Is markerless acquisition of speech production accurate ?

Slim Ouni, Sara Dahmani

Is markerless acquisition of speech production accurate ?

Slim Ouni
Université de Lorraine, LORIA, UMR7503, Villers-lès-Nancy, F-54600, France
Slim.Ouni@loria.fr

Sara Dahmani
Inria, Villers-lès-Nancy, F-54600, France
Sara.Dahmani@inria.fr

Abstract: In this study, the precision of markerless acquisition techniques have been assessed when used to acquire articulatory data for speech production studies. Two different markerless systems have been evaluated and compared to a marker-based one. The main finding is that both markerless systems provide reasonable result during normal speech and the quality is uneven during fast articulated speech. The quality of the data is dependent on the temporal resolution of the markerless system.

PACS numbers: 43.70.-h, 43.70.Jt

1. Introduction

In recent years, the techniques available for acquiring articulatory data have matured steadily. New developments have significantly improved both the quality and quantity of data that can be acquired, driving great advances in speech production research, and with increasing impact in related fields such as speech technology (8), clinical assessment of speech (13; 9), and language learning (5; 7). Electromagnetic Articulography (EMA)(10; 4), a prominent method to record articulation, captures movements in three dimensions (3D) with a high temporal resolution that is widely used by speech production community and different articulographs were recognized as accurate systems (14; 2; 11). Such a system allows tracking tiny sensors attached to speech articulators such as the tongue, teeth, and lips (13; 7; 9). Optical motion capture is also in active use (12; 6; 13). All these acquisition techniques are intrusive, and based on markers or sensors that are glued to the articulators. Some of them are very expensive. In some cases, it is difficult to use them with some specific populations as children. These factors may limit the widespread exploitation of articulatory data to several research fields. Compared to marker-based tracking, markerless techniques are less invasive, low cost and can be used in different environments. The technique is based on a markerless system, using mainly RGB and depth sensors, as Microsoft Kinect, RealSense, PrimeSense and Asus Xtion. It is possible to track the shape of the face, extract the shape of the mouth and get some information about speech articulation and facial expressions. It should be noted that several articulatory events take place within the vocal tract and thus are not visible and cannot be tracked by these markerless systems. Instead, they are appropriate to capturing visible articulators as the lips and the jaw. Our main concern is whether these markerless techniques are suitable for articulatory speech research, and whether they are well adapted to speech production studies. In this paper, we assess the precision of the markerless acquisition technique when used in spoken communication context.

2. Methods

We compare two markerless systems: PrimeSense Carmine and Intel RealSense. The main difference of the two systems is the frame rate (30 fps vs. 60 fps) and to some extent the depth sensor range of the camera. We use as a reference system a marker-based tracker with Vicon cameras. The Vicon system is widely used in the movie industry and also in facial animation and body gestures research. To be confident that it can be considered as a reference in this context, we first assess its accuracy by comparing it to an articulograph (AG501), which is a sensor-based tracking system with a high temporal resolution. We should note that it was not possible to use the articulograph directly as a reference instead of the Vicon system, as the wires of the sensors may introduce some interferences in the estimation of the depth map by the markerless system.

2.1. Setups

We have performed three comparisons : (1) articulograph vs. Vicon system ; (2) PrimeSense vs. Vicon and (3) RealSense vs. Vicon. For these comparisons we made three acquisitions where we have recorded a speaker uttering several sentences and sequences of phonemes. We have considered tracking 22 points on the face : 2 on the temple, 6 on the eyebrows, 2 on the nose, 2 on the cheek, 8 on the lips and 2 on the chin. Figure 1-A shows the positions of these points on a template.

Articulograph - Vicon In this acquisition, the sensors of the articulograph have been glued on the face of the speaker according to the configuration presented in Figure 1. The reflective markers for the Vicon of 3 mm in diameter have been glued on top of the sensors, as shown in Figure 1-B. We have used an articulograph AG501 with 24 sensors and at a sampling rate of 250 Hz. The articulograph software provides the 3D spatial position of each sensor. We have used a Vicon system based on 4 cameras (MX3+) using modified optics for near range. The cameras were placed at *approx.* 150 cm from the speaker. Vicon Nexus software provides the 3D spatial position of each reflective marker at a sampling rate of 100 Hz. Figure 1-B shows the positions of the sensors of the articulograph and the markers of the Vicon on the face of the speaker. Extra sensors and extra markers have been used to remove the head movement, and thus it is possible to compare the markers and the sensors in the same reference frame. The audio was acquired simultaneously with the spatial data using a unidirectional microphone.

