THE ELECTROMAGNETIC ARTICULOGRAPHY MANDARIN ACCENTED ENGLISH (EMA-MAE) CORPUS OF ACOUSTIC AND 3D ARTICULATORY KINEMATIC DATA

An Ji[1], Jeffrey J. Berry[2], Michael T. Johnson[1]

[1]Speech and Signal Processing Laboratory, Department of Electrical and Computer Engineering
[2]Speech and Swallowing Laboratory, Department of Speech Pathology and Audiology
Marquette University, Milwaukee, USA
{an.ji, jeffrey.berry, mike.johnson}@marquette.edu

ABSTRACT

There is a significant need for more comprehensive electromagnetic articulography (EMA) datasets that can provide matched acoustics and articulatory kinematic data with good spatial and temporal resolution. The Marquette University Electromagnetic Articulography Mandarin Accented English (EMA-MAE) corpus provides kinematic and acoustic data from 40 gender and dialect balanced speakers representing 20 Midwestern standard American English L1 speakers and 20 Mandarin Accented English (MAE) L2 speakers, half Beijing region dialect and half are Shanghai region dialect. Three dimensional EMA data were collected at a 400 Hz sampling rate using the NDI Wave system, with articulatory sensors on the midsagittal lips, lower incisors, tongue blade and dorsum, plus lateral lip corner and tongue body. Sensors provide three-dimensional position data as well as two-dimensional orientation data representing the orientation of the sensor plane. Data have been corrected for head movement relative to a fixed reference sensor and also adjusted using a biteplate calibration system to place the data in an articulatory working space relative to each subject's individual midsagittal and maxillary occlusal planes. Speech materials include isolated words chosen to focus on specific contrasts between the English and Mandarin languages, as well as sentences and paragraphs for continuous speech, totaling approximately 45 minutes of data per subject. A beta version of the EMA-MAE corpus is now available, and the full corpus is in preparation for public release to help advance research in areas such as pronunciation modeling, acoustic-articulatory inversion, L1-L2 comparisons, pronunciation error detection, and accent modification training.

Index Terms Articulator motion, Electromagnetic Articulography, non-native speech, English-Mandarin contrasts

1. INTRODUCTION

Electro-Magnetic Articulography (EMA) has been a rapidly growing technology for accurate measurement of articulatory kinematics [1-3]. This technology is based on measuring the induced current caused by motion of encapsulated miniature toroid coils in a system of electromagnetic fields, and is now available through multiple manufacturers, including the Northern Digital, Inc. Wave system [4] and Carstens AG-series [5, 6]. Our NDI Wave system captures both position and orientation at a sampling rate of up to 400 Hz, with position error on the scale of +/- 0.5mm. A single sensor captures 5 Degree of Freedom (DOF) information, including 3 dimensional position information plus the 2-dimensional orientation of the sensor plane. A 6 DOF sensor can be constructed using dual non-planar coils to capture full orientation information.

Prior to EMA, a number of technologies have been used historically for recording articulator movements. X-ray cinematography [7, 8] is effective, but the radiation to the subject's head is a concern. Cine MRI [9, 10] can provide dynamical 3D measurement of the vocal tract but it is somewhat cumbersome and expensive. Ultrasound [11, 12] is able to capture the surface of the tongue, but noise, echo artifacts and refractions may affect the results. EMA has a number of significant advantages over these technologies, including full three-dimensional representation, a sufficient temporal resolution to capture articulatory dynamics, relatively low measurement error, and low cost. EMA technology has seen rapid growth as a cost-effective platform for collection of synchronous acoustic and articulatory kinematic data.

There are several publicly available datasets of articulatory movements, such as the X-Ray Microbeam Speech Production Database [13], the MOCHA-TIMIT (one female and one male native English speaker) and MNGU0 (one male British English speaker) databases [14-15], the EUR-ACCOR multi-language articulatory database (10 speakers from seven languages) [16], the multimodal real-time MRI articulatory corpus MRI-TIMIT (two female and two male American English speakers) [17], and the most recent Edinburgh speech production facility DoubleTalk

corpus (six native English speakers with varying accents) [18]. However, these are all relatively small datasets, and do not contain L2 speakers, which makes investigation of language learning and differences in L1 and L2 groups infeasible.

