Proceedings of Meetings on Acoustics

Volume 11, 2010 http://acousticalsociety.org/

160th Meeting
Acoustical Society of America
Cancun, Mexico
15 - 19 November 2010
Session 4pSC: Speech Communication

4pSC10. Speech articulator movements recorded from facing talkers using two electromagnetic articulometer systems simultaneously

Mark Tiede*, Rikke Bundgaard-Nielsen, Christian Kroos, Guillaume Gibert, Virginie Attina, Benjawan Kasisopa, Eric Vatikiotis-Bateson and Catherine Best

***Corresponding author's address: Haskins Laboratories, 300 George Street, New Haven, CT 06511, tiede@haskins.yale.edu**

Two 3D electromagnetic articulometer (EMA) systems, the Carstens AG500 and Northern Digital WAVE, have been used simultaneously without mutual interference to record the speech articulator movements of two talkers facing one another two meters apart. A series of benchmark tests evaluating the stability of fixed distances between sensors attached to a rotating rigid body was first conducted to determine whether the two systems could operate independently, with results showing no significant effect of dual operation on either system. In the experiment proper, two native speakers of American English participated as subjects. Sensors were glued to three points on the tongue, the upper and lower incisors, lips, and left and right mastoid processes for each subject. Independent audio tracks were recorded using separate directional microphones, which were used to align the kinematic data from both subjects during post-processing. Data collected were of two types: extended spontaneous conversation, and repeated incongruent word sequences (e.g., talker one produced "cop top..."; talker two "top cop..."). Both talkers show strong positive correlations between speech rate (in syllables/sec) and head movement. The word sequences also show error and rate effects related to mutual entrainment. [Supported by ARC Human Communication Science Network (RN0460284), MARCS Auditory Laboratories, NIH]

Published by the Acoustical Society of America through the American Institute of Physics

 [DOI: 10.1121/1.4772388]
Received 14 Nov 2012; published 30 Nov 2012

Introduction

Due to the inherent difficulties associated with direct observation of the speech articulators, existing data tracking their movements have to date been obtained almost exclusively from 'laboratory' (typically read) speech elicited from single speakers. But speech is at its most natural in face-to-face conversational settings, in which prosody, turn-taking and other aspects of spontaneous production are best observed. Pioneering a logical extension of previous acoustic studies of such dyadic speaker interactions, we have in this work for the first time recorded the vocal tract kinematics of face-to-face speakers engaged in conversation using two electromagnetic articulometer (EMA) devices.

Recent unpublished work conducted at the Edinburgh Speech Production Facility (described in Turk et al., 2010) has used EMA systems to observe the speech articulators of paired speakers engaged in goal-directed conversation. However, because the EMA devices that they use are of the same type, conversational partners must be recorded in separate rooms, without visual contact, in order to avoid device interference. But several studies have shown the importance of visual cues in speech; e.g., for structuring conversation (Ashenfelter et al., 2009) and in modulating the extent of phonetic convergence (Miller et al., 2010) among others.

Here we demonstrate that EMA systems from two different manufacturers can be used without mutual interference to simultaneously record the speech articulator movements of interacting talkers with unimpeded visual contact. We collected data of two types: repeated incongruent word sequences (e.g., "*top cop* ..." vs. "*cop top* ...") and extended spontaneous conversation. In addition to establishing the viability of this dual-EMA approach, these data show clear examples of entrainment phenomena (discussed below), articulatory speech errors (cf. Pouplier, 2003; Goldstein et al., 2007), and kinematic aspects of phonetic convergence.

Electromagnetic Articulometry (EMA)

EMA is a point source tracking technique in which small sensors are attached with dental adhesive to various fleshpoints within a speaker's vocal tract (e.g., tongue, lips, maxilla, jaw). Radio-frequency transmitters induce voltages in the sensor coils positioned within the field of the device, and sensor position and orientation are subsequently reconstructed by comparing these

voltages to known reference values (Perkell et al., 1992; Zierdt et al., 1999). The technique provides good spatial (~ .3 mm) and temporal (100-400 Hz) resolution, and the kHz radio frequency radiation levels are well below those considered to be biologically significant.