PrimeSense - Vicon In this acquisition, the Vicon configuration is exactly the same as in the Articulograph-Vicon acquisition described above, where we have also used reflective markers. For the markerless technique we have used PrimeSense Carmine 1.09. This is a short-range markerless system well adapted to facial tracking. The depth sensor range is between 35 cm and 140 cm, the VGA depth map is 640x480 pixels. The frame rate is 30 FPS. The Primesense is placed at a distance of 70 cm from the speaker's head.

RealSense - Vicon The configuration in this acquisition is exactly the same as in PrimeSense-Vicon acquisition, where we have used Intel RealSense instead of PrimeSense. RealSense is also a short-range system. The depth sensor range is between 20 cm and 120 cm, the VGA depth map is 640x480 pixels. The frame rate is 50 FPS (which is different from the theoretical frame rate of 60 announced by the manufacturer). The RealSense is placed at a distance of 40 cm from the speaker's head.

2.2. Material and Procedure

We have recorded a set of 100 French sentences uttered as normally articulated speech by a male native speaker of French. We have also asked our speaker to repeat a sequence of phonemes starting at normal speed and progressively uttering the same sequence as fast as possible. The set of sequences were: (1) /ba/ for an example of an

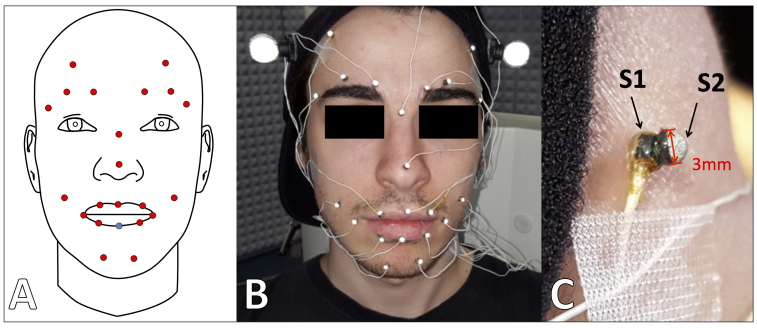

Fig. 1. (Color online) Positions of the 22 markers and sensors on a facial template (A) and on the human speaker face (B) ; (C) The Vicon marker (S2) is glued on the top of the articulograph sensor (S1).

open bilabial, (2) /ʃwa/ for an example of a protrusion and (3) /papipu/ for a bilabial gesture with three different vowels. For this recording the duration of /ba/ varied from .35 s to .09 s ; that of /ʃwa/ from .52 s to .15 s ; and that of /papipu/ from 1.165 s to .33 s. To synchronize each pair of modalities, we have used an in-house electronic device that triggers simultaneously a led lamp, an infrared lamp and high-pitched sound generated by a piezoelectric buzzer. One or more signal will be captured at least by one system (articulograph, Vicon, PrimeSense, RealSense). The synchronization signal for the articulograph and PrimeSense/RealSense is an acoustic signal, and for the Vicon, is the infrared light.

2.3. Data processing

Each pair of modalities has been synchronized using the appropriate signal for each one. As a result, the different streams for each acquisition were synchronous. For the marker-based systems, the spatial data were obtained directly as 3D coordinates for each marker or sensor. Both the Vicon and the articulograph, have a proprietary software that process the data. The spatial data obtained from each system have been merged together into the same reference using successive spatial transformations, in such a way, the marker and the sensor positions of the first frame are identical. For the markerless systems, we have used the commercial software Faceshift (3) that supports both PrimeSense and RealSense. Faceshift is a 3D motion capture software based on a markerless system. First, it adapts a generic facial model to the human speaker face, and then it is possible to reproduce facial gestures of the speaker. Faceshift provides as a post-processing step a fine tracking and manual correction. Virtual markers can be attached to the 3D mesh and then can be extracted as 3D spatial data. The main idea of using Faceshift is to extract the data from both markerless systems using the same tracking method, to focus on the evaluation of the systems themselves and thus the evaluation will be valid. The virtual markers that have been extracted are those that best match the Vicon marker positions.

