5-4-2014

Sensorimotor Adaptation of Speech Using Real-time Articulatory Resynthesis

Jeffrey J. Berry
Marquette University, jeffrey.berry@marquette.edu

Cassandra North
Marquette University

Michael T. Johnson
Marquette University, michael.johnson@marquette.edu

Accepted version. Published as part of the proceedings of the conference, *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014: 3196-3200. DOI.

Sensorimotor Adaptation of Speech Using Real-time Articulatory Resynthesis

Jeff Berry
Speech Pathology & Audiology, Marquette University, Milwaukee, WI
Cassandra North
Electrical & Computer Engineering, Marquette University, Milwaukee, WI
Michael T. Johnson
Electrical & Computer Engineering, Marquette University, Milwaukee, WI

Abstract: Sensorimotor adaptation is an important focus in the study of motor learning for non-disordered speech, but has yet to be studied substantially for speech rehabilitation. Speech adaptation is typically elicited experimentally using LPC resynthesis to modify the sounds that a speaker hears himself producing. This method requires that the participant be able to produce a robust speech-acoustic signal and is therefore not well-suited for talkers with dysarthria. We have developed a novel technique using electromagnetic articulography (EMA) to drive an articulatory synthesizer. The acoustic output of the articulatory synthesizer can be perturbed experimentally to study auditory feedback effects on sensorimotor learning. This work aims to compare sensorimotor adaptation effects using our articulatory resynthesis method with effects from an established, acoustic-only method. Results suggest that the articulatory resynthesis method can elicit speech adaptation, but that the articulatory effects of the two methods differ.

SECTION 1.

INTRODUCTION

2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (May 4-9, 2014): 3196-3200. DOI.

Sensorimotor adaptation is an important focus in the study of speech, but has yet to be studied for rehabilitation applications for individuals with dysarthrias (motor speech disorders).[1,2] Speech adaptation is a form of involuntary sensorimotor learning that can be elicited experimentally using acoustic signal processing techniques to perturb the sounds that a speaker hears himself producing. Novel sensorimotor learning takes the form of involuntary, compensatory changes in speech-articulatory movements. Established techniques for making perturbations to the speech signal in order to manipulate auditory feedback and elicit sensorimotor adaptations are not effective for talkers with dysarthria because participants are required to produce acoustically high-quality speech. Finding a viable method for eliciting sensorimotor adaptation from individuals with dysarthria will support the development of novel approaches to speech rehabilitation following stroke, traumatic brain injury, or other neurological impairments.

We have developed a novel technique for eliciting speech adaptation that uses articulatory resynthesis. Because it does not depend on a high-quality speech signal, this technique is viable even with individuals with severe dysarthria.[3] An EMA system is used to drive an articulatory speech synthesizer. The acoustic output of the synthesizer is perturbed using an established, acoustic-based method.[4] The perturbed, resynthesized speech is sent back to the participant via headphones to provide auditory feedback based on the talker's articulatory movements. Graded acoustic perturbations to the synthesized speech can be used to elicit involuntary changes in articulation, characteristic of sensorimotor adaptation effects elicited using established, acoustic-only methods that directly perturb a talker's speech.[5] The current work aims to compare and contrast speech adaptation effects obtained using our novel articulatory resynthesis technique with an established method that acts directly on the talker's own voiced speech.

SECTION 2.

METHODS

The NDI Wave EMA system was used to register participants' tongue, lip, and jaw movements.[6] Five (5 degree-of-freedom) sensors were attached along the midsagittal plane (two on the dorsal surface of the tongue, one on each lip, and one at the juncture of the central mandibular incisors near the gingival border). A single six degree-of-freedom reference sensor (attached to the bridge of a pair of plastic glasses frames worn by the subject) was used to correct for head movements.

2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (May 4-9, 2014): 3196-3200. DOI.

Five typically-functioning young adults (3 male and 2 female) participated in two (randomized) experimental runs using two different elicitation methods (see Figure 1). Method A used the participant's own voiced speech (transduced via medium diaphragm condenser microphone and processed through the *Audapt* acoustic perturbation software) for auditory feedback.[4] Method B used the acoustic signal from an articulatory synthesizer (driven by the talker's articulatory movements and processed through the acoustic perturbation software) for auditory feedback.

