# Speech planning as an index of speech motor control maturity

*Guillaume Barbier [1], Pascal Perrier [1], Lucie Ménard [2], Yohan Payan [3], Mark K. Tiede [4], Joseph S. Perkell [5]*

[1] Speech and Cognition Department, GIPSA-lab & Grenoble University, Grenoble, France
[2] Department of Linguistics, Université du Québec à Montréal, Montréal, Québec, Canada
[3] TIMC-IMAG Laboratory, Grenoble University & CNRS, La Tronche, France
[4] Haskins Laboratories, New Haven, Connecticut, USA
[5] Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

guillaume.barbier@gipsa-lab.grenoble-inp.fr

## Abstract

This paper investigates speech motor control maturity in 4-year-old Canadian French children. Acoustic and ultrasound data recorded from four children, and for comparison, from four adults, are presented and analyzed. Maturity of speech motor control is assessed by measuring two characteristics: token-to-token variability of isolated vowels, as a measure of motor control accuracy, and extra-syllabic anticipatory coarticulation within $V_1$-C-$V_2$ sequences. In line with theories of optimal motor control, anticipatory coarticulation is assumed to be based on the use of internal models of the speech apparatus and its efficiency is considered to reflect the maturity of these representations. In agreement with former studies, token-to-token variability is larger in children than in adults. An anticipation of $V_2$ in $V_1$ was found in all adults but in none of the children studied so far. These results indicate that children's speech motor control is immature from two perspectives: insufficiently accurate motor control patterns for vowel production, and inability to anticipate forthcoming gestures. Both aspects are discussed and interpreted in the context of the immaturity of the internal representations of the speech motor apparatus in 4-year-old children.

**Index Terms**: speech production development, speech motor control, co-articulation, planning.

## 1. Introduction

The process of speech motor learning is not finished when a child produces his first words. On the contrary, the ability to control the spatio-temporal organization of speech gestures is at a crucial stage of its development. The maturation of speech motor control, for which co-articulation of gestures is an index, is a long process that seems to be fully accomplished in late adolescence only [1, 2].

Lingual coarticulation in children has been investigated in numerous studies (e.g., [3-9]), but remains poorly understood because these studies have provided contradictory conclusions. Some of the possible reasons for the confusion are the small numbers of participants, a large spread in age groups, and the use of acoustic measurements only (except for [7-9]). To help resolve this debate for young children, we designed a study combining articulatory (ultrasound tongue imaging) and acoustic measurements, focused on a narrow age group (from 4 years to 4 years 11 months) and involving a substantial number of participants (20 children, 10 young adults).

In this study, we focused on token-to-token variability in isolated vowel production and on anticipatory co-articulation of $V_2$ in $V_1$ during the production of $V_1$-C-$V_2$ sequences. In a theoretical context which assumes that the production of speech sequences involves gesture planning [10], extra-syllabic anticipatory co-articulation is considered to be a potentially useful index of the maturity of speech motor control. It is assumed to involve the capacity to predict the effect of motor commands on speech gestures and sounds and the ability to integrate those predictions into the planning strategy. In this work we evaluate the hypothesis of immature speech motor control in by comparing the performance of 4-year-old children to that of adults.

In current speech production models (e.g., [11]), the process of speech production relies on feedforward and feedback control mechanisms to achieve auditory and somatosensory goals (see [12] for a recent review). These mechanisms make use of internal models [13] that are implemented as artificial neural networks. In this theoretical framework, it is assumed that experience with the sensory consequences of motor acts is learned and stored in a forward model. This forward model predicts the sensory consequences of motor acts by generating efferent copies for comparison with feedback from those actually executed. Extensive use of forward internal models enables the learning of inverse models that generate motor commands from the specification of desired motor goals [14]. Motor learning is also assumed to involve optimal planning aiming at minimizing a measure of effort in a sequence of movements [15].

