Forschungsberichte des
Instituts für Phonetik und
Sprachliche Kommunikation
der Universität München
(FIPKM) 35 (1997) 165-170

Accuracy of Lip Movement Analysis. Comparison between Electromagnetic Articulography and an Optical Two-Camera Device.

Ingo Hertrich, Hermann Ackermann

University of Tübingen, Department of Neurology
Hoppe-Seyler-Str. 3, D-72076 Tübingen, Germany
E-mail: ingo.hertrich@uni-tuebingen.de

ABSTRACT

The aim of the present study was to determine the accuracy of Electromagnetic Articulography (EMA) under in-vivo conditions by comparing its results to measurements performed with an optoelectronic system (ELITE). The Elite system, based on two cameras emitting infra-red light, allows the tracking of 3-dimensional movements of reflectors attached to the points of interest relative to the space in which the cameras are fixed. In contrast, EMA provides a two-dimensional representation of the trajectories of small receiver coils within the mid-sagittal plane of a helmet to which three larger transmitter coils are attached. For the purpose of the present study two EMA receiver coils were attached to the upper and lower lip, respectively. Similarly, two ELITE reflectors were stuck on top of these coils. The test material consisted of German sentences of the type "Ich habe gepVpe gelesen" (engl. "I have read gepVpe") in which the target vowel V was substituted by /a/, /i/, /u/, or /y/. Eight repetitions of the sentences were produced by a female speaker in randomized order, a) with normal, b) fast, and c) slow rate. The bilabial closing gesture terminating the target vowel was chosen for analysis. The average absolute difference between the EMA and the ELITE data was about .5 mm for movement amplitude, 8 mm/s for peak velocity, and 8 milliseconds for the duration of the closing gesture. With respect to amplitude and velocity, the deviation between the two systems was considerably lower than the standard deviation across the repetitions within each vowel category and speech rate condition. Thus, both methods confirm each other as being adequate to investigate speech gestures.

INTRODUCTION

Within the last decade the kinematic assessment of speech articulation has made some advances. In spite of all problems arising from X-ray techniques, the traditional cinegraphic X-ray representation seems to be the most insightful method to visualize the articulatory gestures as a whole (Sock, Perrier, Bensaber, Bothorel, Brock, and Serignat, 1995). However, the development of dynamic models of speech production requires quantification of kinematic parameters such as peak velocity and movement amplitude. In case of visible movements, e.g. lip gestures, analysis can rely on optoelectronic systems (Kelso, Vatikiotis-Bateson, Saltzman, and Kay, 1985; Munhall, 1993; Zmarich, Magno-Caldognetto and Vagges, 1995; Magno-Caldognetto, Vagges and Zmarich, 1995) or on mechanical movement transducers (Müller and Abbs, 1979). Intra-oral movement tracking, however, requires methods such as the X-ray microbeam technique (Kiritani, Itoh, and Fujimura, 1975; Westbury 1994), ultrasonic methods (Keller and Ostry, 1983; Sonies, Shawker, Hall, Gerber, and Leighton, 1981; Ostry and Munhall, 1985), or electromagnetic (midsagittal) articulography (EMA or EMMA) (Schönle, Gräbe, Wenig, Höhne, Schrader, and Conrad, 1987; Perkell, Cohen, Svirsky, Matthies, Garabieta, and Jackson, 1992). Even though the latter method is relatively inexpensive as compared to the microbeam system, its temporal resolution is considerably higher. With respect to spatial accuracy, several difficulties have to be met, e.g. thermal instability, electromagnetic noise, rotation of sensors, deviation from the midsagittal plane, and interaction of system components. Each of these sources of error may lead to measurement deviances on the order of .1 to 1 mm (Gracco and Nye, 1993; Perkell et al., 1992; Tuller, Shao, and Kelso, 1990). "Bench tests" with known receiver coil positions after a more advanced calibration procedure reported measurement errors of less than .5mm (Hoole, 1993), which can be regarded as a considerable improvement as compared previous studies on EMA accuracy (e.g. Schönle, Müller and Wenig, 1989) using the Carstens AG100 or its precursor model. However, an empirical validation under realistic conditions is still outstanding (Hoole, 1993). In fact, one study has been performed comparing EMA data with ultrasound recordings (Honda and Kaburagi, 1993), reporting an accuracy of about 1 mm. However, this measurement was performed using only the standard calibration procedure of the Carstens Articulograph AG100 (Carstens Medizinelektronik, Göttingen, Germany), not the extended version including a complete set of calibration data for each sensor coil separately (see Hoole, 1993). Furthermore, the two methods, EMA and ultrasonography, produce quite different types of data: fleshpoint trajectories in case of EMA, and contour images in case of the ultrasound method. Therefore, in the first instance they may be regarded as complementary (Stone, 1990) rather than validating each other.

