Adaptation of trajectory shapes during conversation

Mark Tiede[1], Christine Mooshammer[2,1], Dolly Goldenberg[3,1], Douglas N. Honorof[1]

[1]*Haskins Laboratories, New Haven, USA*
[2]*Humboldt-Universität, Berlin, Germany*
[3]*Yale University, New Haven, USA*

tiede@haskins.yale.edu, mooshamc@hu-berlin.de, dolly.goldenberg@yale.edu, honorof@haskins.yale.edu

Abstract

Speech audio and articulatory movements of age- and gender-matched speaker pairs have been recorded in face-to-face spontaneous conversation using two electromagnetic articulometer (EMA) systems simultaneously. Elicited tasks included synchronized, imitated and spontaneous speech interspersed with repeated baseline utterances for evaluation of mutual accommodation. Convergence measures have been obtained using between-speaker (for magnitude) and within-speaker (for symmetry) distances computed on coda velar gestures and the preceding vowel from baseline tasks produced before and after conversational interaction. The linearly normalized velar trajectories have also been compared for evidence of increased similarity in shape. Results are equivocal, indicating that in some cases convergence can persist as an aftereffect, but more generally supporting the view that it is active interaction that drives mutual accommodation, and without it the effect collapses, leaving production to drift.

Keywords: phonetic convergence, speech kinematics, EMA

1. Introduction

When two speakers interact in a conversation their patterns of speech may gradually become more similar. This adaptation effect, known as convergence, has been observed on many linguistic levels (Pickering & Garrod, 2004). To date most evidence for convergence on the phonetic level has been acquired from acoustic studies and perceptual similarity ratings (e.g. Pardo 2006). Recently we have begun a project to collect kinematic data from interacting speaker pairs to investigate whether convergence can also be observed in the underlying articulation.

Although interaction between two speakers may fail to produce convergence, operate asymmetrically, or even result in diverging speech patterns, in general it has been observed that convergence is facilitated when two individuals are aligned along such parameters as age, gender, and social background (Babel 2012). It is also known that convergence is enhanced by direct access to visual cues provided by the face (e.g. Hazan et al. 2005). Our view is that if convergence does occur it is driven by production-perception links that organize a loose coupling between two interacting systems (Beek et al. 1992). This entrainment serves as a forcing function on internal dynamics, leading to adjustments in existing patterns of behavior within their intrinsic range and resulting in relative coordination between the two speakers (as contrasted with absolute coordination between physically coupled systems). Convergence observed in the acoustic signal is thus a surface manifestation of underlying alignment in articulation.

In the current study we are interested in whether adaptation can be found on purely motoric levels, as in the tongue dorsum movements associated with velar stops. It has been previously observed that following back vowels the tongue dorsum slides forward along the palate during closure (Mooshammer et al. 1995). These so-called forward loops have been found in many languages and show speaker-specific patterns in the extent of the forward movement. Conversely, following front vowels the tongue dorsum does not typically slide during velar closure. Listeners are aware of the naturally occurring shape variation of the trajectory and rate voiced velar stops in back vowel context as more natural if they are produced with a forward loop compared to a straight movement or a backward loop (cf. Nam et al. 2013). In this work we compare the gestural characteristics and shape of velar loops in various contexts obtained from interacting speaker pairs pre- and post-adaptation to a series of conversational tasks. We expect the effects to be most pronounced in those cases where changes towards increased similarity are also observed within such acoustic parameters as vowel quality.

2. Method

2.1. Participants

Ten native speakers of American English were matched for age, gender, social background, personality (assessed from responses to a questionnaire), and dialect. Each passed a 20 dB pure-tone hearing test and had no self-reported speech deficits. One male and one female pair are presented here.

2.2. Experimental tasks

In each experiment participants performed a range of tasks that included synchronized (choral production), imitated and spontaneous speech, interspersed with repeated baseline utterances used to evaluate mutual accommodation. These baseline tasks were produced separately by each participant and consisted of focus words (Table 1) produced within a consistent context (“Say ___ it again”).