For this project two types of commercially-manufactured EMA systems were employed. The AG500 (Carstens Medizinelektronik, GmbH) uses six narrowly tuned transmitters that operate continuously at different frequencies (7.5kHz – 13.75kHz); it has been assessed for speech purposes by Yunusova et al. (2009). The WAVE (Northern Digital, Inc.) uses eight strobed transmitters, all operating at 3kHz; it has been assessed by Berry (2011). 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.

Methods

The experiment was conducted in a large room at the MARCS Auditory Laboratories[1] of the University of Western Sydney. The two EMA systems were positioned 2 meters apart, as measured from the center of the AG500 cube to the center of the WAVE field generator.

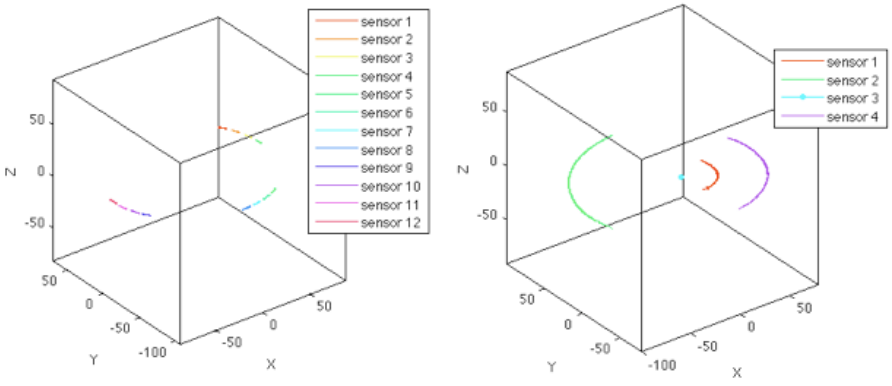

Figure 1: Sample joint operability tasks for AG500 (left) and WAVE (right). Stability was evaluated on Euclidean distances between sensors fixed to a moving rigid body.

Joint Operability Tests

To assess potential mutual interference from their joint operation a series of test recordings were made, first with each device operating alone and then with both active. In these tests sensors were mounted in a fixed position relative to a rigid body, which was then

[1] The MARCS Auditory Laboratories have since been renamed The MARCS Institute.

manipulated through systematic displacements and rotations in the field. The Euclidean distances between all possible sensor pairings were evaluated to obtain the (worst-case) maximum deviation from their measured distance, inter-quartile range, and standard deviation. We found no significant differences in these measures in comparing values obtained from the systems operated alone and operated together. Figure 1 shows sample tasks and Table I summarizes these results.

Status	Range	IQR	S.D.
AG500 only	1.72	0.38	0.28
AG500 + WAVE	1.62	0.37	0.26
WAVE only	1.21	0.22	0.18
WAVE + AG500	1.04	0.19	0.14

Table I: Comparable Euclidean distance measures (mm) evaluated with each system operated alone and together. AG500 values are maxima of 66 inter-sensor comparisons from 12 sensors; WAVE values are maxima of 6 inter-sensor comparisons from 4 sensors. **No significant differences observed with dual operation**.

Experiment

Two native speakers of American English participated in the subsequent experiment, one male and one female, both with normal speech and hearing. Each was seated such that their anterior vocal tracts were centered within the respective measurement fields (AG500 female, WAVE male), for a face-to-face distance of slightly less than 2 meters. Each participant had a clear view of the other speaker's face.

Figure 2: Dual-EMA experimental setup. AG500 on the right, NDI WAVE on the left. Separate directional microphones were used to isolate audio from each speaker.

Sensor deployment was the same for both speakers[2]: articulation was tracked through sensors placed on the tongue, jaw (lower incisors) and lips; head movement was tracked through references placed on the upper incisors (UI) and the mastoid process behind each ear.