This paper introduces a new EMA dataset that includes gender and dialect balanced data from both native English speakers and native Mandarin speakers speaking English. Data have been corrected for head movement relative to a fixed reference sensor and also adjusted using a biteplate calibration system to place the data in an articulatory working space relative to each subject's individual midsagittal and maxillary occlusal planes.

Following this introduction, Section 2 of the paper will introduce the data collection methodologies and speech tasks used with this dataset. Section 3 will describe the data processing methods, including head correction, bite plate correction, and palatal mesh processing. Section 4 will describe the annotation and transcription methods, while Section 5 will give an overview of the software tools included with the dataset. Sections 6 and 7 summarize conclusions and acknowledgements.

2. DATA COLLECTION

2.1 Subjects

The EMA-MAE corpus includes 40 subjects, including two primary subject groups designated L1 and L2. The L1 group consists of 10 male and 10 female native speakers of English, with an upper Midwest American English dialect background. The L2 group consists of 10 male and 10 female native speakers of Mandarin Chinese who speak English as a second language. Within the L2 group is a further dialectal division into subjects with a northern Beijing-region dialect background, and subjects with a southern Shanghai-region dialect background, with 5 male and 5 female speakers from each of these subgroups.

Subjects are between the ages of 18-40 with no history of speech, language, or hearing pathology, no history of orofacial surgery (other than typical dental extractions), and no history of use of anticonvulsant, antipsychotic, or anti-anxiety medications (as these factors may affect motor performance).

2.2 Speech Tasks

The corpus includes approximately 45 minutes of synchronized acoustic and kinematic data for each speaker, including word, sentence, and paragraph level speech samples.

2.2.1 MAE minimal contrast word pairs

Subjects read 330 text-prompted words in single-word citation form [19-20]. Words were blocked into approximately 25 words per record, to allow monitoring of sensor adhesion and give participants regular rest and adjustment periods.

Words included Rogers' list of minimally contrasting words with emphasis on the likely acoustic-phonetic confusions characteristic of Mandarin Accented English, as well as a set of words and pseudo words covering the phonetic space of English vowels [20, 21].

2.2.2 Sentences

Text-prompted read sentence materials include selected sentences from the TIMIT database [22] and from the Harvard Intelligibility Sentences [23], as well as sentences containing words with contrastive stress.

2.2.3 Connected speech

Connected speech materials include several read paragraphs, chosen to emphasize intelligibility, breath group utilization, accented-English intelligibility, speaking rate, and segmental timing [24-27] .

2.3 Data collection methodology

The EMA-MAE corpus includes synchronous acoustic sampled at 22kHz and three-dimensional kinematic data sampled at 400Hz. Data were collected in an acoustic booth with time-syncronized acoustic records obtained using a cardioid pattern directional condenser microphone positioned approximately 1 m from the center of the electromagnetic field. Participants were seated in a custom plastic chair designed to allow subjects to maintain a comfortable speaking posture.

As shown in Figure 1, articulatory sensors included the jaw (MI) lower front incisor), lower lip (LL), upper lip (UL), tongue body (TD), and tongue tip (TT), all placed in the midsagittal plane. In addition, there were two lateral sensors, one (LC) at the left corner of the mouth to help indicate lip rounding and one (LT) in the left central midpoint of the tongue body to help indicate lateral tongue curvature.

A reference sensor (RE) was located near the bridge of the nose using a pair of plastic glasses. The reference sensor was a 6 DOF sensor, providing three dimensional position as well as three-dimensional orientation data. All other sensors in the system were 5 DOF sensors, since these are significantly smaller and have less interference with natural subject articulation. Five DOF sensors provide three dimensional position information but only two dimensional orientation data. This identifies the orientation, i.e., pitch and roll, of the sensor plane (the plane of the toroidal sensor coil, perpendicular to the toroid axis) but no information about yaw of this plane. Position data are given in millimeters. Orientation data are given in quaternion rotation format, indicating rotation axis and angle relative to a base orientation.

