The stability of mouth movements for multiple talkers over multiple sessions

Chris Davis 1, Jeesun Kim 1, Vincent Aubanel, Greg Zelic, Yatin Mahajan

1 The MARCS Institute, University of Western Sydney, Australia
Chris.davis; j.kim; v.aubanel; g.zelic; y.mahajan@uws.edu.au

Abstract
To examine the stability of visible speech articulation (a potentially useful biometric) we examined the degree of similarity of a speaker’s mouth movements when uttering the same sentence on six different occasions. We tested four speakers of differing language background and compared within- and across speaker variability. We obtained mouth motion data using an inexpensive 3D close range sensor and commercial face motion capture software. These data were exported as c3d files and the analysis was based on guided principal components derived from custom Matlab scripts. We showed that within-speaker repetitions were more similar than between speaker ones; that language background did not affect the stability of the utterances and that the patterns of articulation from different speakers were relatively distinctive.

Index Terms: biometrics, speech articulation, mouth motion; 3D face motion capture

1. Introduction
Biometric person recognition utilizes an intrinsic or behavioral trait of a person to allow access to a closed system, or for monitoring purposes. Biometric identification allows access to a system without the need for a password, physical key, identification card, and so on. Typical biometric markers involve measuring an aspect of a person’s physiology (e.g., finger-prints). Although such physiological markers can facilitate precise identification, typically they require physical contact with an input device; as such their use is more suited to handheld devices (e.g., mobile phones) and may be intrusive in other situations.

In this regard, some sources of biometric information offer the potential for unobtrusive recognition and have enormous potential for security applications. Indeed, behaviourally-based biometrics (e.g., voice, gait recognition) are less intrusive and more covert than physiologically-based ones. However, because such behavioral biometrics are dynamic, they are often less robust than stable physiological ones [1].

To be effective, a biometric marker generally needs to have the following properties: universality, distinctiveness, and permanence, while being easy to collect and socially acceptable. Identification based upon behavioural information obtained via a single modality tends to be fragile since the signal can be distorted (e.g., background noise, reverberation, uneven lighting, etc.). A standard proposal to mitigate this problem is to collect data from multiple modalities. The assumption is that limitations in unimodal based biometric measures can be offset by using information from a modality providing complementary information. This probably best describes how we ourselves identify people in our everyday life where multiple redundant cues are often available and used.

In regards to person recognition or verification, two signals have attracted the bulk of attention: the face and the voice. Although most research on face recognition has focused on static images, here we examine the dynamic visual speech signal. Ideally, a speech based biometric would employ both auditory and visual speech information. Indeed, examining the combination of these two signals offers considerable opportunity for biometric information that may be quite person specific. However, to our knowledge, some basic properties of the suitability of the visual speech signal in terms of biometrics have not been established. For example, how consistent are a talker’s visible articulatory movements over multiple renditions? How distinctive is one talker’s speech articulation from others. These are important questions, because a key issue confronting the use of dynamic behavioral signals as biometrics concerns their stability (in the sense of whether properties of the signal are consistent when these have been collected at different occasions) and their distinctiveness. On a practical level, another issue concerns how cheaply a signal can be acquired.

To investigate the stability of visual speech articulation over multiple sessions, we collected head, face, lips and jaw motion from four talkers saying the same ten sentences in six separate sessions (over a period of two weeks). These four speakers were selected to contrast English language experience and first language (see Participant section below). The motivation for this was to determine whether English language experience and language background would affect the stability of uttering English language sentences.

To address the issue of the cost, we used a commercially available inexpensive 3D infra-red scanner and commercial 3D facial motion capture software (Faceshift®). To make the reporting of the data more manageable, the current paper reports the stability of only one component of the visual speech motion signal: mouth motion when a talker pronounces a sentence. We used both time- and amplitude-based methods to assess the consistency of articulation. This was done by parametrically comparing the utterance of each session with the other sessions, and by time normalizing renditions, creating a motion mean, and determining the degree to which each session’s utterances deviated from this.

2. Method

2.1. Participants
Four male speakers (aged 32, 32, 34 and 55 years) took part in the recording sessions. The speakers had different first
languages: Sp_1: Australian English; Sp_2 & 3 French and Sp_4 Punjabi. All currently reside in Australia. The Punjabi speaker had learned English at an early age (this was the language of school instruction) and has resided in Australia for over six years; the French speakers began learning English at age 14 (for one, Sp_2 it was not the primary foreign language learned) and have lived in Australia for the last two years.