The purpose of the present study is to evaluate electromagnetic articulography under in vivo speech conditions by use of synchronous optoelectronic measurements. Both systems can measure lip trajectories, but their working principles are completely different and independent of each other. The optoelectronic system is based on two cameras allowing the tracking of 3-dimensional trajectories of reflectors attached to the points of interest, relative to the space in which the cameras are fixed (Borghese and Ferrigno, 1990; Ferrigno and Pedotti, 1985). In contrast, EMA provides a two-dimensional representation of the trajectories of small receiver coils within the mid-sagittal plane of a helmet to which three larger transmitter coils are attached. Since the intention was to determine the error which is actually relevant to phonetic studies, measurement parameters were chosen which might also be used in 'real' studies: movement amplitude, movement time and peak velocity of a bilabial closing gesture. In addition, some statistical 'laboratory' measurement data are presented in order to show the extent of technical random noise.

METHODS

Electromagnetic articulography

The working principle of the Articulograph (AG100, Carstens Medizinelektronik, Göttingen, Germany) has been described elswhere (Hoole, 1993; Perkell et al., 1992; Schönle, 1988; Schönle et al., 1987). Three transmitter coils, the axes of which have an orthogonal orientation to the midsagittal plane, are mounted on a helmet near the speaker's chin, neck, and forehead. Each of them produces an alternating electromagnetic field at a frequency of about 16, 18, and 20 kHz, respectively. Small transducer coils are attached to the speaker's articulators of interest, receiving the transmitted frequencies with intensities which are approximately inversely proportional to the third power of the distances from the respective transmitter coils. Calibration was performed with the 'MKal32' (Carstens Medizinelektronik) calibration device which is comparable to the one used in Hoole (1993). For data acquisition a sampling rate of 200 Hz was chosen. Before the main experiment was performed, random measurement noise on the measured distance between two receiver coils was determined in the present study.

In the main experiment two EMA receiver coils were attached to the upper and lower lip, respectively. For the purpose of another study, three further coils were placed on the nasion, the tip of the nose, and the mandible just below the lower incisors. The mandible and lip sensors were attatched using an histoacrylic adhesive, and the two nose coils were fixed with a plaster.

Optoelectronic measurements

The optoelectronic system used in the present study (ELITE, BTS Bioengineering Technology & Systems, Milano, Italy), consists of two cameras, the objectives of which are surrounded by infra-red emitting diodes. Reflecting markers are attached to the moving structures of interest. The reflector locations are automatically detected within the 2-d representation provided by each camera. The 3-D-reconstruction is done in a separate session under manual control for correct marker assignment.

For the purpose of the present study the two cameras were mounted about 2 meter in front and 1 meter to the left and right, respectively, of the speaker's head while she was sitting on a chair. Four reflectors (diameter 6 mm) were attached on top of the four visible EMA receiver coils, i.e. the two nose and the two lip sensors. For the purpose of another study, two additional reflectors were attached to the left and right corner of the mouth, respectively. In order to achieve maximal temporal resolution, zoom objectives were used allowing the calibrated space to be confined to about 400x400x400 mm with the speaker's head in the center. The sampling rate was set to the maximum of 100 Hz. The spatial resolution of the system, i.e. the smallest detectable displacement, has been reported by the manufacturer to be about .1 mm which is comparable to the resolution of other optical systems (Munhall, 1993). Three further laboratory measures of accuracy, i.e. random noise, distortion of the calibrated space, and the error due to movement of targets, were determined under the calibration conditions of the present study by repeatedly measuring the distance between two reflecting markers mounted at a fixed distance from each other.

