EXTRINSIC TALKER NORMALIZATION ALTERS SELF PERCEPTION DURING SPEECH

Nicolas J. Bourguignon[1,2,4], Shari R. Baum[3,4], Douglas M. Shiller[1,2,4]

[1]*Département d'orthophonie et d'audiologie, University of Montreal, Canada;* [2]*Centre de recherche du CHU Sainte-Justine, Montreal Canada;* [3]*School of Communication Sciences and Disorders, McGill University, Montreal, Canada;* [4]*Centre for Research on the Brain, Language and Music, Montreal, Canada.*

ABSTRACT

In this study, we investigated the effects of changes in formant structure of externally presented speech signals on participants' auditory perception of their *own* speech output (i.e., feedback) during a word production task. The study involved a novel combination of two previously established research paradigms: (1) sensorimotor adaptation to altered auditory feedback during speech, and (2) extrinsic talker normalization of vowel perception through the presentation of carrier-phrases spoken with different formant patterns. The results suggest that the formant frequencies of a carrier-phrase presented immediately prior to word production serve as a frame of reference for the perception of self-generated speech outcomes, thereby influencing subsequent speech targets. This finding extends other recent evidence indicating that the auditory processing of speech sounds guiding speech production is highly flexible and adaptive under a range of different conditions.

Keywords: Speech perception, speech motor control, speaker normalization, sensorimotor adaptation.

1. INTRODUCTION

As speech typically takes place under noisy and variable conditions, the human language processing system is equipped with mechanisms for reducing the impact of linguistically irrelevant variation in the incoming speech signal [1]. Among these is a capacity for speaker normalization (SN), whereby talker-related phonetic variability – for example pertaining to age or gender – is accounted for such that segments will be correctly identified regardless of who produces them [2, 3]. In support for the existence of "extrinsic" SN processes that rely on context (prior exposure to talker properties), Ladefoged & Broadbent [4, 5] and others [6] have demonstrated an influence of the vowel formant frequencies of an introductory carrier phrase (e.g., "*Please say what this word is…*") on listeners' identification of a subsequently presented word containing an ambiguous vowel (e.g., intermediate in F1 between /bƐt/ and /bIt/). Specifically, participants were more likely to identify the ambiguous vowel as one associated with a lower F1 (e.g., /I/) when the carrier-phrase's F1 was relatively high, and as one associated with a higher F1 (e.g., /Ɛ/) when the carrier-phrase's F1 was relatively low. The vowel formants of the carrier-phrase thus provide listeners with a perceptual frame of reference, which in turn influences the listeners' categorization of a subsequently presented vowel. Such findings support the idea that listeners adapt their acoustic-phonetic representations of vowel sounds to accommodate previously perceived differences in talker vocal tract properties, thus preserving their ability to categorize speech sounds in the face of inter-speaker variation.

These rapid, context-dependent changes in acoustic-perceptual representations of speech sounds stand somewhat at odds with the characterization of representations of speech sounds in models of *speech production*, where such sounds are generally considered as accurate, stable sensory targets that serve as the primary goals of speech movements [9,10]. Studies demonstrating speech motor adaptation to perturbations of auditory feedback during speech [7, 16] have provided substantial evidence for the stability of acoustic speech targets. More recent studies, however, have begun to challenge this notion by demonstrating that the auditory targets of speech production can be readily altered through reinforcement-based perceptual training [12, 15] or top-down lexical effects [13], directly impacting talkers' patterns of speech motor adaptation to auditory feedback manipulations.