In this context, it is assumed that both adults and children are likely to use neural representations of their motor systems to predict the sensory consequences of motor acts. Presumably, mature speakers have acquired implicit knowledge of the amount of produced variability that is compatible with correct perception of the produced sounds by listeners. They use this tolerance of variability to plan and execute a sequence of speech gestures with minimized articulatory effort [16]. Our hypothesis is that 4-year-old children do not have sufficient experience with the sensory consequences of motor acts to be able to implement this effort-minimizing strategy effectively, particularly with respect to variability in sound categories. This idea emerges from a literature review of studies of arm movement and speech production in children (e.g, [17-20]) and from speech perturbation studies in children (e.g.[21, 22]). As a consequence, we predict that a child’s ability to plan upcoming gestures could be either limited or inaccurate. This hypothesis is tested here through the analysis of extra-syllabic anticipatory co-articulation by comparing the performance of 4-year-old children to that of adults.

# 2. Material and Methods

## 2.1. Participants

Twenty young 4-year-old Canadian French children (4 years 0 months to 4 years 11 months) and 10 Canadian French adults (18-28 years old) were recruited in Montréal for the experiment. Canadian French was the first language of all participants. All children lived in monolingual French families and were educated in French only. Participants reported no history of speech or hearing problems. All participants showed normal audition, by passing a bilateral pure tone screening test at 20dB at 250Hz, 500Hz, 1000Hz, 2000Hz and 4000Hz before the experiment. All participants and participant's parents, in the case of children, were informed about all the procedures before the experiment and gave their consent. This study was approved by the ethical committee of the Université du Québec à Montréal (UQÀM). This paper presents partial results based on 4 children and 4 adults (complete results will be presented at the conference).

## 2.2. Data acquisition and processing

Ultrasound is a noninvasive imaging technique. It is adaptable for use with very young children, and enables a real-time 2D view of most of the tongue, with good temporal (15Hz-200Hz) and spatial (~1mm) resolution [23, 24]. To make reliable measurements of tongue movement independent of any head movement we used the HOCUS system (Haskins Optically Corrected Ultrasound System, [25]), which uses optical tracking (Optotrak, NDI Certus) of infrared emitting diodes (iREDs), positioned both on the ultrasound probe and on the head of the participant, to provide a representation of the data in a movement-corrected head-centric frame of reference. This approach is appropriate for developmental studies, in that it preserves some freedom of movement for the participants. In this study, an iRED was placed on the chin to allow tongue movements to be dissociated from jaw movements by providing an index of jaw motion.

Synchronous recordings of tongue movement in the midsagittal plane (at NTSC 29.97 Hz) and of the speech signal (at 44.1kHz) were made by the ultrasound device (Sonosite 180 Plus) and a directional microphone. The Optotrak system was used to record audio and the positions of the iREDs concurrently. Synchronization of these data was obtained during post-processing through cross-correlation between the two audio signals. After head-movement correction and alignment to a coordinate system centered on the upper incisors, the data are mapped onto a 3D view in which the position of the iREDs and the tongue imaging plane are visible. Images sampled too far from the midsagittal plane or at an angle to the midsagittal plane exceeding 5° were removed and not used in the study.

## 2.3. Task

Data were collected on-site at Montréal day care centers and at the Laboratoire de Phonétique, UQÀM. Participants were seated in front of the Optotrak device, disguised as a puppet theater, with the ultrasound probe held under their chins by a microphone stand. One experimenter checked that the head of the speaker was not moving too much with reference to the ultrasound probe, and that most of the tongue was visible on the screen; another experimenter controlled the recording (Optotrak and ultrasound) and checked that all the iREDs were visible during the trials.

The task was presented as a puppet game, with a third experimenter serving as puppet master. Puppets were presented in different pairs. The order of appearance of the pairs was randomized. The game took place as follows: the first puppet appeared alone, and asked the participant to pronounce its name. The second puppet did the same. When the participant correctly recognized the names of the two puppets, the game began. The participant's task was to pronounce the name of the puppet when it appeared. In this way, participants had to recall, plan and execute a speech gesture or a sequence of speech gestures.