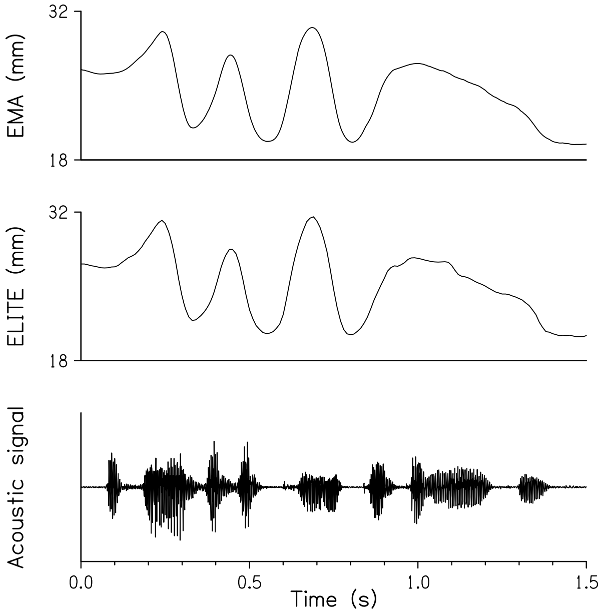

Figure 1. Test sentence "Ich habe gepape gelesen": Acoustic signal (bottom) and the parameter 'lip distance', i.e. the distance between the midsagittal upper and lower lip markers, measured with Electromagnetic articulography (TOP) and with the optoelectronic ELITE system.

Speech material and procedure

The test material used in the main experiment consisted of German sentences of the type "Ich habe gepVpe gelesen" (Engl. "I have read gepVpe") in which the target vowel V was substituted by /a/, /i/, /u/, or /y/. Eight repetitions of the sentences were produced by a female speaker in randomized order, a) with normal, b) fast, and c) slow rate. Altogether 4 (vowels) x 3 (rate conditions) x 8 (repetitions) = 96 sentences were recorded.

Primary parameters

The primary measurement parameter was the 'lip distance' defined as the 2-dimensional distance between the upper and lower lip receiver coils in case of EMA data and the 3-dimensional distance between the two lip reflectors attached to these coils in case of the ELITE data. In order to provide comparability with respect to the temporal resolution and upper frequency cutoff of the two systems, the ELITE data were low-pass filtered with a 3-point averaging filter and the EMA data with 5-point averaging. Figure 1 shows an example of the lip distance measured with the two systems during a test utterance. 'Movement velocity' was defined as the first derivative of the lip distance using a 3-point method: vel(i)= (x(i+1)-x(i-1))/2dt where x(i) is the i-th sample of the lip distance parameter, vel(i) the corresponding velocity, and dt the time interval between two adjacent samples.

Derived parameters

'Movement amplitude' of the bilabial closing gesture terminating the target vowel was defined as the difference of the lip distance parameter between maximum opening during the target vowel V and maximum closure during the succeeding /p/ stop. 'Movement time' was defined as the duration of this gesture, and 'peak velocity' as the highest velocity value occuring during this gesture.

RESULTS AND DISCUSSION

Pre-experiment 1: Random noise of the EMA system

Two receiver coils were placed within the helmet in a distance of about 4 cm from each other. 12 recordings of 1-2 seconds each were performed. Between recordings the location of the sensors within the calibrated space was changed while their physical distance from each other remained approximately the same. For each of the 12 recordings the standard deviation of the measured distance was computed. It ranged from 0.07 mm to .62 mm. On average the intra-recording standard deviation was .20 mm