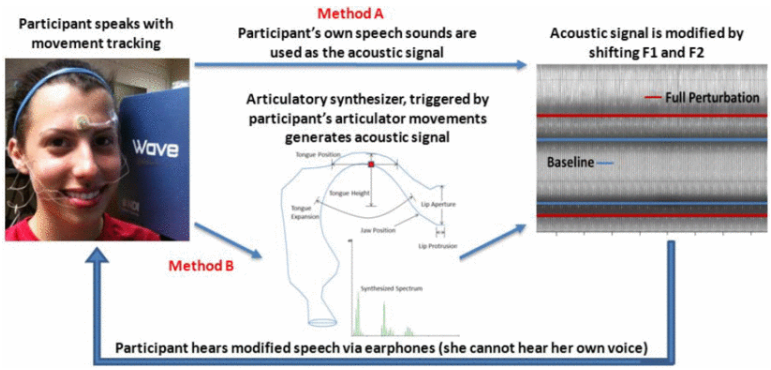

Figure 1: Schematic of experimental setup for two methods

During the Method B experiments, participants were asked to move the articulators as they would during speech, but to refrain from vocalization. Articulator movements were transformed into control parameters for a software articulatory speech synthesizer[7,8] using a simple, piece-wise linear mapping that defined absolute boundaries of the articulatory working space, and positions associated with the corner vowels. This mapping approach was speaker-independent and a common vocal tract model was used across participants, though speaker-specific calibrations were necessary.[5] Calibration data were obtained from each participant from brief kinematic records characterizing articulatory position extrema (i.e., "open your mouth as wide as you can") and articulatory positions during sustained corner vowels ([i], [a], [u], [ae]). Articulatory position extrema were mapped onto articulatory synthesis parameter extrema and sustained vowel positions were mapped onto ideal parameter settings for the corner vowels, derived from Maeda.[7] The mapping method resulted in some normalization of the acoustic working space across participants and a common voice quality (source parameters of the synthesizer were unchanged). Since participants heard no natural pitch variation in the Method B condition, all auditory feedback presented during Method A was also processed to eliminate fundamental frequency variations.

2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (May 4-9, 2014): 3196-3200. DOI.

For both experimental methods, participants were seated in a sound booth, with EMA sensors attached to the articulators. Stimuli were displayed on a computer screen in large font and participants were simply asked to produce each vowel sound when it appeared on the screen. The acoustic output of the microphone (Method A) or the articulatory synthesizer (Method B) was perturbed with the software *Audapt* used in previous speech adaptation experiments.[4] These perturbations were designed to elicit involuntary changes in participant articulator movements.[9] Each experimental run was divided into five contiguous phases (see Figure 2): 1) baseline (80 total tokens alternating [e] and [0]); 2) ramp (40 [e] tokens with progressively increasing formant perturbation magnitude); 3) full perturbation (40 [e] tokens at the maximum formant perturbation); 4) masking (40 tokens alternating [e] and [0] with pink noise masking auditory feedback); 6) return (20 tokens alternating [e] and [0] with auditory feedback returned to baseline condition). The current work focused on adaptation of isolated productions of the vowels [e] and [0].