The present study aimed to extend these recent findings by investigating another way in which extrinsic information may influence sensory processing during speech production, namely extrinsic SN. The study involved a novel combination of two distinct paradigms: (1) the normalization of vowel perception to differences in formant properties of extrinsically presented speech (i.e., the approach developed by Ladefoged and Broadbent [4]), and (2) sensorimotor adaptation of speech production to altered auditory feedback (AAF). Subjects read aloud single words containing the vowel /Ɛ/ (e.g., “bet”, “head”) under conditions of normal or altered auditory feedback. The real-time feedback alteration involved a *decrease* in F1 frequency, resulting in a vowel perceived to be closer to /I/ (e.g., “bit”, “hid”). Before each word production, subjects heard a brief phrase spoken with one of three different formant patterns, simulating differences in vocal tract properties of three different talkers (nearly identical to carrier phrases previously shown to induce changes in the perception of an ambiguous vowel between /Ɛ/ and /I/, [5]). We predicted that if the carrier phrase similarly influenced subjects’ perception of their *own* vowel formants during word production, the resulting change would impact the degree of motor adaptation to their F1-altered auditory feedback. Specifically, subjects exposed to the carrier phrase containing relatively *high* formant frequencies should perceive their own vowel as comparatively *lower* in F1 (i.e., closer to /I/), thus enhancing the perceived auditory feedback manipulation and increasing the degree of speech motor adaptation (see Figure 1). Conversely, subjects exposed to the carrier phrase containing relatively *low* formant values should perceive their own vowel as comparatively *higher* in F1 (i.e., closer to /Ɛ/), diminishing the perceived auditory feedback manipulation and thereby reducing the degree of motor adaptation.

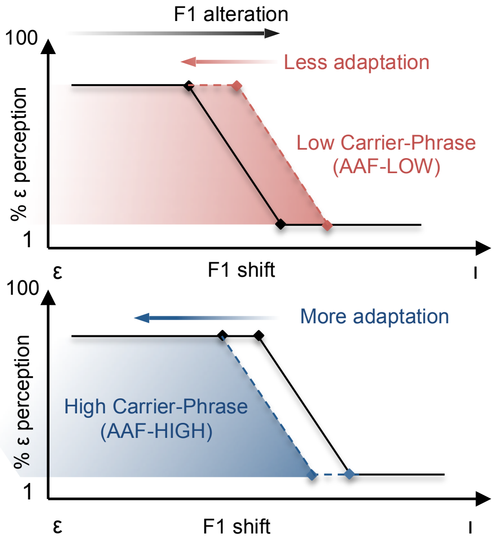

Figure 1 – Expected group differences in motor adaptation to AAF in *Phase 2*. Increased adaptation was expected to occur in the group exposed to the *High Carrier-Phrase* (blue) relative to participants hearing the *Low Carrier-Phrase* (red).

2. THE STUDY

2.1. Materials and methods

2.1.1. Subjects

Twenty male, native speakers of English (18-30 years) without history of speech, language or hearing disorders took part in the study. All subjects passed a pure-tone hearing screening (threshold < 20 dB HL at octave frequencies between 250 and 4000 Hz) and provided written informed consent prior to testing. Procedures were approved by the Institutional Review Board, Faculty of Medicine, McGill University.

2.1.2. Stimuli and group assignment

Participants were randomly assigned to one of two groups (*High Carrier-Phrase* and *Low Carrier-Phrase*, 10 subjects per group) and underwent an identical series of tasks involving the production of mono-syllabic words containing the vowel [Ɛ] (e.g., “bed”, “head”), first under normal-auditory-feedback conditions (NAF) and then during a period of altered auditory feedback (AAF). A real-time acoustic manipulation carried out while subjects produced these words yielded a perception of the vowel [Ɛ] as being closer to [I] (see section 2.1.3 for details). Each word production trial began with the auditory presentation of the carrier sentence “*Please say what this word is…*”, after which a target word appeared on a computer monitor, which the participants read aloud. Subjects listened to their own amplified auditory feedback through headphones. Three different versions of the carrier sentence were used in the experiment (*Neutral*, *Low* and *High*), characterized by the following average F0, F1 and F2 formant frequencies (Table 1).

Table 1 – Formant values of the various versions of the carrier phrases as used in the present study

	F0	F1	F2	F3
NEUTRAL	95.7	435.9	1556.0	2515.3
LOW	98.1	417.7	1345.1	2525.0
HIGH	112.4	455.7	1639.0	2616.9

2.1.3. Speech motor adaptation

Subjects in both groups produced a total of 260 target words chosen at random from a stimulus list of 10 possible /Ɛ/ words[1]. All subjects underwent the following sequence of auditory feedback and carrier-phrase conditions: (1) an initial set of 30 trials under normal auditory feedback and preceded by the *Neutral* carrier sentence (NAF-Neutral); (2) a set of 100 practice trials under conditions of AAF preceded by the High or Low carrier phrase, depending on the group (AAF-High or AAF-Low); (3) a set of 100 words under AAF preceded by the Neutral carrier sentence (AAF-Neutral), and finally (4) a washout phase of 30 trials under NAF with the Neutral carrier sentence (NAF-Neutral). The auditory feedback manipulation corresponded to a 30% decrease in F1 (average shift: 180.2 Hz), inducing the perception of a vowel closer to [I]. The system used to carry out the feedback manipulation combines a digital signal processor (DSP) designed for manipulating vocal signals in near-real-time (VoiceOne, TC Helicon) with low-/high-pass filters to restrict the formant shift to F1. A detailed description of the system has been published previously [8, 13].