The corpus was collected as follows, with 8 to 10 repetitions for each isolated vowel or $V_1$-C-$V_2$ sequence:

- Isolated vowels /i e ɛ a u/
- $V_1$-C-$V_2$ sequences with
  C = /b d g/, $V_1$= /ɛ a/ and $V_2$ = /i a/

The isolated vowels were used to measure the dispersion in the F1-F2 plane of vowel token-to-token variability. The $V_1$-C-$V_2$ sequences were composed of $V_1$ vowels for which a certain variability (/a/ and /ɛ/) was expected, and with a high vowel /i/ and a low vowel /a/ as $V_2$. These sequences were designed to measure the effects of anticipating $V_2$ within $V_1$.

## 2.4. Data Analysis

The acoustic signal was downsampled to 16000 Hz in order to achieve more accurate formant detection. Automatic acoustic measurements of the formants in the midpoint of the vowels were made with a Linear Predictive Coding (LPC) algorithm. Because formant tracking is difficult in child speech, with the potential for detection errors, we combined the measures of the frequencies of the maxima in the LPC spectra with the frequencies of the poles in the LPC filter. A range of acceptable formant values for each vowel was used to remove outliers.

The acoustic data were labeled with *Praat* [26]. For vowels, the beginning of the vowel was defined as the first descending zero-crossing of the signal after the clear emergence of F2, and the end of the vowel was defined as the first descending zero-crossing after the disappearance of F2. For stops, the beginning of the consonant was defined as the time where F2 of the preceding vowel disappeared, and the end corresponded to the beginning of the release burst. The beginning and the end of the burst were also labeled. The transition from the stop to the subsequent vowel begins at noise onset and ends at the time when F2 appears.

For articulatory data, the ultrasound images corresponding to the midpoint of the vowels were used. The midsagittal tongue contour was extracted using a semi-automatic procedure developed for the purpose, *GetContour*, similar to other edge extraction tools such as *EdgeTrak* [27]. Contours were converted to 3D head-centric coordinates using the HOCUS procedures described above.

# 3. Results

We present here initial results based on analysis of 4 children and 4 adults. The first result concerns token-to-token variability of isolated vowels (Figures 1-2). Token-to-token variability in the formant space corresponded to the standard deviation. The second result concerns the production of $V_1$ in

$V_1$-C-$V_2$ sequences, depending on the $V_2$ context (Figures 3-8). These two results provide information about the maturity of speech motor control mechanisms in 4-year-old children.

Figures 1 and 2 show the token-to-token variability of isolated vowels in the F1-F2 plane for an adult and for a child.

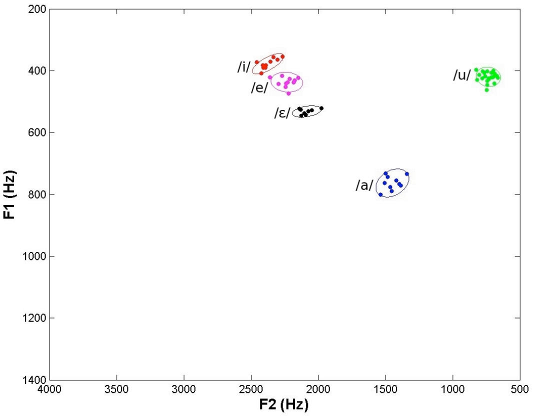

Figure 1: *Token-to-token variability of isolated vowels for an example adult speaker (and dispersion ellipses at 2σ).*

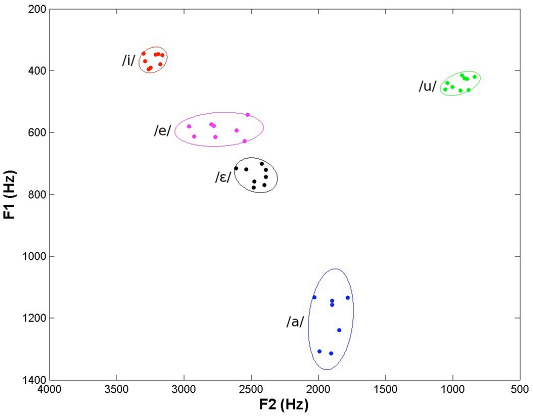

Figure 2: *Token-to-token variability of isolated vowels for an example child speaker (and dispersion ellipses at 2σ).*