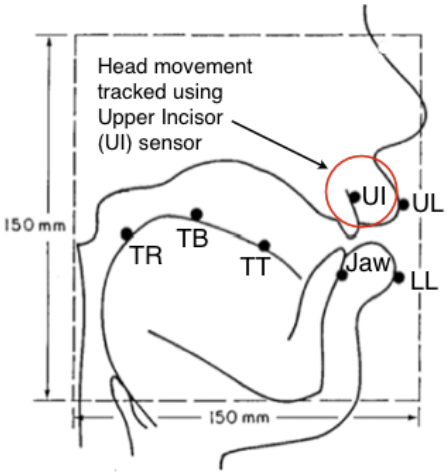

Figure 3: Sensor placement, used for both speakers: Tongue Tip (TT), Blade (TB) and Rear (TR); Upper and Lower Lips (UL, LL); Jaw (Lower Incisors); and Upper Incisors reference (UI).

Separate directional microphones located about 60 cm from the mouth were used to isolate audio for each speaker. Output was split in each case between input to the respective EMA device (for synchronization with transduced movement) and a separate two channel audio recording combining the inputs.

TRIAL	MALE	FEMALE	
trial 01	"cop cop..."	"top top..."	
trial 02	"top top..."	"cop cop..."	
trial 03	"cop top..."	"top cop..."	
trial 04	"top cop..."	"cop top..."	
trial 05	"copper copper..."	"topper topper..."	30 secs
trial 06	"topper topper..."	"copper copper..."	
trial 07	"topper copper..."	"topper copper..."	
trial 08	"copper topper..."	"topper copper..."	
trial 09	"topper copper..."	"copper topper..."	
trial 10	spontaneous conversation	spontaneous conversation	
trial 11	competing monologues	competing monologues	120 secs
trial 12-16	spontaneous conversation	spontaneous conversation	

Table II: Experiment speech tasks.

[2] The AG500 supports up to 12 sensors; the WAVE system installed at MARCS has an optional second sensor control unit supporting up to 16 sensors in total.

Table II summarizes the speech tasks performed by the participants. In the first part of the experiment each speaker was instructed to repeat a two-word sequence during 30 second trials, breathing as necessary, with one speaker using order AB and the other BA, as in for example "*top cop*" vs. "*cop top*." In the second part, speakers engaged in free conversation that bridged the 120-second trials.

The acquisition software proprietary to each system was run on separate computers. At the onset of each trial any necessary instructions were provided to the participants orally, then the AG500 was triggered first, followed by the WAVE. WAVE data acquisition was terminated automatically after a timed interval, following which AG500 sampling was toggled off manually. The two-channel audio was collected continuously at a 44100 Hz sampling rate.

Post-processing

Data for each speaker were aligned to a movement-corrected standard orientation through the following steps. First, to reduce noise, reference sensor trajectories were low-pass filtered at 5 Hz and movement sensor trajectories at 20 Hz (a copy of the UI sensor used to characterize head motion was also filtered at 20 Hz). Next the mean sensor locations from an initial interval prior to the onset of production in the first trial were used to establish a head reference position. The mapping from this reference position to a standard orientation was then established by a constrained optimization, resulting in a coordinate system aligned with the triangle determined by the mastoid and upper incisor references, and with origin at the upper incisors. Finally, for each trial sample, the rotation/translation matrix needed to align the reference sensors to the standard orientation/offset was computed using the method of Horn (1987) and applied as movement correction to all sensors at that sample.

The result of this series of steps was two sets of data per trial, one derived from the AG500 (female speaker) and one from the WAVE (male speaker). While each system recorded audio synchronized with sensor movement, because of the different starting and stopping times the AG500 data were not aligned with the WAVE data. To accomplish this alignment four additional steps were performed. First, audio from the three sources (AG500 16 kHz, WAVE 22050 Hz, and 2-channel 44100 Hz) were resampled to a consistent 11025 Hz sampling rate.

Next the AG500 200 Hz movement rate was decimated to match the WAVE 100 Hz rate. The peak of the cross-correlation between corresponding audio signals was then used to find the appropriate alignment of the shorter WAVE trials within the longer AG500 trials, and the latter were truncated accordingly. Finally, a single composite dataset was constructed for each trial comprised of data from both sources, using aligned audio drawn from the higher-quality 2-channel recording.