2.1.4. Acoustic analysis

For each word produced in the speech adaptation task, a 30 ms segment centred around the midpoint of the vowel was selected. Mean F1 and F2 frequency for each segment was then estimated by LPC analysis in Matlab. LPC parameters were chosen on a per-subject basis to minimize the occurrence of spurious formant values. F0 was also estimated for each vowel centre using an autocorrelation method [14]. Values of F0, F1 and F2 frequency were used to directly compare vowel acoustic properties between conditions (NAF and AAF), and between groups (Low-Sentence and High-Sentence). Vowel acoustic changes during the speech adaptation task were computed as the proportion change in frequency relative to the mean values during the baseline NAF phase (averaged over trials 11-30). Differences in speech adaptation between the two groups were evaluated at two key time points: (1) at the end of the first practice phase under conditions of altered feedback (AAF-High or AAF-Low; averaged over trials 111-130), and (2) at the end of the second phase under altered feedback (AAF-Neutral; averaged over trials 211-230).

2.2. Results

2.2.1. Baseline

In order to ensure that the two groups were comparable in their production of /Ɛ/ during the NAF-Neutral baseline phase, mean F0, F1 and F2 values were compared between groups using independent-samples *t*-tests. No reliable baseline differences were observed between the groups for F0 ($t(18) = 1.63$, $p = 0.12$), F1 ($t(18) = 0.50$, $p = 0.62$), or F2 ($t(18) = 0.80$, $p = 0.43$).

2.2.2. Speech Adaptation

The results of the speech adaptation task for the two groups (*High Carrier-Phrase* and *Low Carrier-Phrase*) are shown in Figure 2 with mean changes in formant values relative to baseline at the three time-points in the testing sequence shown in Figure 3. Overall, a compensatory increase in F1 frequency can be observed in response to the F1 auditory feedback manipulation for both groups. By the end of the first AAF phase, during which the two carrier sentences differed for the two groups, the magnitude of the F1 compensation can be seen to diverge between the two sentence conditions, with the *High Carrier-Phrase* group showing a relatively large F1 change, and the *Low Carrier-Phrase* showing a smaller change. By the end of the second AAF phase, during which both groups were presented with the *Neutral* carrier phrase, the magnitude of the compensatory change can be seen to converge once again.

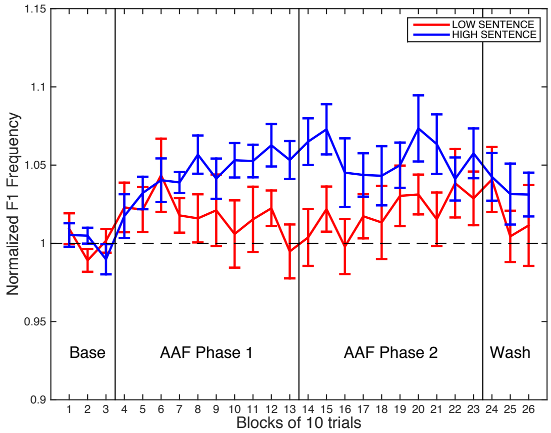

Figure 2 – Top: Mean change in F1 (proportion relative to baseline) is shown for the High-sentence (blue) and Low-sentence (red) groups during the entire speech adaptation task.