These two figures illustrate that token-to-token variability is lower for adults than for children. The mean standard deviation of vowel production, across all vowels and the four speakers, is 20 Hz in F1 and 55 Hz in F2 for adults, and is 43 Hz in F1 and 114 Hz in F2 for children. As the vowel space of children is larger than that of adults, we normalized these dispersion measurements using the /i/-/a/ distance for F1 and the /i/-/u/ distance for F2. After normalization, the dispersion of children's vowel categories is 1.18 times greater in F1 and 1.42 times greater in F2 than that of adults.

Figures 3-6 present the F1-F2 patterns at the vowel midpoint for /a/ as $V_1$ (Figures 3 and 4) and for /ɛ/ as $V_1$ (Figures 5 and 6), for two different anticipated vowels /i/ and /a/. In these figures, the effects of $V_2$ on $V_1$ can be seen as a difference in $V_1$ productions depending on the upcoming vowel. If $V_1$ differs from one context to another, and if this difference occurs in the direction of the upcoming vowel, we can say that an anticipation of $V_2$ in $V_1$ is observed. We expect, in adults, an anticipation of /i/ in /a/ mostly in the antero-posterior direction, that is in F2 (see Figure 3).

Figure 3 shows that for an adult, a clear anticipation of /i/ (as $V_2$) can be seen in the production of /a/ (as $V_1$) in the antero-posterior direction. Figure 4 shows a lack of anticipation of $V_2$ in $V_1$for a child.

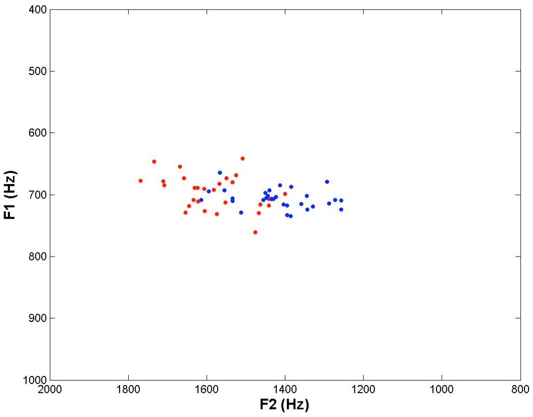

Figure 3: *F1-F2 patterns for /a/ as $V_1$ for all VCV sequences in the context of $V_2$ = /a/ (blue) and $V_2$ = /i/ (red), for an example adult speaker.*

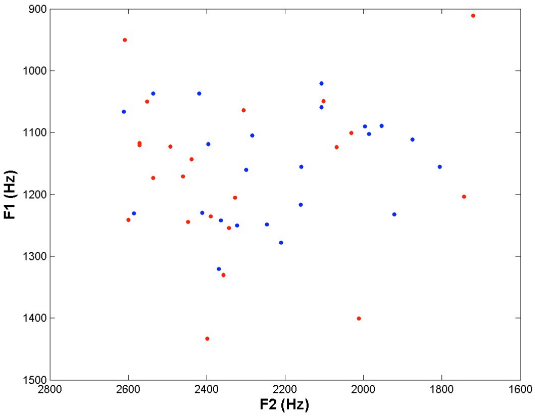

Figure 4: *F1-F2 patterns for /a/ as $V_1$ for all VCV sequences in the context of $V_2$ = /a/ (blue) and $V_2$ = /i/ (red), for an example child speaker.*

Figure 5 shows clear anticipation of $V_2$ in the realization of /ɛ/ (as $V_1$) in the infero-superior direction for an adult, and Figure 6 shows an absence of anticipation for a child.