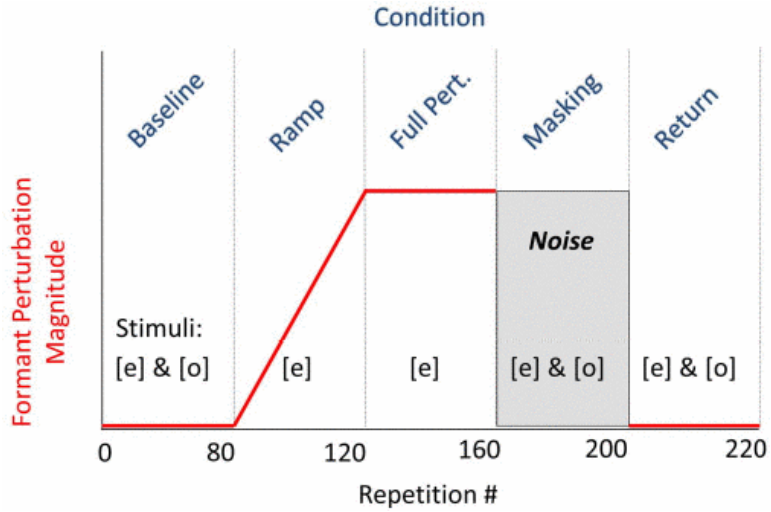

Figure 2: Schematic of adaptation experiment

Adaptation was trained on the vowel [e] by gradually shifting the first formant down in frequency and the second formant up in frequency. This acoustic shift supports the perception of increasing vowel height, resulting in an [e] (off-glide) that sounds progressively more like the vowel [i]. Adaptation was hypothesized to be compensatory, resulting in an involuntarily lower and more posterior tongue position.[9,10] This hypothesis is consistent with compensation for the perception of increasing vowel height and advancement. The five participants were engaged in two experimental runs of the speech adaptation protocol, differing only to the extent that the articulatory synthesizer was used to supplant the subject's speech (Method A versus Method B).

2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (May 4-9, 2014): 3196-3200. DOI.

SECTION 3.

RESULTS

Figure 3 shows acoustic data obtained from one participant using both methods. Formant frequency values were obtained using pitch-synchronous LPC (26 coefficients) via the TF32 acoustic analysis software.[11] F1 and F2 values were measured at the time of the peak F2 (typically the off-glide steady-state of [e]). The axes of the data in Figure 3 are oriented to roughly approximate the articulatory working space. Thus, increasing F2 values along the positive x-axis approximate more forward tongue positions and decreasing F1 values along the (inverted) positive y-axis approximate higher tongue positions. Acoustic data are coded by experimental phase: black circles indicate baseline performance; red triangles indicate performance during full perturbation ([e] maximally shifted to sound like [i]), green squares indicate performance during masking (auditory feedback eliminated with noise); and yellow diamonds indicate performance when feedback is returned to the baseline state. Sensorimotor adaptation is characterized by two component changes: 1) compensation - measured by the articulatory change from baseline to full perturbation; and 2) adaptation - measured by the change from baseline to masking. The hypothesized effect would be reflected in red triangles and green squares that tend to fall below and to the left of black circles (roughly corresponding to a lower and more posterior tongue position compared to baseline). Based on Figure 3, both methods appear to elicit some tendency toward this effect, though the data are quite variable.

Figure 4 summarizes the average F1 and F2 values associated with articulation during each experimental phase for each participant using both elicitation methods. While the hypothesized effect predicted compensatory decreases in F2 and increases in F1 (roughly equivalent to lower and more posterior tongue positions during full perturbation and masking compared to baseline) it is apparent that participant responses were idiosyncratic for Method A. Moreover, a substantial variability across subjects in the size of the acoustic working space is evident. By contrast, the results for Method B indicate greater effect homogeneity across participants, with all subjects demonstrating the predicted effect of reduced F1 and increased F2 in compensation for the perceived perturbation of [e] toward [i]. Moreover, the acoustic working space of the vowels produced by the participants is substantially normalized compared to the data acquired for Method A. This benefit of our novel method is demonstrated by the reduced scaling of both F1 and F2 axes as well as the greater overlap of average data points across speakers.

2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (May 4-9, 2014): 3196-3200. DOI.