The reliability of these effects was evaluated using a 2-way mixed-factorial ANOVA, with GROUP (*High* vs. *Low*) and PHASE (AAF-Phase 1 and AAF-Phase 2) as factors. The main effect of GROUP and PHASE were both not significant (GROUP: $F_{1,18} = 0.13$, $p = 0.13$; PHASE: $F_{1,18} = 0.36$, $p = 0.36$), however the interaction effect was

[1] The set of stimulus words included the following: *Trek*, *Bet*, *Pet*, *Ten*, *Peck*, *Pen*, *Tech*, *Neck*, *Mess*.

reliable ($F_{1,18} = 0.042$, $p < 0.05$). Post-hoc pairwise comparisons were carried out using the Holm-Bonferroni procedure, revealing a reliable difference between groups at the end of AAF-Phase 1 ($p < 0.05$), but not at the end of AAF-Phase 2 ($p = 0.68$). While a between-group effect was noted for F1, no reliable difference was observed between groups for F0 and F2 (Figure 4).

A 2-way mixed-factorial ANOVA showed no significant main or interaction effects for either F0 (GROUP: $F_{1,18} = 0.11$, $p = 0.75$; PHASE: $F_{1,18} = 0.323$, $p = 0.23$; Interaction: $F_{1,18} = 0.18$, $p = 0.68$) or F2 (GROUP: $F_{1,18} = 0.28$, $p = 0.60$; PHASE: $F_{1,18} = 0.90$, $p = 0.35$; Interaction: $F_{1,18} = 0.04$, $p = 0.84$).

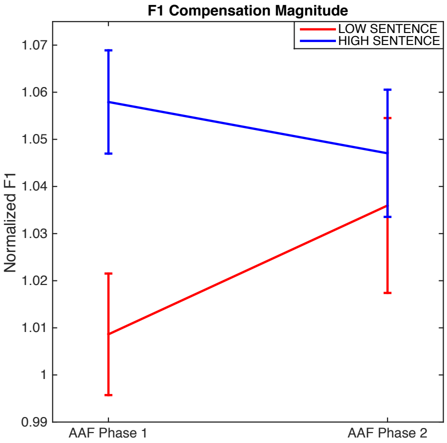

Figure 3 - The mean compensation effect at the end of AAF-Phase 1 and AAF-Phase 2. Consistent with predictions, the High-sentence group exhibited greater compensation than the Low-sentence group during AAF-Phase1. The effect is seen to diminish by the end of AAF-Phase 1.

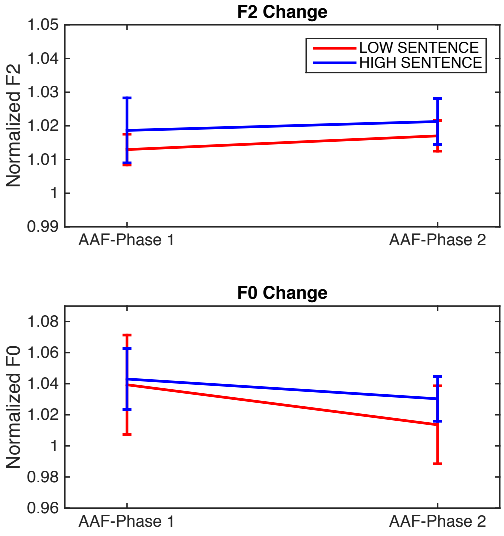

Figure 4 – Mean change in F2 and F0 (proportion relative to baseline) is shown for the High-sentence (blue) and Low-sentence (red) groups at the end of AAF-Phase 1 and AAF-Phase 2. In contrast with F1, no difference between groups was observed for either of these acoustic measures during AAF-Phase1 or AAF-Phase 2.

3. DISCUSSION

The present research sought to expand on the findings of several recent studies demonstrating a surprising degree of plasticity in the processing of auditory feedback during speech production [12, 13, 15]. Here, we tested whether the vowel formants characterizing an introductory carrier-phrase would serve as a perceptual frame of reference for a talker's own speech, thereby influencing their degree of motor adaptation to a real-time alteration of auditory feedback during the production of /Ɛ/-words (lowering F1, resulting in a vowel perceived to be closer to /I/). More specifically, we predicted that a carrier phrase with higher formants would yield a perception of the self-produced vowel as comparatively low in F1 frequency (further toward /I/), thereby increasing the perceived auditory error and enhancing the degree of speech motor compensation. In contrast, a carrier sentence with lower formants was predicted to yield a perception of the self-produced vowel as comparatively high in F1 (closer to /Ɛ/), thereby decreasing the perceived error and diminishing the motor compensatory response. The results were consistent with the predictions, showing a difference in the degree of F1 motor compensation between groups when they were exposed to different carrier-phrases under AAF, and a subsequent convergence in compensation magnitude when both groups were exposed to the same (Neutral) phrase under AAF. Note that the effect of carrier phrase was confined to F1, with no effect of carrier sentence on F2 or F0. This suggests that subjects in the two sentence groups did not simply converge acoustically [11] toward their respective High- or Low- carrier phrases (which varied in a range of spectral properties). Rather, subjects appear to have more specifically interpreted their own vowel acoustic error, which was confined to F1 frequency, within a frame of reference provided by the carrier phrase.