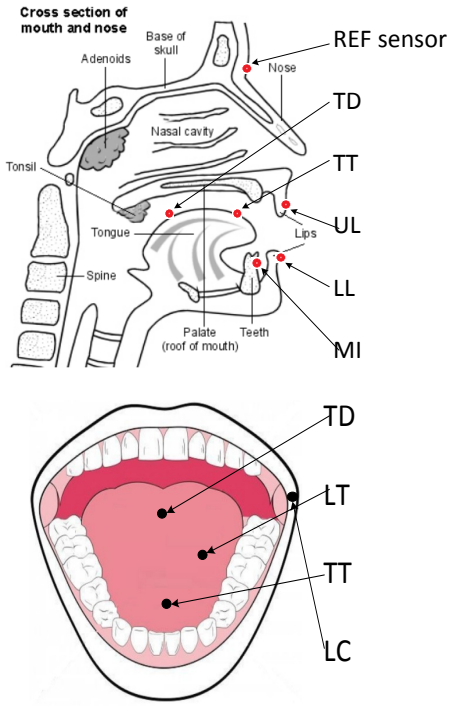

Figure 1 Sensor placement.

Subjects each underwent an initial calibration process in which softened dental wax is formed into a bite plate around a tongue depressor and a dental impression is taken. Biteplate sensors are placed at the front incisor and at the mid-point of the back molars to indicate the midsagittal and maxillary occlusal planes relative to the reference sensor, which is used in data processing to form a consistent articulatory working space. Some of the subjects also have a third lateral biteplate sensor used with alternative articulatory space calibrations, an addition made about midway through the data collection procedure. Biteplate configuration is pictured in Figure 2.

Subjects also underwent a palatal calibration in which the experimenter used a sensor-tipped palate wand to collect palatal reference data. As described in the next section, this palate information can be used to provide palate information for determining vocal tract configuration relative to tongue sensors.

In addition to the biteplate and palatal calibration processes, subjects were given an acclimation period and opportunity to read some practice materials once sensors had been attached.

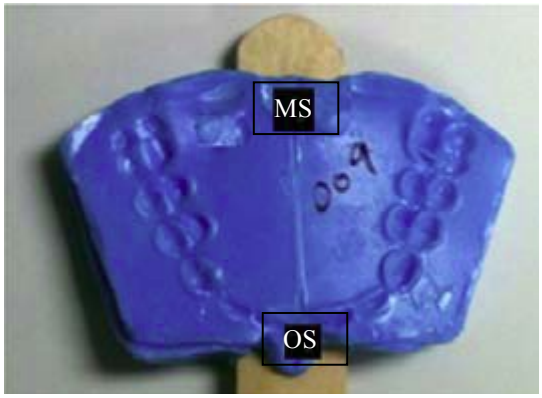

Figure 2 Biteplate calibration

3. DATA PROCESSING

Raw data from the EMA system are in a global coordinate space relative to the system's electromagnetic field. There is significant data processing required to compensate for subject movement and physiology and provide data in an appropriate articulatory working space. In addition, palatal data can be used to determine vocal tract measures, e.g. the vertical distance from the tongue to the hard palate or estimates of cross-sectional areas.

3.1 Internal head-correction

Transformation of the global coordinate data into a local coordinate space relative to a fixed reference sensor is handled in real-time by the NDI Wave software. As described in Section 2.3, a reference sensor mounted on a pair of plastic glasses is used with all subjects to determine and compensate for head movements. Position data are adjusted by a direct linear translation, and orientation data are adjusted through a quaternion rotation relative to the reference sensor's orientation.

To establish some measures of head correction and biteplate calibration variance, about half of the subject data include an additional calibration step in which subjects were asked to nod their heads up and down and move their heads back and forth with the bite plate in their mouths.

3.2 Biteplate calibration

Biteplate calibration refers to the translation and rotation of the head-referenced coordinate system into a subject-referenced articulatory working space. Through the biteplate calibration process described in Section 2.3, the exact positions of the OS and MS reference sensors are determined. Together these sensors represent the line of intersection between the subject's midsagittal plane and the maxillary occlusal plane, and together with the REF sensor they define the necessary working space.

The biteplate calibration process begins by translating the origin of the coordinate system from the REF to the OS sensor, including a small adjustment for incisor and dental

width in order to place the origin at the inside edge of the upper front incisor. Following this, a single unique rotation is identified that will place the triangle formed by the REF, OS, and MS sensors onto the XY plane, with OS at the origin and MS along the X axis. This creates an articulatory working space in which the XY plane is the midsagittal plane and the XZ plane is the maxillary occlusal plane.

Biteplate calibration places all position and orientation data into a physiologically-referenced articulatory working space. In addition, the biteplate record also provides reference position and orientation data for all sensors in an "at-rest" position. This reference information can be used to determine sensor placement consistency or to help calculate sensor plane angles during speech. Detailed derivations of the biteplate calibration method can be found in [28].

