. At the end of the perturbation a transfer was observed in the same syllable in any position of the sequence. The transfer partially spread to temporally nearby syllables. Selective perturbation in terms of position in the sequence would raise new perspectives in studying how speech units are put into sequence from a motor control perspective.

4.3. Adaptation to the perturbation and limitations

Twenty-seven speakers out of 36 (e.g. 75%) adapted to the perturbation by significantly increasing F2-F1. The 9 remaining speakers were all women. The perturbation did not seem to work properly in 5 of them and the 4 remaining speakers did not show significant increase of F2-F1. Although this depletion rate can seem quite high, it is very similar to what has been observed in previous studies using the same perturbation system. In Cai et al. (2010) [12] 31 speakers out of 40 were included in the results, 4 speakers had formants that were not well detected by the system, which therefore did not apply the targeted perturbation properly. The last 5 speakers were reported to follow the perturbation instead of compensating for it. It is not mentioned whether these speakers were males or females. Some studies with other perturbation systems report lower depletion rate. In MacDonald et al. (2010) [22] only one male out of 20 did not show any compensation.

The parameters of our perturbation system may be better fine tuned to adjust to women's voices, but formants detection is inherently more difficult in high pitched voices since the gap between two harmonics is higher.

Beyond technical limitations, some speakers may not have adapted to the perturbation because they rely more on their somatosensory feedback than on their auditory feedback during speech production. The existence of a sensory preference was suggested by a study in which alterations of the auditory and the somatosensory feedbacks were performed simultaneously in speakers [23].

Another possible explanation for the absence of compensation in some speakers could be that they do not make any distinction between /e/ and /ɛ/ in production as these phonemes are often confused in French speakers. However no correlation was established between the amplitude of speakers' adaptation and their contrast in production between /e/ and /ɛ/ as assessed in the pretest.

5. Conclusions

The objective of this study was to contrast the contribution of the word and the syllable in the definition of speech motor goals. We artificially modified the production of the syllable /be/ in French speakers by altering their auditory feedback. After speakers significantly adapted to the perturbation, the transfer to /bebe/ was measured under normal feedback and compared to the after-effect in /be/. The results show that the change in /be/ due to adaptation partially transferred to the production of /bebe/, suggesting that both units amongst others may contribute to defining speech motor goals. The difference of transfer amplitude observed in /bebe/ between the first and the second syllable raises new questions with regards to serial order. Several factors, from a potential prosody effect to a possible influence of the monosyllabic structure of the training word could contribute to this last effect. Further investigation is required to test these assumptions.

6. Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2013 Grant Agreement no. 339152).

The authors would like to thank Christophe Savariaux and Takayuki Ito (CNRS, Gipsa-lab, F-38000 Grenoble, France) for their contribution to this work, in particular their help in the experimental setup. Many thanks to Frederic Field who helped enhancing the wording in this article.

7. References

[1] C. Fougeron and P. A. Keating, "Articulatory strengthening at edges of prosodic domains.," *J. Acoust. Soc. Am.*, vol. 101, no. 6, pp. 3728–40, 1997.

[2] J. L. Schwartz, A. Basirat, L. Ménard, and M. Sato, "The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception," *J. Neurolinguistics*, vol. 25, pp. 336–354, 2012.

[3] W. J. M. Levelt, "Models of word production," *Trends Cogn. Sci.*, vol. 3, no. 6, pp. 223–232, 1999.

[4] R. Välimaa-Blum, "The phoneme in cognitive phonology: episodic memories of both meaningful and meaningless units?," *CogniTextes*, vol. 2, 2009.

[5] M. Vihman and W. Croft, "Phonological development: Toward a 'radical' templatic phonology," *Linguistics*, vol. 45, no. 4, pp. 683–725, 2007.

[6] A. Cristia and P. Hallé, "Global and detailed speech representations in early language acquisition," in *Speech Planning and Dynamics*, Peter Lang, 2012, pp. 11–38.