3. Results

To compare the quality, we have used the EMA as a reference to evaluate the Vicon system, and the Vicon as a reference to evaluate PrimeSense and RealSense systems. To compare the systems in consideration to the reference trajectories, we have used the root-mean-square error (**RMSE**) and the cross-correlation (**CC**) with a lag of zero (measure of similarity). The RMSE and CC have been computed for every marker and sensor and an overall means have been computed. As each system in

consideration has different sampling rate, for each pair of systems, we have down-sampled the higher sampling rate of the two, and we used it in computation. In the setup Articulograph-Vicon, the overall RMSE and CC on the different axes were as follows : X-axis (RMSE=.27 mm, CC=.91) ; Y-axis (RMSE=.38 mm, CC=.99), and Z-axis (RMSE=.29 mm, CC=.98), For each group. The RMSE is thus very low and both systems provide trajectories that are highly correlated. This shows clearly that it is reasonable to consider the Vicon system as very accurate compared to the articulograph as it is within the same range of precision as the AG501 (11). The results for the two other setups (Vicon-PrimeSense and Vicon-RealSense) are detailed in Table 1. We present the RMSE and CC results for each axis (X, Y and Z), for PrimeSense (**PS**) and for RealSense (**RS**) compared to Vicon. The results were grouped as follows: (1) normal speech: when uttering the 100 sentences at normal speaking rate; (2) Protrusion repetition: when uttering the protruded /ʃwa/ ; (3) Bilabial-open repetition: when uttering the open bilabial /ba/ ; (4) Bilabial repetition: when uttering the bilabial sequence (/papipu/). The repetitions were performed at different speaking rates (from normal to very fast, as described in *section* 2.2). We only present the results for the lower part of the face (makers on the lips and the chin). In fact, as the uttered speech was not expressive, the upper part of the face is hardly moving. For normal speech the RMSEs of both RS and PS were barely above 1 mm, but RS has slightly lower RMSE, which is a good result. During speech repetition, almost all the results have RMSE below 2 mm. For protrusion, the overall errors were higher than 1.5 mm, on the three axes for both PS and RS. However the RMSE of RS were lower than that of PS. The main movements during protrusion were on the Z-axis (protruding/advancing the lips) and Y-axis (vertical movement). For bilabial-open/bilabial repetitions, the main movement was on the Y-axis. We note that during repetitions, the highest RMSEs were on the Y axis. RMSE on the Z-axis was overall lower for RS than for PS. It has been shown in (1) that the depth measurement accuracy of the PS (the Z-axis in our setup) is not constant and presents some measurement errors, which is confirmed by the finding of the current study. When examining the correlation, we note that for the X-axis, PS and RS present almost no correlation with the reference trajectories. This can be explained, as above, that the major movements were on the Y- and Z-axes, the overall movements on the X-axis can be considered small but present more inherent differences between the trajectories. For the Y- and Z-axes, RS has a higher CC than PS. However, both systems presented lower correlation when compared to the first setup (Articulograph-Vicon comparison). The CC results varied from .72 to .86 for Y-axis (compared to .99), and from .80 to .89 for Z-axis(compared to .98). The effect of the correlation can be illustrated in Figure 2, where the trajectories of a marker in the middle of the lower lip (the blue marker on the template presented in Figure 1-A) on Y- and Z-axes, for the protrusion repetition and for bilabial-open repetition, were shown. The result is presented for both PS and RS, and compared to that of the Vicon.

When the speaking rate increases, we can see clearly that PS is failing to track the fast gestures. Nevertheless, RS presented overall higher performances and to a lesser extent for Z-axis trajectories.