Table 1: *Focus words produced in baseline tasks.*

bag	pad	pack	
beg	peg	peck	deck
big	pig	pick	Dick
bog	pod	pock	dock

Within the baseline tasks each word was elicited three times in randomized order.

2.3. Experimental procedures

To record speech articulator movements for two interacting speakers simultaneously we rely on separate electromagnetic articulometer (EMA) devices, one for each talker. Two types of commercially-manufactured EMA systems are used together. The AG500 (Carstens Medizinelektronik, GmbH) has six narrowly tuned transmitters that operate continuously at different frequencies (7.5kHz – 13.75kHz). The WAVE (Northern Digital, Inc.) uses eight strobed transmitters, all operating at 3kHz. Both systems resolve three spatial and two angular orientation measurements per sensor at sampling rates of at least 100 Hz. Crucially both systems permit unrestricted head movement and provide an unimpeded view of the face.

In pilot work to validate the assumption that the different operating frequencies of the AG500 and WAVE devices would support simultaneous operation, a series of benchmark tests were performed. With the measurement centers of each system positioned 2 meters apart, the stability of fixed distances between sensors attached to a rotating rigid body within the field of each device was assessed, with and without the other device in active operation. Results based on rotational symmetry and the standard deviations between pairs of fixed sensors showed no significant effect of dual operation on either system (Tiede et al. 2012).

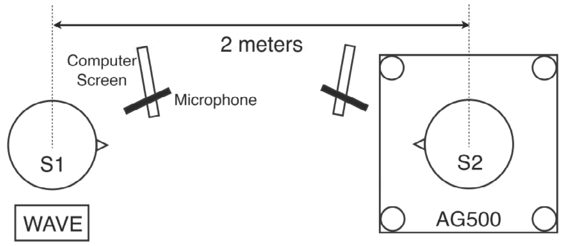

Figure 1: Experimental setup showing dual-EMA arrangement. Separate stimulus display monitors and directional microphones were positioned for each speaker.

For each participant sensors were glued to two points on the tongue (blade and dorsum), the lower incisors, and the upper and lower lips. Additional sensors placed on the upper incisors and left and right mastoid processes were used to correct for head movement. Independent audio tracks were recorded at 44.1 kHz using separate directional microphones located about 50 cm from the mouth. Participants were seated such that their anterior vocal tracts were centered within the respective device fields, for a face-to-face distance of slightly less than 2 meters, and with a clear view of their partner's face. All tasks were presented on separate monitors specific to each speaker. Stimulus presentation and data acquisition were coordinated by custom software (*Marta*, Haskins Laboratories). Figure 1 shows the experimental arrangement.

Data for each speaker were remapped to a movement-corrected standard orientation aligned with the occlusal plane established with a biteplane trial. To reduce noise, reference sensor trajectories were low-pass filtered at 5 Hz and movement sensor trajectories at 20 Hz. Alignment between the two EMA data streams was effected through cross-correlation of their respective audio.

2.4. Measurements

F0 and formant values were computed at the midpoint of each focus word vowel. Tongue dorsum movements for the velar gestures of the focus words were labeled using *Mview* (Haskins Laboratories). The /g/ of "again" in the context sentence was also labeled. A 20% thresholding criterion applied to local peak sensor velocity was used to identify the closing gesture onset (GONS) and opening offset (GOFFS), together with closing and opening peak velocities and the point of maximum constriction. Gestural 'stiffness' (STIFF$_{close}$, STIFF$_{open}$) was computed as the peak velocity divided by the sensor distance traveled during the closing and opening movements. Normalized trajectories for morphological comparisons were obtained using linear interpolation to a standard number of samples over the GONS : GOFFS range.

Phonetic convergence has been quantified in three ways, using in each case corresponding focus words from the initial (PRE) and final (POST) series of baseline tasks. The first measure, called BETWEEN, is the average between-speaker Euclidean distance between each pair's matched values contrasting the initial and final distances; it shows the extent of any convergence (speakers who converge will show smaller BETWEEN values for the POST task). Distances are computed on [F1xF2] pairings in mel space for formant measures, and as absolute differences for monodimensional values.

The second measure, called WITHIN, is the within-speaker Euclidean distance between initial and final matched values for each speaker. It is used to assess the asymmetry of convergence: speakers who accommodate to one another symmetrically will show approximately equal WITHIN values.