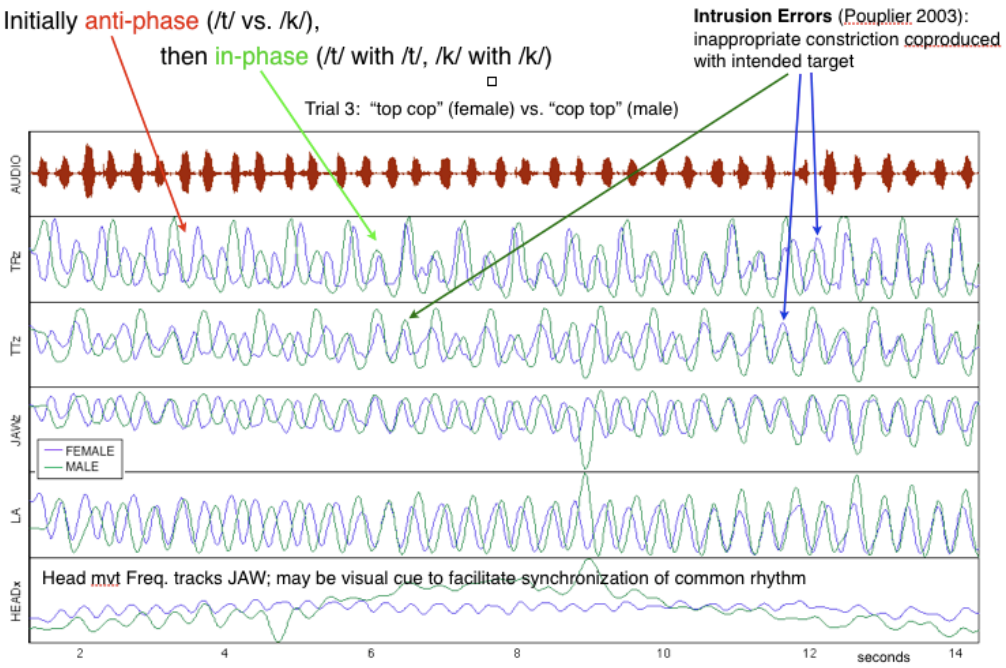

Figure 4: Incongruent sequence example, showing onset of in-phase entrainment and intrusive production errors. LA is the Euclidean distance between UL and LL; x shows posterior/anterior and z inferior/superior movement.

Discussion

Although intended primarily as a proof-of-concept demonstration, data resulting from this experiment nonetheless show several interesting features. In each of the incongruent word pair trials, despite starting out of phase with one another as instructed speakers rapidly entrained to one another in both phase and speech rate (see Figures 4 and 5). In addition, they readily tracked one another through spontaneous changes in speech rate (see Figure 6). Substantial head movement was observed by both speakers aligned with and tracking the frequency of jaw movement, and it is possible that this provided a visual cue to facilitate synchronization of a common production rhythm. Numerous 'covert' intrusive speech errors (inappropriate constriction coproduced with intended target; Pouplier, 2003) were observed with no overt