4. Discussion

For both PS and RS the overall RMSE for the different axes is in a relatively low range, with better performances for RS. Although both systems provided a lower correlation compared to Articulograph-Vicon comparison, the CC is higher for RS than that of PS. The main difference between PS and RS is the acquisition sampling rate (30 Hz for PS and 50 Hz for RealSense). The sampling rate has an important impact in the case of fast articulation. In fact, PS failed to track articulatory movements. However, RS seems to be capable of keeping track of fast movement, but not as good as Vicon when compared to the articulograph. We should note that in this study we are comparing

Table 1. RMSEs and CC for PS and RS, for each axis, and for normal speech, variable speaking rate repetitions of protrusion, bilabial opening and bilabial sequences.

Type	System	RMSE X	RMSE Y	RMSE Z	CCX	CCY	CCZ
Normal Speech	PS	1.293	1.196	1.190	-.089	.785	.813
	RS	1.269	1.071	1.091	.076	.832	.884
Protrusion Rep.	PS	1.972	2.035	1.875	.028	.718	.805
	RS	1.560	1.854	1.533	.208	.866	.893
Bilabial-open Rep.	PS	1.671	1.706	1.138	-.125	.721	.883
	RS	1.110	1.554	.950	.092	.844	.887
Bilabial Rep.	PS	1.189	1.421	1.185	-.039	.832	.809
	RS	1.260	1.750	.884	-.107	.809	.851

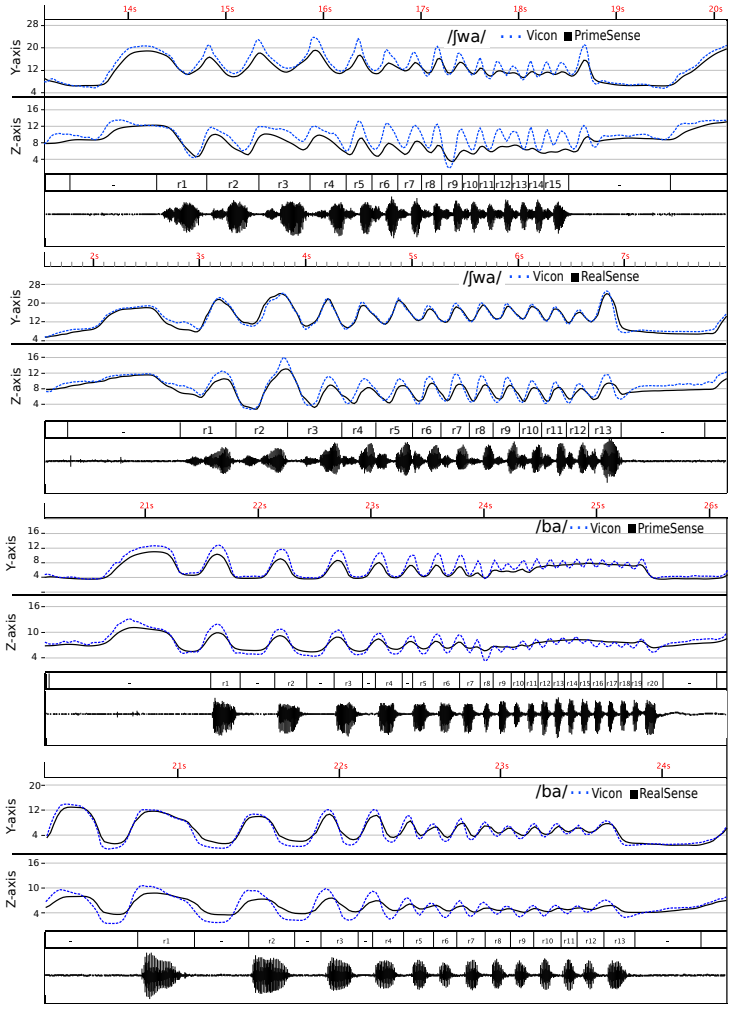

Fig. 2. (Color online) Articulatory trajectories of the lower lip during /ʃwa/ and /ba/ repetition from normal to fast speaking rate. Each graph presents a comparison between Markerless (PS and RS) and marker-based (Vicon) trajectories.

the PS and RS systems and not the tracking method that is in use. It might be possible that other tracking methods provide better results but, the improvement should be beneficial very likely for both systems, and this will not question the finding of this article. This is the main difference between marker-based and markerless techniques. In fact, marker-based techniques provide a direct measurable spatial data, whereas markerless methods need to implement a mathematical or algorithmic model using machine learning techniques to retrieve measurable data. Naturally, the quality of the acquisition data is also highly dependent on the used tracking model.