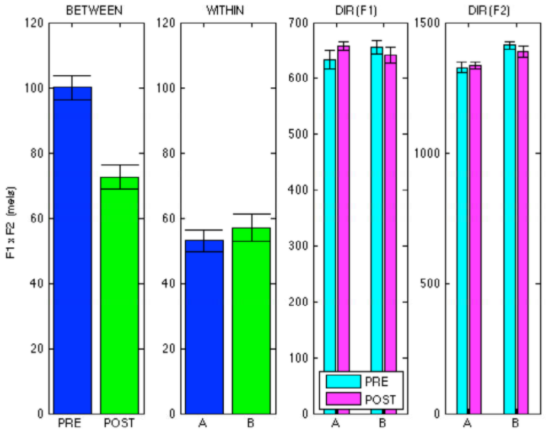

Figure 2: F1xF2 measures for focus words with /ɛ/ (male pair C01). Note the PRE/POST convergence BETWEEN speakers, while similar WITHIN speaker differences show convergence in this instance is effectively symmetric.

The third measure, called DIRection, contrasts the averaged initial and final values of the measured parameter separately by speaker. It is used to determine the direction of any change, in particular whether such change was in the (converging) direction of the partner's values.

Differences between velar loops are computed on the normalized trajectories by summing the Euclidean distances between corresponding samples, with comparisons made BETWEEN and WITHIN speakers as described above.

3. Results

Despite the instance of vowel-specific convergence shown in Figure 2, results obtained for male pair C01 in general support an asymmetric pattern for the acoustic measures (i.e., larger changes in WITHIN values for the second member of the pair) and noisy results for the kinematics:

```
F0            PRE    POST
  BETWEEN:    7.5     6.9   t(10170) =   4.8 *
   WITHIN:    6.9     8.3   t(10130) = -12.2 *
FMTS
  BETWEEN:  114.1   108.5      t(52) =   0.2
   WITHIN:   37.0   108.8      t(52) =  -2.6 *
STIFF_close
  BETWEEN:    0.07    0.08     t(38) =  -0.9
   WITHIN:    0.02    0.01     t(38) =   1.1
STIFF_open
  BETWEEN:    0.07    0.06     t(43) =   0.8
   WITHIN:    0.04    0.03     t(43) =   0.9
```

The normalized trajectories for the male pair show a pattern of divergence, both in the summed distances (Table 2) and in the shapes of the loops (Figure 3).

Table 2: *Summed Euclidean distances between corresponding samples of normalized trajectories (male pair C01)*

	BETWEEN		**WITHIN**	
	pre	**post**	**pre**	**post**
/a/	199.4	410.0	209.2	320.8
/æ/	232.8	473.1	421.6	267.8
/ɪ/	330.7	273.1	210.7	289.3
/ɛ/	303.7	346.4	335.6	356.8
'again'	342.1	514.3	281.1	201.8

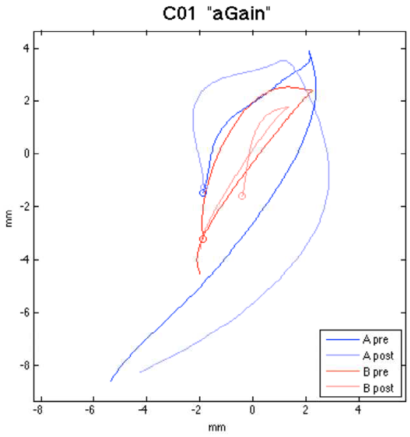

Figure 3: Mean normalized trajectories (male pair C01). "o" marks initial sample.

The female pair (C09) shows some evidence for acoustic convergence, but it is asymmetric with the second member of the pair adapting more. The gestural stiffness measures show divergence:

```
F0            PRE    POST
  BETWEEN:  21.0    18.9  t(11412) =   7.8 *
   WITHIN:  15.3    18.2  t(11376) = -12.1 *
FMTS
  BETWEEN: 397.6   256.1     t(52) =   1.4
   WITHIN: 248.9   343.1     t(52) =  -0.9
STIFF_close
  BETWEEN:   0.02    0.06    t(42) =  -5.9 *
   WITHIN:   0.04    0.02    t(42) =   2.3 *
STIFF_open
  BETWEEN:   0.02    0.07    t(42) =  -4.1 *
   WITHIN:   0.05    0.03    t(42) =   1.1
```

However, normalized trajectories show convergence not only on the summed distance measures (Table 3), but in increased similarity of loop shape as well (Figure 4).