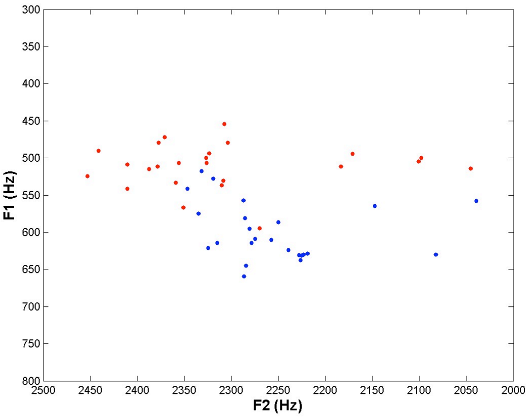

Figure 5: *F1-F2 patterns for /ɛ/ as $V_1$ in all VCV sequences in the context of $V_2$ = /a/ (blue) and $V_2$ = /i/ (red), for an example adult speaker.*

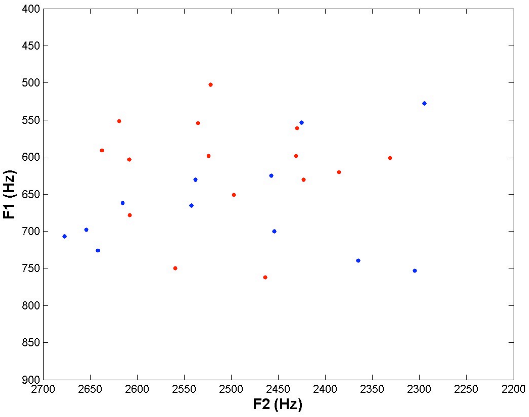

Figure 6: *F1-F2 patterns for /ɛ/ as $V_1$ in all VCV sequences in the context of $V_2$ = /a/ (blue) and $V_2$ = /i/ (red), for an example child speaker.*

These results apply to the 4 adults and the 4 children analyzed thus far. Two univariate ANOVAs were performed on F1 and on F2 separately, to test whether $V_1$ was significantly different according to the $V_2$ contexts. For all adults, this anticipatory effect is statistically significant (at $p<0.01$), with the largest effect in F2 for three adults, and in F1 for one adult with $V_1$ = /a/. This indicates that the anticipation of $V_2$ (for $V_1$ = /a/) was mainly performed in the antero-posterior direction for three of the adults, and in the infero-superior direction for the one adult. This pattern was confirmed by looking at articulatory data. For $V_1$ = /ɛ/, the anticipation was observed mostly in F1 for adults. For all children, no statistically significant effect (in either F1 or in F2) was found (at p=0.01) for both vowels. This indicates that none of the children so far studied anticipate $V_2$ within the /a/ and /ɛ/ realizations of $V_1$.

To illustrate this finding with ultrasound tongue imaging, we present in Figures 7 and 8 examples of tongue contours taken at the midpoint of /a/ as $V_1$ in /ada/ versus /adi/ sequences for a child and an adult.

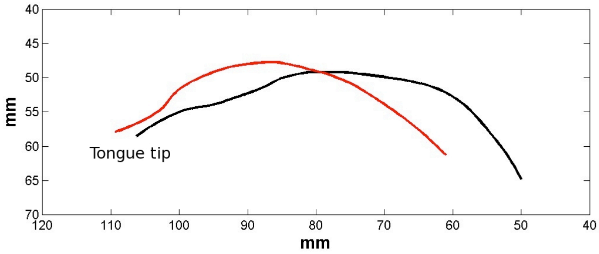

Figure 7: *Example of tongue contours for /a/ as $V_1$ in the contexts $V_2$ = /a/ (black) and $V_2$ = /i/ (red) for an example adult speaker facing left.*