Speech production can significantly benefit from the constant progress in markerless technique development, to improve both the quality and quantity of visual articulatory data and can widen its application to a variety of fields. Based on the results of our study, we found that both markerless systems provide overall good performances. However, we cannot recommend using the current available markerless technologies, and more specifically those with sampling rate below 50 Hz for fine modeling lip movement for speech production simulation purposes or fine articulatory synthesis (as a physical simulation). Furthermore, markerless techniques have a major limitation that they do not track the inner lips, but track only the outer contour, which may omit important information as complete closure of the lips. A technique based on the articulograph can overcome this and also it is possible to track the vocal tract movement, presented mainly by the tongue. Nevertheless, the markerless techniques can be perfectly suitable for tracking facial expressions as in emotions or laughing, or affect, as these movements have relatively slow variation in comparison to speech articulation.

Acknowledgments

This work was supported (in part) by the EQUIPEX Ortolang, Inria (ADT Plavis), Region Lorraine (Corexp).

References and links

[1] A. Bandini, S. Ouni, P. Cosi, S. Orlandi, and C. Manfredi,"Accuracy of a markerless acquisition technique for studying speech articulators", Interspeech 2015, Dresden, Germany. Sept. 2015

[2] J. J. Berry, "Accuracy of the NDI Wave Speech Research System". J Speech Lang Hear Res, 54(5), 1295-1301, (2011)

[3] Faceshift - http://www.faceshift.com. - Faceshift is not available anymore since September 2015.

[4] P. Hoole and A. Zierdt, "Five-dimensional articulography", Speech motor control: New developments in basic and applied research, 331-349, (2010).

[5] Y. Iribe, S. Manosavan, K. Katsurada, R. Hayashi, Z. Chunyue and T. Nitta, "Improvement of animated articulatory gesture extracted from speech for pronunciation training," in Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp.5133-5136, 25-30, (2012).

[6] J. Jiang , A. Alwan, P. A. Keating, E. T. Auer and L. E. Bernstein, "On the Relationship between Face Movements, Tongue Movements, and Speech Acoustics", EURASIP Journal on Advances in Signal Processing, 2002(11), p 1-15, (2002)

[7] W.F. Katz, T.F. Campbell, J. Wang, E. Farrar, J.C. Eubanks, A. Balasubramanian, B. Prabhakaran, and R. Rennaker, "Opti-speech: A real-time, 3D visual feedback system for speech training", Interspeech'2014, Singapore, (2014).

[8] ZH Ling, K Richmond and J Yamagishi, "Articulatory control of HMM-based parametric speech synthesis using feature-space-switched multiple regression", Audio, Speech, and Language Processing, IEEE Transactions on 21 (1), 207-219, (2012).

[9] A. Mefferd, "Articulatory-to-Acoustic Relations in Talkers With Dysarthria: A First Analysis". J Speech Lang Hear Res, 58(3), 576-589, (2015).

[10] J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, and M. T. Jackson, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements", The Journal of the Acoustical Society of America, 92(6), 3078-3096, (1992).

[11] M. Stella, A. Stella, F. Sigona, P. Bernardini, M. Grimaldi and B. Gili Fivela, "Electromagnetic Articulography with AG500 and AG501", in Interspeech'2013, Lyon, 13161320, (2013).

[12] E. Vatikiotis-Bateson, K. Munhall, and D. Ostry, "Optoelectronic measurement of orofacial motions during speech production", The Journal of the Acoustical Society of America, 93, 2414-2414, (1993)

[13] B. Walsh and A. Smith, "Basic parameters of articulatory movements and acoustics in individuals with Parkinsons disease, Movement Disorders, vol. 27, no. 7, pp. 843-850, (2012).

[14] Y. Yunusova, J. R. Green and A. Mefferd, "Accuracy Assessment for AG500, Electromagnetic Articulograph". Journal of Speech, Language, and Hearing Research: JSLHR, 52(2), 547555, (2009)