Pre-experiment 2: Accuracy of the Elite system

Laboratory assessment of the accuracy of the Elite system was performed by fixing two reflecting markers at a distance of about 76 mm onto an object which could be moved including translation and rotation such that the markers remained visible to both cameras. The object was placed at 50 different positions within the calibrated space. In each position a recording of one second duration (=100 samples) was made while the object did not move. For each recording the mean three-dimensional distance between the two markers across the 100 samples was computed as well as the respective standard deviation. The average (across the 50 recordings) intra-recording standard deviation was used as a measure of random noise. It amounted to 0.09 mm ranging from 0.02 to 0.34 mm. As a measure of error due to orientation and placement, the standard deviation across the 50 recording means was considered, which was 0.54 mm. This quite large error, which might be due to extreme rotation and placement at the margins of the calibrated space, is larger than the error of 1/4000 of the field of view, reported by Borghese and Ferrigno (1990).

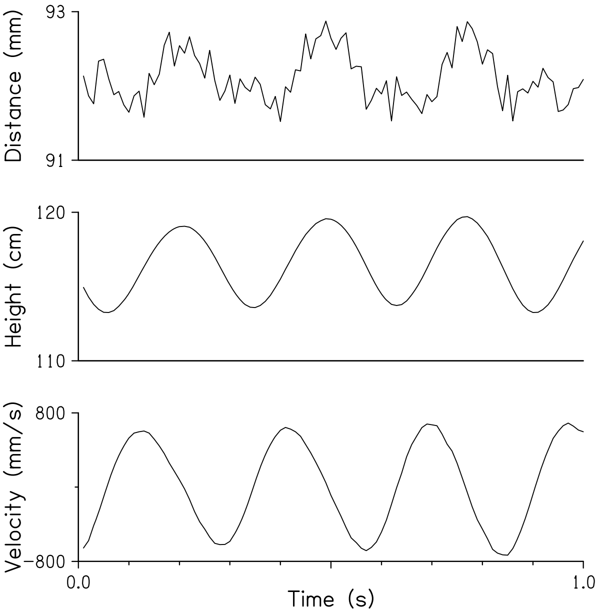

Figure 2. Effect of target movement on measurement accuracy: The top panel shows the measured distance between the two markers fixed on the stick while the stick was moved up and down. In the mid panel the actual up and down movement is displayed, i.e. the vertical coordinate of one of the two markers, in cm above the floor of the recording room. The lower panel contains the corresponding velocity curve, i.e. the first derivative of the curve displayed in the mid panel.

In order to determine the system's accuracy which is actually relevant for speech movements, two reflectors were mounted on a small stick at a fixed distance. During 10 recordings the stick was held by a person without volontary movement, during 10 succeeding recordings the person was instructed to move the stick about 50 mm up and down as fast as possible. The actual movement amplitude in the vertical direction ranged from about 40 to 80 mm, the corresponding peak velocity was about 400-800 mm/s, which may be regarded as the upper limit occurring in labial articulation gestures. The average standard deviation of the distance between the two markers during the first ten recordings was 0.19 mm, whereas under the movement instruction it was 0.34 mm. Post-hoc inspection of the movement data showed that the measured distance varied in phase with absolute displacement and not with movement velocity (Figure 2). Therefore, a particular error due to movement of targets may be neglected.

Main Experiment: *Synchronous measurement EMA and ELITE*

As a first step to compare two methods of kinematic data acquisition, the Pearson correlation coefficients between the two sets of parameters across all 96 test utterances were computed. For the amplitude of the bilabial closing gesture the correlation between ELITE and EMA data was $r = 0.985$, for movement velocity it was $r = .981$, and for movement time it was $r = 0.838$ ($p < .0001$ for all three correlations). The relatively low correlation of the durational measure is probably due to the limited sampling rate of the Elite system.