The EMA-MAE dataset includes photographs of the biteplates for each subject, against a grid-lined background. This provides reference data for determining internal and external dental perimeters that, together with palate data, enable spatial normalization of the articulatory working space.

3.3 Palate trace processing

For each subject, palate data includes a trace of the mid-sagittal palate line, a series of transverse traces across the palate, and both inner perimeter and outer perimeter dental traces at the gum line. Together with the bite plate data, this information provides reference data that can be used to calculate physiologically-referenced vocal tract measures.

In addition to the raw palate data, the EMA-MAE dataset includes a referential palate mesh computed on a grid in the articulatory working space provided by the bite-corrected data. This mesh is computed using the thin-plate spline method with a smoothing factor of 0.05 as recommended by error and variance analysis [29], with a vertical half-sensor offset to account for the wand sensor thickness. Both the mesh grid and the software used to compute it are provided.

4. TRANSCRIPTION AND ANNOTATION

For all subjects, both phoneme-level transcriptions and orthographic transcriptions of the target utterances are provided. Transcription was completed by trained graduate students in Marquette's Speech Pathology and Audiology program using an IPA notation American English phoneme set. All transcriptions were completed by trained listeners with a common upper Midwest American English dialect. Multiple listener transcriptions are included for L2 subjects, to use for estimating perceived phoneme variability and perceived intelligibility. For the connected speech data, timestamps of clear pause locations (breath group and/or sentence boundaries) are included so that the paragraph-level utterances and transcriptions can be easily subdivided into sentence level data if desired.

Additional transcription information includes a look-up table of phoneme usage for each subject, to provide a simple mechanism for L1-L2 comparisons of selected phonemes.

5. SOFTWARE TOOLS

A basic set of software tools for processing the NDI Wave kinematic data is included with the EMA-MAE dataset. Stand-alone Windows platform tools are provided to translate between global coordinate space and head-corrected space and also between head-corrected space and the biteplate corrected articulatory working space. The biteplate correction tool includes optional offsets to change the origin of the working space as well as toggle switches to change the positive/negative direction of the coordinate axes.

In addition, a Matlab toolbox is included with functions to do these same coordinate space transformations, plus some additional functionality. This includes basic tools to read the data files, to construct a palate map function from the palate cloud record using the thin plate spline method, and to construct a simple tongue mesh from the three tongue sensors (two midsagittal, one lateral) that incorporates both position and orientation information. A tool demonstrating the combined use of the palate mesh and tongue mesh data to compute a grid of palate-to-tongue distances and cross-sectional vocal tract areas is also included.

6. CONCLUSIONS

The EMA-MAE dataset is intended to fill a need for comparative acoustic and three-dimensional kinematic data across L1 and native-Mandarin L2 speaker sets. A beta version of the EMA-MAE corpus is now available for download at http://speechlab.eece.mu.edu/emamae , and the full corpus is in preparation for public release to help advance research in areas such as pronunciation modeling, acoustic-articulatory inversion, L1-L2 comparisons, pronunciation error detection, and accent modification training.

7. ACKNOWLEGEMENTS

Funding for this work was provided by the National Science Foundation under NSF IIS-1320892.

8. REFERENCES

[1] R. Schwestka-Polly, W. Engelke and D. Engelke, "The importance of electromagnetic articulography in studying tongue motor function in the framework of an orthodontic diagnosis," Fortschritte Der Kieferorthopädie, pp. 3-10, 1992.

[2] H. Horn, G. Göz, M. Bacher, M. Müllauer, I. Kretschmer and D. Axmann-Krcmar, "Reliability of electromagnetic articulography recording during speaking sequences," European Journal of Orthodontics, pp. 647-655, 1997.

[3] S. Fuchs, P. Perrier and B. Pompino-Marschall, "Speech production and perception: Experimental analyses and models," ZAS Papers in Linguistics, 2005.

[4] J. J. Berry, "Accuracy of the NDI Wave speech research system," Journal of Speech, Language and Hearing Research, vol. 54, pp. 1295, 2011.

[5] C. Kroos, "Evaluation of the measurement precision in three-dimensional Electromagnetic Articulography (Carstens AG500)," Journal of Phonetics, vol. 40, pp. 453-465, 2012.

[6] Y. Yunusova, J. R. Green and A. Mefferd, "Accuracy assessment for AG500 electromagnetic articulograph," Journal of Speech, Language and Hearing Research, vol. 52, pp. 547-555, 2009.