[7] R. Shadmehr, "Generalization as a Behavioral Window to the Neural Mechanisms of Learning Internal Models," *Hum. Mov Sci.*, vol. 29, no. 6, pp. 997–1003, 2012.

[8] J. F. Houde and M. I. Jordan, "Sensorimotor adaptation in speech production.," *Science*, vol. 279, no. 1998, pp. 1213–1216, 1998.

[9] D. W. Purcell and K. G. Munhall, "Compensation following real-time manipulation of formants in isolated vowels.," *J. Acoust. Soc. Am.*, vol. 119, no. January, pp. 2288–2297, 2006.

[10] J. F. Houde and M. I. Jordan, "Sensorimotor Adaptation of Speech I : Compensation and Adaptation," *JSLHR*, vol. 45, no. April 2002, pp. 295–310, 2002.

[11] A. Rochet-Capellan, L. Richer, and D. J. Ostry, "Nonhomogeneous transfer reveals specificity in speech motor learning," *J. Neurophysiol.*, vol. 107, pp. 1711–1717, 2012.

[12] S. Cai, S. S. Ghosh, F. H. Guenther, and J. S. Perkell, "Adaptive auditory feedback control of the production of formant trajectories in the Mandarin triphthong /iau/ and its pattern of generalization.," *J. Acoust. Soc. Am.*, vol. 128, pp. 2033–2048, 2010.

[13] N. J. Bourguignon, S. R. Baum, and D. M. Shiller, "Lexical-perceptual integration influences sensorimotor adaptation in

speech.," *Front. Hum. Neurosci.*, vol. 8, no. April, p. 208, 2014.

[14] A. F. Frank, "Integrating Linguistic, Motor, and Perceptual Information in Language Production," *Diss. Abstr. Int. B Sci. Eng.*, vol. 72, p. 2454, 2011.

[15] B. New, C. Pallier, M. Brysbaert, and L. Ferrand, "Lexique 2: a new French lexical database.," *Behav. Res. Methods. Instrum. Comput.*, vol. 36, no. 3, pp. 516–524, 2004.

[16] S. Cai, M. Boucek, S. S. Ghosh, F. H. Guenther, and J. S. Perkell, "A System for Online Dynamic Perturbation of Formant Trajectories and Results from Perturbations of the Mandarin.pdf," *Int. Semin. Speech Prod. 2008*, pp. 65–68, 2008.

[17] A. J. Yates, "Delayed auditory feedback," *Psychol. Bull.*, vol. 60, no. 3, pp. 213–232, 1963.

[18] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Objective Measure of Certain Speech Signal Degradations Based on Masking Properties of Human Auditory Perception," in *Frontiers of Speech Communication Research*, B. (ed. & pref. . Lindblom and S. E. G. (ed. Öhman pref., & introd.), Eds. New York: Academic, 1979, pp. 217–229.

[19] D. W. Purcell and K. G. Munhall, "Adaptive control of vowel formant frequency: evidence from real-time formant manipulation.," *J. Acoust. Soc. Am.*, vol. 120, pp. 966–977, 2006.

[20] S.-A. Jun and C. Fougeron, "Realizations of accentual phrase in French intonation," *Probus*, vol. 14, no. 1, pp. 147–172, 2002.

[21] L. A. Hoffmann and S. J. Sober, "Vocal Generalization Depends on Gesture Identity and Sequence," *J. Neurosci.*, vol. 34, no. 16, pp. 5564–5574, 2014.

[22] E. N. MacDonald, R. Goldberg, and K. G. Munhall, "Compensations in response to real-time formant perturbations of different magnitudes.," *J. Acoust. Soc. Am.*, vol. 127, no. February, pp. 1059–1068, 2010.

[23] D. R. Lametti, S. M. Nasir, and D. J. Ostry, "Sensory Preference in Speech Production Revealed by Simultaneous Alteration of Auditory and Somatosensory Feedback," *J. Neurosci.*, vol. 32, no. 27, pp. 9351–9358, 2012.