In order to assess the absolute deviation between the two systems the mean absolute difference between EMA and EILTE data was computed. It amounted to .51 mm. This deviation between the two systems consists of a systematic and a random component: The mean difference between ELITE data minus EMA data was .43 mm, i.e. movement amplitudes measured with the ELITE system were consistently larger than the respective EMA amplitudes. The systematic difference may be the result of three different sources: (1) The ELITE data are three-dimensional whereas EMA data are two-dimensional. Therefore, in case of a small movement component perpendicular to the midsagittal plane of the EMA helmet, the ELITE data are expected to be larger. (2) The lip sensor trajectories are not absolutely straight because of a rotational component of the mandibular movement. Since the ELITE reflectors were mounted on top of the EMA receiver coils, the movement radius of the reflectors and thus the length of the actually measured chord of the performed arc exceeds that of the EMA coils. (3) An additional sytematic component due to spatial distortion can not be ruled out.

After the EMA data were corrected for this systematic difference by simply adding .43 mm to all measured EMA amplitudes, the residual, nonsystematic error amounted to .33 mm. With respect to peak velocity, the mean absolute difference between the EMA and ELITE data was 8.1 mm/s, with respect to movement time it was 8.1 ms.

Table 1. Means of the kinematic data across the subject's repetitions

		Movement Amplitude (mm)		Movement time (ms)		peak velocity (mm/s)	
Rate	Vowel	ELITE	EMA	ELITE	EMA	ELITE	EMA
fast	a	8.9	8.3	85	86	188	179
	i	6.9	6.0	80	81	149	140
	u	4.6	4.4	95	95	86	83
	y	4.4	3.8	81	85	87	86
normal	a	11.1	10.6	113	119	192	187
	i	8.5	7.9	96	97	168	163
	u	4.6	4.4	104	110	83	72
	y	5.2	5.1	113	109	85	88
slow	a	8.6	8.1	125	124	149	148
	i	6.5	6.0	105	106	124	121
	u	4.1	3.8	129	128	57	59
	y	4.0	3.6	123	119	67	65

In order to determine whether the reliability of the acquired data is acceptable, a kind of 'signal-to-noise' ratio may be considered. Within the present experiment several sources of variability contributed to the observed movement data: (1) The 'interesting' effects to be tested, i.e. vowel categories and the speech rate conditions, (2) small variation in the speaker's behavior across repetitions, and (3) measurement error. The standard deviation across the means of the 12 (4 vowels x 3 rate conditions) test conditions was 2.37 mm in case of the ELITE and 2.25 mm in case of the EMA data. The standard deviation across the speaker's repetitions of identical test sentences, averaged across test conditions, amounted to .76 mm (ELITE) and .69 mm (EMA), respectively. This is about twice as high as the unsystematic difference of .33 mm between the two measurement systems. Therefore, measurement reliability may be regarded as sufficient in order to detect phonetic effects. If, however, these effects are small and thus difficult to test statistically, the experimenter should increase the number of repetitions and/or speakers rather than improve measurement reliability (because the measurement system is more reliable than the human speaker).

Finally, Table 1 shows the phonetic effects within the two data sets of the present study. In spite of some absolute deviances, the two measurement systems provide quite similar results with respect the relational structure within the data corpus. For example, with respect to movement amplitude the rank orders among both, vowel categories as well as speech rate conditions, are identical in the two data sets.

CONCLUSION

The results of the present study show that both methods of kinematic data acquisition, the electromagnetic articulography and the optoelectronic ELITE system provide comparable results with respect to the assessment of lip movements. The deviation between the two systems is about .5 mm, which can be expected considering previous studies of measurement reliability under laboratory conditions. Measurement differences between the two systems were considerably smaller than the variability across the speaker's repetitions of identical items.

REFERENCES

Borghese, N.A., and Ferrigno, G. (1990). An algorithm for 3-D automatic movement detection by means of standard TV cameras. *IEEE Transactions on Biomedical Engeneering,* **37**, 1221-1225.

Ferrigno, G., and Pedotti, A. (1985). ELITE: A digital dedicated hardware system for movement analysis via real-time TV signal processing. *IEEE Transactions on Biomedical Engeneering,* **32**, 943-950.

Gracco, V.L., and Nye, P.W. (1993). Magnetometry in speech articulation research: Some misadventures on the road to enlightenment. *Forschungsberichte des Instituts für Phonetik und sprachliche Kommunikation der Universität München,* **31**, 91-104.