[7] K. G. Munhall, E. Vatikiotis-Bateson and Y. Tohkura, "X-ray film database for speech research," Journal of the Acoustical Society of America, pp. 1222-1224, 1998.

[8] J. S. Perkell, "Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study," 1969.

[9] S. Masaki, M. K. Tiede, K. Honda, Y. Shimade, I. Fujimoto, Y. Nakamura and N. Ninomiya, "MRI-basd speech production study using a synchronized sampling method," Journal of the Acoustical Society of America, vol. 20, pp. 375-397, 1999.

[10] S. Narayanan, K. Nayak, S. Lee, A. Sethy and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," Journal of the Acoustical Society of America, vol. 115, pp. 1771-1776, 2004.

[11] M. L. Stone, B. C. Sonies, T. H. Shawker, G. Weiss and L. Nadel, "Analysis of real-time ultrasound images of tongue configuration using a grid-digitizing system," Journal of Phonetics, vol. 11, pp. 207-218, 1983.

[12] T. Kaburagi and M. Honda, "An ultrasonic method for monitoring tongue shape and the position of a fixed-point on the tongue surface," Journal of the Acoustical Society of America, vol. 95, pp. 2268-2270, 1994.

[13] J. R. Westbury. "X-ray microbeam speech production database user's handbook." Madison, WI: University of Wisconsin Press, 1994

[14] A. A. Wrench and H. J. William, "A multichannel articulatory database and its application for automatic speech recognition", 5th Seminar on Speech Production: Models and Data, Bavaria, pp. 305-308, 2000.

[15] K. Richmond, P. Hoole, and S. King, "Announcing the electro-magnetic articulography (day 1) subset of the mngu0 articulatory corpus," Interspeech, Florence, Italy, pp. 1505-1508. 2011

[16] http://www.cstr.ed.ac.uk/research/projects/artic/accor.html

[17] http://sail.usc.edu/span/mri-timit/

[18] J. M. Scobbie, A. Turk, C. Geng, S. King, R. J. Lickley and K. Richmond. "The ediburgh speech production facility doubletalk courpus." Interspeech, Lyon, France, pp. 764-766, 2013

[19] C. Rogers, "Intelligibility of Chinese-accented English (Doctoral dissertation, Indiana University, 1997)," Dissertation Abstracts International, vol. 58, 1997.

[20] J. Hillenbrand, L. A. Getty, M. J. Clark and K. Wheeler, "Acoustic characteristics of American English vowels," Journal of the Acoustical Society of America, vol. 97, pp. 3099, 1995.

[21] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," Journal of the Acoustical Society of America, vol. 24, pp. 175-184, 1952.

[22] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," National Technical Information Service, 1993.

[23] IEEE Subcommittee, "IEEE Recommended Practice for Speech Quality Measurements," IEEE Transactions on Audio and Electroacoustics, vol. 17, pp. 225-246, 1969.

[24] J. R. Green, D. R. Beukelman and L. J. Ball, "Algorithmic estimation of pauses in extended speech samples of dysarthric and typical speech," Journal of Medical Speech-Language Pathology, vol. 12, pp. 149, 2004.

[25] T. H. Crystal and A. S. House, "Articulation rate and the duration of syllables and stress groups in connected speech," Journal of the Acoustical Society of America, vol. 88, pp. 101, 1990.

[26] R. Patel, K. Connaghan, D. Franco, E. Edsall, D. Forgit, L. Olsen, L. Ramage, E. Tyler and S. Russell, "The Caterpillar: A Novel Reading Passage for Assessment of Motor Speech Disorders," American Journal of Speech-Language Pathology / American Speech-Language-Hearing Association, Jul 30, 2012.

[27] D. Honorof, J. McCullough and B. Somerville, "Comma gets a cure: A diagnostic passage for accent study," Retrieved February, vol. 20, pp. 2007, 2000.

[28] A. Ji, M. T. Johnson and J. Berry, "Articulatory space calibration in 3D electro-magnetic articulography," in China SIP International Conference on Signal and Image Processing, 2013, .

[29] Y. Yunusova, M. Baljko, G. Pintilie, K. Rudy, P. Faloutsos and J. Daskalogiannakis, "Acquisition of the 3D surface of the palate by in-vivo digitization with Wave," Speech Communciation, pp. 923-931, 2012.