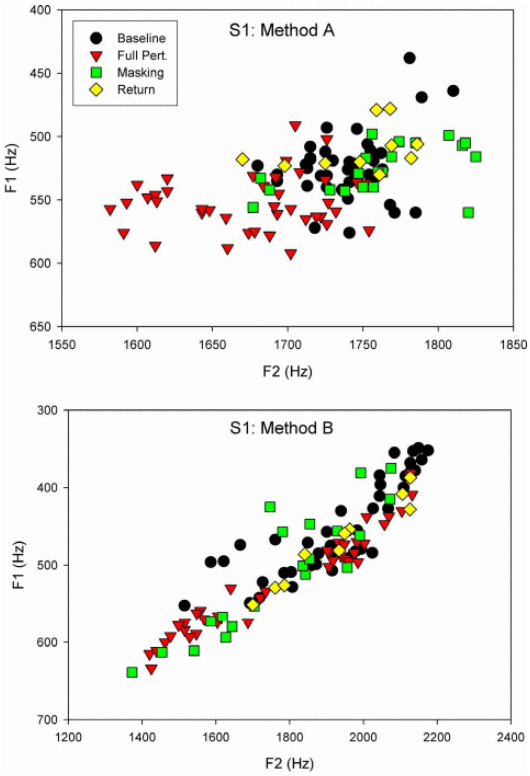

Figure 3: F1-F2 values from subject 1 showing articulatory changes between experimental phases for both methods

Table 1 summarizes the acoustic changes in articulation exhibited by subjects from baseline to full perturbation (compensation) and baseline to masking (adaptation) using the two different elicitation methods. While both methods demonstrate some capacity to elicit compensation and adaptation, neither method works consistently with all subjects. The direction of the effects elicited using Method A vary across subjects. In particular, while it was hypothesized that participants would increase F1 and decrease F2 in response to auditory feedback perturbations, the majority of the significant formant changes elicited using Method A do not follow the direction of the hypothesis. In contrast, the effects elicited using Method B all follow the hypothesized direction (all F1 changes are positive and all F2 changes are negative). Moreover, within subject, all significant changes are comparable across conditions. For example, using Method B, subject 5 compensates by raising F1 and adapts by raising F1. When using Method A, the same subject compensates by lowering F1, but then adapts by raising F1.

2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (May 4-9, 2014): 3196-3200. DOI.

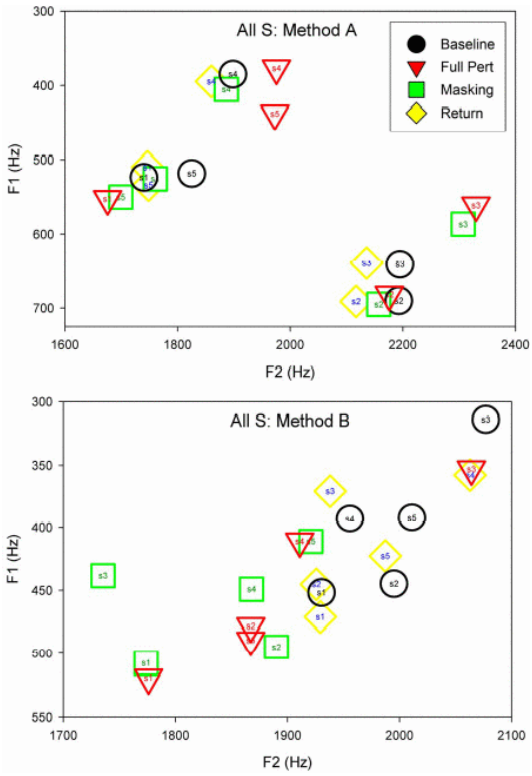

Figure 4: Average F1-F2 values per experimental phase by subject and method

Table 1. Summary of F1 & F2 shifts (Hz) from baseline to full perturbation (compensation) and baseline to masking (adaptation). Asterisks indicate significant effects ($p<0.05$)

Method A

	Compensation		Adaptation	
	F1	**F2**	**F1**	**F2**
S1	29*	-64*	2	20
S2	-9*	-17	5	-36
S3	-80*	135	-54*	112
S4	-8*	77*	20	-12
S5	-81*	148*	32*	-126*

Method B

	Compensation		Adaptation	
	F1	**F2**	**F1**	**F2**
S1	68*	-154*	56*	-157*
S2	33*	-128*	50*	-105*
S3	39	-13	125	-341*
S4	18	-45	56	-88
S5	98*	-144*	19*	-90*

2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (May 4-9, 2014): 3196-3200. DOI.

SECTION 4.