Honda, M., and Kaburagi, T. (1993). Comparison of electromagnetic and ultrasonictechniques for monitoring tongue motions. *Forschungsberichte des*

Instituts für Phonetik und sprachliche Kommunikation der Universität München, **31**, 121-136.

Hoole, P. (1993). Methodological considerations in the use of Electromagnetic articulography in phonetic research. *Forschungsberichte des Instituts für Phonetik und sprachliche Kommunikation der Universität München*, **31**, 43-64.

Keller, E., and Ostry, D. (1983). Computer measurement of tongue dorsum movement with pulsed echo ultrasound. *Journal of the Acoustical Society of America*, **73**, 1309-1315.

Kelso, J.A.S., Vatikiotis-Bateson, E., Saltzman, E., and Kay, B. (1985). A qualitative analysis of reiterant speech production: phase portraits, kinematics, and dynamic modeling. *Journal of the Acoustical Society of America*, **77**, 266-280.

Kiritani, S. (1986). X-ray microbeam method for measurement of articulatory dynamics: techniques and results. *Speech Communication*, **5**, 119-140.

Magno-Caldognetto, E., Vagges, K., and Zmarich, C. (1995). Visible articulatory characteristics of the Italian stressed and unstressed vowels. *Proceedings of the 13th International Congress of Phonetic Sciences at Stockholm, Vol. 1*, 366-369.

Müller, E., and Abbs, J. (1979). Strain gauge transduction of lip and jaw motion in the midsagittal plane: Refinement of a prototype system. *Journal of the Acoustical Society of America*, **65**, 481-486.

Munhall, K.G. (1993). Kinematic analysis of speech using an optoelectronic measurement system. *Forschungsberichte des Instituts für Phonetik und sprachliche Kommunikation der Universität München*, **31**, 163-168.

Perkell, J.S., Cohen, M.H., Svirsky, M.A., Matthies, M.L., Garabieta, I., and Jackson, M.T.T. (1992). Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *Journal of the Acoustical Society of America*, **92**, 3078-3096.

Schönle, P.-W. (1988). *Elektromagnetische Artikulographie. Ein neues Verfahren zur klinischen Untersuchung der Sprechmotorik.* Springer: Heidelberg.

Schönle, P.W., Gräbe, K., Wenig, P., Höhne, J., Schrader, J., and Conrad, B. (1987). Electromagnetic articulography: use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language*, **31**, 26-35.

Schönle, P.W., Müller, C., and Wenig, P. (1989). Echtzeitanalyse von orofacialen Bewegungenn mit Hilfe der Elektromagnetischen Artikulographie. *Biomedizinische Technik,* **34**, 126-130.

Sock, R., Perrier, P., Bensaber, K., Bothorel, A., Brock, G., and Serignat, J.-F. (1995). Elaborating a multimedia platform for analyzing and exploiting speech cineradiographic data. *Paper presented at the ACCOR workshop on Articulatory Databases at Munich, May 1995.*

Sonies, B.C., Shawker, T.H., Hall, T.E., Gerber, L.H., and Leighton, S.B. (1981). Ultrasonic visualization of tongue motion during speech. *Journal of the Acoustical Society of America*, **70**, 683-686.

Stone, M. (1990). A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data. *Journal of the Acoustical Society of America*, **87**, 2207-2217.

Tuller, B., Shao, S., and Kelso, J.A.S. (1990). An evaluation of an alternating magnetic field device for monitoring tongue movements. *Journal of the Acoustical Society of America*, **88**, 674-679.

Westbury, J.R. (1994). *X-ray microbeam speech production database user's handbook*. Waisman Center on Mental Retardation & Human Development. University of Wisconsin Madison, WI.

Zmarich, C., Magno-Caldognetto, E., and Vagges, K. (1995). Variability in the articulatory kinematics of lips and jaw in repeated /pa/ and /ba/ sequences in Italian stutterers. *Proceedings of the 13th International Congress of Phonetic Sciences at Stockholm, Vol. 4*, 536-539.