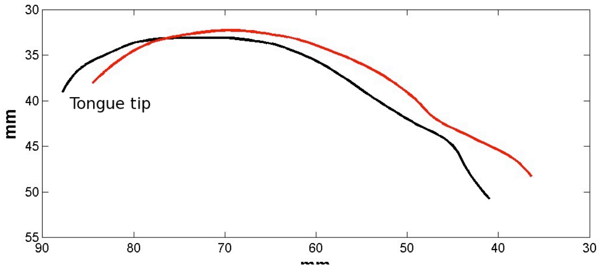

Figure 8: *Example of tongue contours for /a/ as $V_1$ in the contexts $V_2$ = /a/ (black) and $V_2$ = /i/ (red) for an example child speaker facing left.*

These examples illustrate clear anticipation for the adult and the absence of such anticipation for the child.

## 4. Conclusion

Initial results of our study of the maturity of speech motor control in 4-year-old children have shown greater token-to-token variability than in adults. This result is consistent with numerous studies (e.g., [17-20, 28]) showing that humans exhibit decreasing gestural variability with age until adulthood. Our data also show that children tend not to anticipate $V_2$ in $V_1$ during the production of $V_1$-C-$V_2$ sequences. In sum, these results indicate that 4-year-old children's speech motor control is immature from two perspectives: incompletely optimized motor control patterns for vowel production, and inability to anticipate forthcoming gestures. Our interpretation is that the neural representations of 4-year-old children's speech motor systems are immature, particularly in their incapacity to account for the appropriate variability compatible with correct perception of the target sound.

## 5. Acknowledgements

This work is partly supported by FQRNT project N° 147877 (Lucie Ménard) and the ANR project ANR-08-BLAN-0272 (Pascal Perrier). We want to thank Amélie Prémont, Christine Turgeon, Paméla Trudeau-Fisette and Dominique Côté for their crucial help concerning the experiment. We are thankful to the infant school Le Sablier and particularly its co-director Carole Langlois for her involvement in the project. We are also very grateful to all our participants.

# 6. References

[1] Smith, A. (2010). Development of Neural Control of Orofacial Movements for Speech. In *Handbook of Phonetic Sciences*. W. Hardcastle and J. Laver (Eds). Oxford: Blackwell.

[2] Walsh, B. and Smith, A. (2002). Articulatory movements in adolescents: evidence for protracted development of speech motor control processes. *Journal of Speech, Language and Hearing Research*, 45, 1119–1133.

[3] Sereno, J. A. and Lieberman, P. (1987). Developmental aspects of lingual coarticulation. *Journal of Phonetics*, 15, 247–257.

[4] Nittrouer, S., Studdert-Kennedy, M. and Neely, S.T. (1996). How children learn to organize their speech gestures: further evidence from fricative-vowel syllables. *Journal of Speech and Hearing Research*, 39, 379–389.

[5] Siren, K.A. and Wilcox, K.A. (1995). Effects of lexical meaning and practiced productions on coarticulation in children's and adults' speech. *Journal of Speech and Hearing Research*, 38, 351–359.

[6] Goodell, E. W. and Studdert-Kennedy, M. (1993). Acoustic evidence for the development of gestural coordination in the speech of 2-year-olds: A longitudinal study. *Journal of Speech and Hearing Research*. 36, 707–727.

[7] Zharkova, N., Hewlett, N. and Hardcastle W. J. (2011). Coarticulation as an Indicator of Speech Motor Control Development in Children: An Ultrasound Study. *Motor Control*, 15, 118-140.

[8] Zharkova, N., Hewlett, N., and Hardcastle, W. J. (2012). An ultrasound study of lingual coarticulation in /sV/ syllables produced by adults and typically developing children, *Journal of the International Phonetic Association*. 42, 193–208.

[9] Noiray, A., Ménard, L. and Iskarous, K. (2013). The development of motor synergies in children: Ultrasound and acoustic measurements. *Journal of the Acoustical Society of America*, 133, 444–452.

[10] Whalen, D. H. (1990) Coarticulation is largely planned. *Journal of Phonetics* 18, 3-35.