The present finding of speech adaptation to AAF is consistent with the idea that the acoustic correlates of phoneme categories serve as primary targets of speech production. However the present results, along with those of other recent studies, strongly indicate that auditory-sensory feedback processing can be biased or altered by numerous factors, including reinforcement-based perceptual training [12, 15], top-down lexical effects [13], and now context-dependent vowel normalization. An important consideration for future models of speech motor control would be to reconcile such perceptual flexibility with the sensory-dependent feed-forward and feedback mechanisms presumed to drive speech motor adaptation and control [7, 9].

This work was in part funded by the US National Institute of Health (NIDCD – R01DC012502), the Natural Science and Engineering Research Council of Canada and the Fonds de Recherche Québécois, Nature et Technologie.

7. REFERENCES

1. Zion-Golumbic, E.M., D. Poeppel, and C.E. Schroeder, *Temporal context in speech processing and attentional stream selection: A behavioral and neural perspective.* Brain and Language, 2012. **122**: p. 151-161.
2. Peterson, G.E. & H. Barney. (1952). Control Methods Used in a Study of the Vowels. *The Journal of the Acoustical Society of America*. **24**(2): 175-184.
3. Johnson, K. (2008). *Speaker normalization in speech perception*, in D.B. Pisoni & R. Remez (Eds). John Wiley & Sons: Malden.
4. Ladefoged, P. (1989). A note on "Information conveyed by vowels". *Journal of the Acoustical Society of America* **85**(5): 2223-2224.
5. Ladefoged, P. & D. Broadbent. (1957). Information Conveyed by Vowels. *The Journal of the Acoustical Society of America* **29**(1): 98-104.
6. Dechovitz, D. (1977). Information conveyed by vowels: A confirmation. *Haskins Laboratory Status Report on Speech Research* SR-53-54: 213-219.
7. Houde, J. & M.I. Jordan. (1998). Sensorimotor Adaptation in Speech Production*.* *Science* 279: 1213-1216.
8. Mollaei, F., D.M. Shiller & V.L. Gracco. Sensorimotor adaptation of speech in Parkinson's disease. *Movement Disorders* 28(12): 1668-1674.
9. Tourville, J.A. & F.H. Guenther. (2011). The DIVA model: A neural theory of speech acquisition and production*.* *Language and Cognitive Processes* 26(7): 952-981.
10. Houde, J.F. & S.S. Nagarajan. (2011). Speech production as state feedback control*.* *Frontiers in Human Neuroscience* **5**(82).
11. Sato, M., K. Grabski, M. Garnier, L. Granjon, J-L. Schwartz, N. Nguyen. (2013). Converging toward a common speech code: imitative and perceptuo-motor recalibration processes in speech production*.* *Frontiers in Psychology* **4**.
12. Lametti, D.R., S.A. Kroll, D.M. Shiller & D.J. Ostry. (2014). Brief Periods of Auditory Perceptual Training Can Determine the Sensory Targets of Speech Motor Learning. *Psychological Science* 25(7): 1325-1336.
13. Bourguignon, N.J., S.M. Baum & D.M. Shiller. (2014). Lexical-perceptual integration influences sensorimotor adaptation in speech. *Frontiers in Human Neuroscience* 8.

14 Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of sampled sound. Institute of Phonetic Sciences, University of Amsterdam 17: 97-110.

15. Schiller, D.M. & M-L. Rochon. (2014). Auditory-Perceptual Learning Improves Speech Motor Adaptation in Children. *Journal of Experimental Psychology* 40(4): 1308-1315.
16. Villacorta, V.M., J. S. Perkell & F.H. Guenther. (2007). Sensorimotor adaptation to feedback perturbations to vowel acoustics and its relation to perception. *Journal of the Acoustical Society of America* 122(4): 2306-2319.