4. Discussion and conclusion

The preliminary results presented here are useful in illustrating methods for quantifying kinematic convergence while at the same time underscoring their limitations. In particular, the lack of a consistent response across acoustic and kinematic measures calls into question the sensitivity of the PRE vs. POST tasks for detecting possibly subtle convergence effects. The underlying assumption is that once shifted within their respective production ranges speakers would stay shifted as a form of production aftereffect, and indeed the female pair does show some evidence for a persistent shift towards greater shape similarity in the velar trajectories (Figure 4).

However, because participants perform these baseline tasks reading from a screen and without interaction with their partner it now seems more likely to us that in the POST phase the coupling driving the effect collapses, leaving production to drift. To test this idea we are currently investigating whether more systematic effects can be observed on the same focus words embedded in the speech tasks during which partners were actively engaged with one another.

Table 3: *Summed Euclidean distances between corresponding samples of normalized trajectories (female pair C09)*

	BETWEEN		**WITHIN**	
	pre	**post**	**pre**	**post**
/a/	797.7	640.1	506.1	496.2
/æ/	675.7	444.6	350.9	354.3
/ɪ/	591.5	341.6	431.1	294.2
/ɛ/	838.2	502.2	427.9	478.6
'again'	365.9	132.1	195.7	225.4

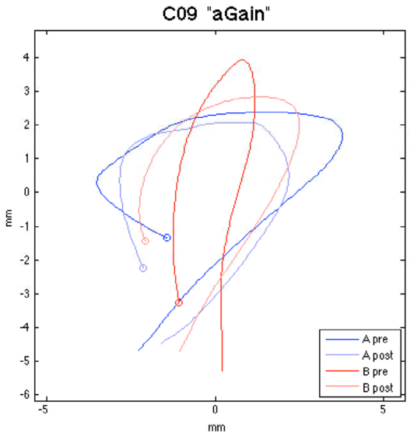

Figure 4: Mean normalized trajectories (female pair C09). "o" marks initial sample.

5. Acknowledgements

This work is supported by a grant from the U.S. National Institutes of Health (NIDCD R21-DC012350).

6. References

Babel, M. (2012). “Evidence for phonetic and social selectivity in spontaneous phonetic imitation”. *Journal of Phonetics*, 40, 177-189.

Beek, P., Turvey, M. & Schmidt, R. (1992). “Autonomous and nonautonomous dynamics of coordinated rhythmic movements”. *Ecological Psychology*, 4, 65-95.

Hazan, V., Sennema, A., Iba, M. & Faulkner, A. (2005). “Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English”. *Speech Communication*, 47, 360-378.

Mooshammer, C., Hoole, P. & Kühnert, B. (1995). “On loops”. *Journal of Phonetics,* 23, 3-21.

Nam, H., Mooshammer, C., Iskarous, K., & Whalen, D. (2013). “Hearing tongue loops: Perceptual sensitivity to acoustic signatures of articulatory dynamics.” *Journal of the Acoustical Society of America*, 134, 3808-3817.

Pardo, J. (2006). “On phonetic convergence during conversational interaction”. *Journal of the Acoustical Society of America,* 119, 2382-2293.

Pickering, M. & Garrod, S. (2004). “Toward a mechanistic psychology of dialogue”. *Behavioral and Brain Sciences,* 27, 169-190.

Tiede, M., Bundgaard-Nielsen, R., Kroos, C., Gibert, G., Attina, V., Kasisopa, B., Vatikiotis-Bateson, E. & Best, C. (2012). “Speech articulator movements recorded from facing talkers using two electromagnetic articulometer systems simultaneously”. *Proceedings of Meetings on Acoustics*, 11, 1-9.