[11] Guenther, F. H., Ghosh, S. S. and Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, 96, 280–301.

[12] Perkell, JS (2012). Movement goals and feedback and feedforward control mechanisms in speech production, *Journal of Neurolinguistics* 25, 382-407.

[13] Jordan, M.I. and Rumelhart, D.E. (1992). Forward models: supervised learning with a distal teacher, *Cognitive Science*, 16, 307-354.

[14] Kawato M., Furawaka K. and Suzuki R. (1987). A hierarchical neural net-work model for the control and learning of voluntary movements. *Biological Cybernetics*, 56, 1 – 17.

[15] Kawato, M., Maeda, Y., Uno, Y and Suzuki, R. (1990). Trajectory Formation of Arm Movement by Cascade Neural Network Model Based on Minimum Torque-Change Criterion, *Biological Cybernetics*, 62, 275-188.

[16] Perkell, J. S., Zandipour, M., Matthies, M. L. and Lane, H. (2002). Economy of effort in different speaking conditions I: a preliminary study of intersubject differences and modeling issues. *Journal of the Acoustical Society of America*, 112, 1627–1641.

[17] Lambert, J. and Bard, C. (2005). Acquisition of visuomanual skills and improvement of information processing capacities in 6- to 10-year-old children performing a 2D pointing task. *Neuroscience Letters*, 377, 1-6.

[18] Jansen-Osmann, P. · Richter, S., Konczak, J. and Kalveram, K.-T. (2002). Force adaptation transfers to untrained workspace regions in children: Evidence for developing inverse dynamic motor models. *Experimental Brain Research*, 143, 212-220.

[19] Hourcade, J. P., Bederson, B. B., Druin, A. and Guimbretière, F. (2004). Differences in Pointing Task Performance Between Preschool Children and Adults Using Mice, *ACM Trans. Comput. Hum. Inter*. 11. 357-386.

[20] Jansen-Osmann, P., Richter, S., Konczak, J. and Kalveram, K.-T. (2002). Force adaptation transfers to untrained workspace regions in children, *Experimental Brain Research*. 143, 212–220.

[21] Baum, S.R. and Katz, W.F. (1988). Acoustic analysis of compensatory articulation in children. *Journal of the Acoustical Society of America*, 84, 1662-1668.

[22] Ménard, L., Perrier, P., Savariaux, C., Aubin, J. and Thibeault, M. (2008). Compensation strategies for a lip-tube perturbation of French [u]: An acoustic and perceptual study of 4-year-old children, *Journal of the Acoustical Society of America*, 124, 1192–1206.

[23] Aubin, J. and Ménard, L. (2006). Compensation for a labial perturbation: An acoustic and articulatory study of child and adult French speakers. In *7th International Seminar on Speech Production*, Ubatuba, Brazil, 209-216.

[24] Song, J.-Y., Demuth, K., Ménard, L. and Shattuck-Huffnagel, S. (2010). Acoustic and gestural characteristics of a 2-year-old's American English coda consonants. Poster presented at *Laboratory Phonology*, 12, New Mexico.

[25] Whalen, D.H., Iskarous, K., Tiede, M.K., Ostry, D.J., Lehnert-LeHouillier, H., Vatikiotis-Bateson, E. and Hailey, D.S. (2005). The Haski ns Opti cal l y Corrected Ul trasound System (HOCUS). *Journal of Speech, Language, and Hearing Research*, 48, 543–553.

[26] Boersma, P., and Weenink, D. (1996). Praat, a system for doing phonetics by computer, version 3.4, *Report No. 132, Institute of Phonetic Sciences of the University of Amsterdam*, 1–182.

[27] Li, M., Kambhamettu, C., and Stone, M. (2005). Automatic contour tracking in ultrasound images, *Clinical Linguistics and Phonetics*, 6, 545–554.

[28] Smith, A., and Goffman, L. (1998). Stability and patterning of speech movement sequences in children and adults, *Journal of Speech, Langage and Hearing Research*, 41, 18–30.