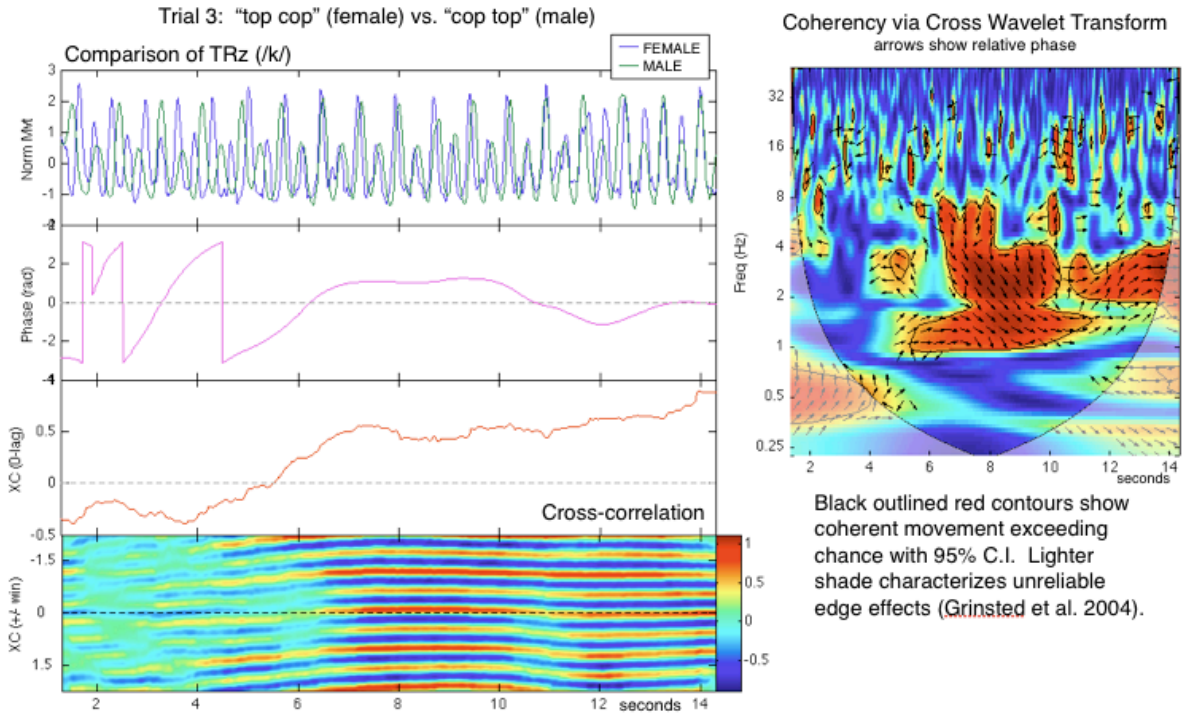

Figure 5: Onset of in-phase entrainment for vertical TR movement (/k/) in Trial 3. Initially unsynchronized and anti-phase, talkers entrain to in-phase synchronized productions; first led by female, then by male. At the 8 second mark coherent movement for both speakers can be observed at both the fundamental alternating frequency (~1.5 Hz) and the anti-phase movement frequency (~3 Hz).