CONCLUSIONS

The purpose of the current work is to evaluate a novel method for eliciting sensorimotor adaptation of speech-articulatory movements. Sensorimotor adaptation experiments were completed using an established method whereby a talker's voiced speech is acoustically resynthesized[4] and our novel articulatory resynthesis method [5]. Both methods were able to elicit involuntary changes in participant articulation during vowel production. While subject responses where idiosyncratic and variable using both elicitation methods, our novel method elicited the most consistent patterns of articulatory change (measured acoustically) across subjects. Additionally, our novel articulatory resynthesis method provided substantial normalization of the acoustic working space across subjects. Given the wide idiosyncrasy of the articulatory-acoustic working spaces across talkers, it is likely that the extent of difference in both the acoustic working space and the specifics of the articulatory-to-acoustic mapping will affect the degree to which the effects of our articulatory resynthesis method will generate effects comparable to the established method.

A primary value of our novel method is that it can used with talkers with dysarthria who cannot produce a robust speech-acoustic signal. This capacity will support the exploration of rehabilitation applications. Further work will address the potential value of scaling the acoustic working space of the articulatory synthesizer to improve the generalizability of elicited effects to real speech, as well as implementation of this novel articulatory resynthesis method for talkers with dysarthria. The novel use of speech technology in helping to address clinical speech problems will be key to advancing speech neurorehabilitation. The current work advances a novel approach for modifying articulatory behavior that may have direct application for rehabilitating the speech of individuals with dysarthria.

Acknowledgements

Funding for this work was provided by the American Speech-Language-Hearing Foundation New Century Scholar Research Grant and the National Science Foundation NSF IIS-1320892. The authors would like to thank Mark Huckvale, Shanqing Cai, and Kevin Reilly for their support.

References

2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (May 4-9, 2014): 3196-3200. DOI.

[1]J.F. Houde and M.I. Jordan, "Sensory adaptation of speech I: compensation and adapation," Journal of Speech, Language, and Hearing Research, vol. 45, pp. 295-310, 2002.

[2]V.M. Villacorta, J.S. Perkell, and F.H. Guenther, "Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception," Journal of the Acoustical Society of America, vol. 122, pp. 2306-2319, 2007.

[3]J. Berry, C. Bechel, C. North, and M.T. Johnson, "Learning novel articulatory-acoustic mappings in dysarthria," presented at the Conference of the American Speech-Language-Hearing Association, Chicago, Illinois, 2013.

[4]S. Cai, S.S. Ghosh, F.H. Guenther, and J.S. Perkell, "Adaptive auditory feedback control of the production of the formant trajectories in the Mandarin triphthong /iau/ and its patterns of generalization," Journal of the Acoustical Society of America, vol. 128, pp. 2033-2048, 2010.

[5]J. Berry, C. North, B. Meyers, and M.T. Johnson, "Sensorimotor learning through a virtual vocal tract," Proceedings of Meetings on Acoustics, vol. 19, 060099, pp. 1-8, 2013.

[6]J. Berry, "Accuracy of the NDI wave speech research system," Journal of Speech-Language-Hearing Research, vol. 54, pp. 1295-1301, 2011.

[7]S. Maeda, "Compensatory Articulation during Speech: Evidence from the Analysis and Synthesis of Vocal-Tract Shapes using an Articulatory Model," in Speech Production and Modelling, pp. 131-149. W.J.Hardcastle & A. Marchal (Eds.), Academic Publishers, Kluwer, 1989.

[8]M. Huckvale, "VTDemo-Vocal Tract Acoustics Demonstrator," [Computer Program] University College London, 2009.

[9]K.G. Munhall, E.N. MacDonald, S.K. Byrne, and I. Johnsrude, "Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate," Journal of the Acoustical Society of America, vol. 125, pp. 384-390, 2009.

[10]D.W. Purcell and K.G. Munhall, "Compensation following real-time manipulation of formants in isolated vowels," Journal of the Acoustical Society of America, vol. 119, pp. 2288-2297, 2006.

[11]P. Milenkovic, "TF32 Speech Analysis Software," [Computer Program] University of Wisconsin-Madison, 1998.

2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (May 4-9, 2014): 3196-3200. DOI.