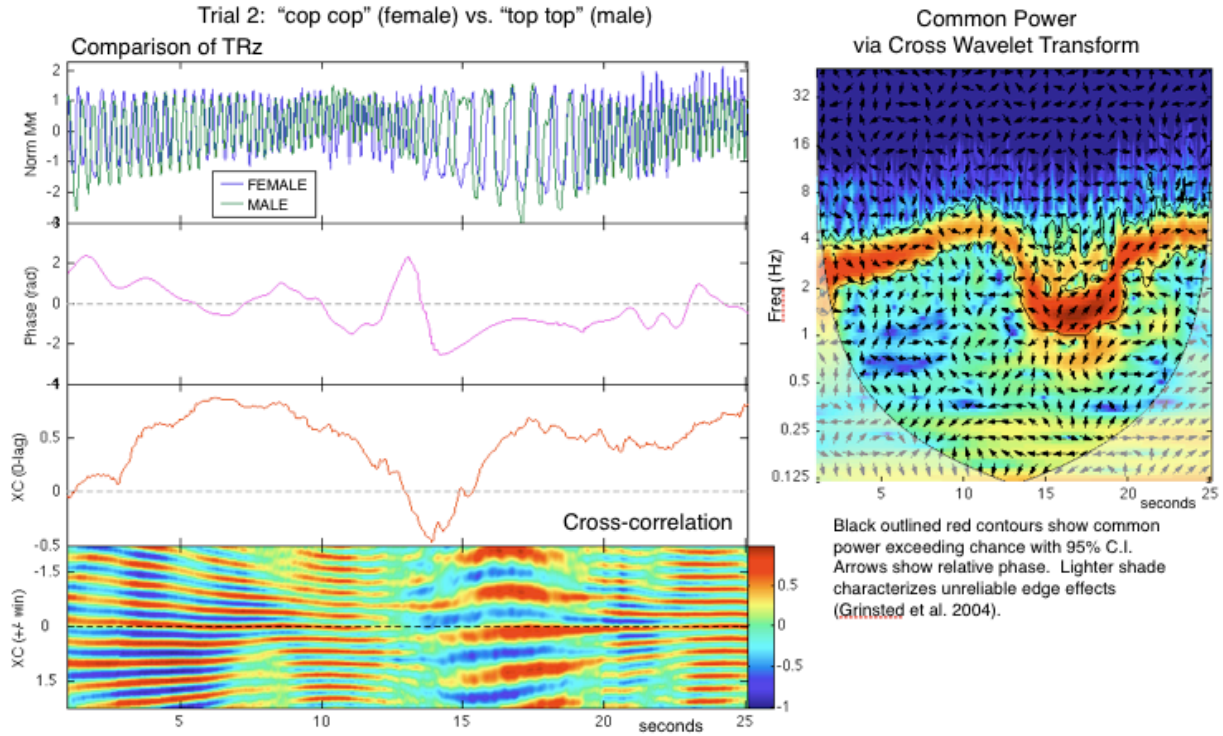

Figure 6: Vertical TR movement for Trial 2. Talkers rapidly converge on shared rhythm in repeated production task, and track one another consistently through spontaneous acceleration / deceleration.

auditory percept, and do not appear to have triggered corresponding 'errors' in interlocutor production. In addition, correlation map analysis of kinematic trajectories applied to the extended conversation trials has shown that the input signals (e.g., TR and TT) are continuously coordinated between speakers over an optimum correlation path, even as this coordination fluctuates over time (Barbosa et al., 2012).

Summary

Unforced conversational interaction depends on interlocutors positioned within reasonable proximity to one another, ideally with a clear view of each partner's face. We have demonstrated the feasibility of recording kinematic data from two facing talkers simultaneously when separated by a distance of two meters, using different models of EMA devices to avoid mutual interference. Preliminary data show error and rate effects related to spontaneous synchronization in repetitive word sequences. Results suggest that studies of phonetic convergence can be usefully extended to observation of potential shifts in articulation driven by conversational alignment.

References

Ashenfelter KT, Boker SM, Waddell JR, Vitanov N (2009) Spatiotemporal symmetry and multifractal structure of head movements during dyadic conversation. *J. Exp. Psych: Human Perception and Performance*, 35, 1072-91.

Barbosa AV, Déchaine R-M, Vatikiotis-Bateson E, Yehia HC (2012) Quantifying time-varying coordination of multimodal speech signals using correlation map analysis. *J. Acoust. Soc. Am*., 131, 2162-72.

Berry JJ (2011) Accuracy of the NDI Wave Speech Research System. *J. Speech, Language, and Hearing Res.*, 54, 1295-301.

Goldstein L, Pouplier M, Chen L, Saltzman E, Byrd D (2007) Dynamic action units slip in speech production errors. *Cognition*, 103, 386-412.

Grinsted A, Moore J, Jevrejeva S (2004) Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Processes in Geophysics*, 11, 561-66.

Horn BKP (1987) Closed-form solution of absolute orientation using unit quaternions. *J. Opt. Soc. Am.*, 4, 629-42.

Miller RM, Sanchez K, Rosenblum LD (2010) Alignment to visual speech information. *Attention, Perception, & Psychophysics*, 72, 1614-25.

Perkell JS, Cohen MH, Svirsky MA, Matthies ML, Garabieta I, Jackson MTT (1992) Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *J. Acoust. Soc. Am*., 92, 3078-96.

Pouplier, M (2003) The dynamics of error. *Proc. 15th ICPhS*, 2245-48.

Turk A, Scobbie J, Geng C, Macmartin C, Bard E, et al. (2010) The Edinburgh Speech Production Facility's articulatory corpus of spontaneous dialogue. *J. Acoust. Soc. Am*., 128, 2429.

Yunusova Y, Green JR, Mefferd A (2009) Accuracy assessment for AG500, electromagnetic articulograph. *J. Speech, Language, and Hearing Res.*, 52, 547-55.

Zierdt A, Hoole P, Tillmann HG (1999) Development of a system for three-dimensional fleshpoint measurement of speech movements. *Proc. 14th ICPhS*, 1, 73-75.