2.2. Equipment

2.2.1. Image/motion capture

3D data was captured and constructed using a Carmine 1.09 close range sensor (0.35m - 1.4m). The spatial x/y Resolution 640 x 480 (VGA) (2-Sigma Values) at 0.5m = 0.9 mm; the depth Resolution (2-Sigma Values) at 0.5m = 0.1 cm; the depth Image Field-of-View: Horizontal at 0.5 m = 53.6 degrees; Vertical = 45 degrees. In addition, color Image sequences were captured at 640 x 480 (VGA).

2.2.2. Image/motion registration and processing

Faceshift Studio® 2014 facial motion software [2] was used to register and process the 3D sensor data (see procedure below). Auditory speech was recorded using this software from a AKG C417 PP professional lavalier microphone input to a Roland Duo capture EX soundcard.

2.2.3. Recording

The recording sessions took place in a typical office that was lit with two Bowens UNI-LITE BW3370 flood fill lights (with semi opaque diffusers).

2.3. Materials

The materials were 10 sentences selected from the Harvard list of phonetically balanced sentences (IEEE 1969 [3]). These sentences were six to nine words in length (M = 7.3; SD. 0:94) and consisted of descriptions of fairly mundane events (e.g., “The big red apple fell to the ground”).

2.4. Procedure

2.4.1. Faceshift preparation

Faceshift requires that custom specific expression model for each individual. This model consists of 51 blend-shapes that are captured as a person produces different training postures. Each of the four participants created a model by modelling 23 face postures (Neutral; Open mouth; Smile; Brows down; Brows up; Sneer; Jaw left; Jaw right; Jaw front; Mouth left; Mouth right; Dimple smile; Chin raise; Kiss; Mouth funnel; Frown; mouth M-shape; Check puff; Chew; Mouth press; Mouth stretch; Lower lip down; Upper lip up).

2.4.2. Recording

Each speaker was recorded individually. Speakers were seated in a quiet room with the Carmine 1.09 close range sensor positioned directly in front at face level and at approximately 0.6 m distance (see Figure 1). Speakers were given a written list of the 10 sentences which they learned by heart. Speakers controlled the Faceshift acquisition themselves, looking directly at the camera without shifting their pose, uttering the sentence aloud and then stopping the capture with a mouse click. This procedure was repeated until each of the 10 sentences had been recited (along with two ‘wag’ trials to establish the centre of head rotation).

Figure 1: A schematic depiction of the data capture procedure.

2.4.3. Data Processing

The quantification of speech related articulatory movements: Faceshift uses an input device (here a Carmine 1.09 close range sensor) to construct a depth map by analysing a speckle pattern of infrared laser light. Virtual marker positions can be tagged to this depthmap and used to parameterize motion. Here we exported the FaceRobot® virtual marker set in c3d format and selected a subset of markers to use in a data reduction process (see left panel, Figure 2).

Given the high dimensionality of the recorded corpus, dimensionality reduction was then performed. Guided principal component analysis (gPCA, [4]) was used. This style of PCA employs linear decomposition to extract a set of a priori defined components representing biomechanically plausible articulatory control parameters (six components are typically sufficient to explain the majority of articulatory data [5]).

Figure 2: A depiction of the virtual marker positions exported in c3d format from Faceshift and the subset of markers used for the gPCA (shown in red in the left panel). The right panel shows the tool used to confirm the gPC 3D reconstruction (moving the sliders, bottom right, showed the influence of each PC).
The shape-normalised motion data was processed using gPCA to reduce the dimensionality of the data set to eight non-rigid components, along with three rigid translations and three rigid rotations (pitch, roll and yaw) of the whole head. To minimise the overrepresentation of particular marker configurations (e.g., the neutral position at the start and end of each utterance), a database of unique movements was generated. Using the ‘wag’ trials, the six rigid motion parameters around the estimated centre of rotation were determined (using the quaternion method [6]) and extracted from the database. The remaining non-rigid movements were then analysed applying gPCA. The gPCA solutions were inspected in 3D space using a tool in which the influence of each PCA was visualized (see right panel, Figure 2). Following inspection, the gPCA parameters were output as vectors of the contribution of a gPCA per frame (time). Note, as mentioned above, the current paper only reports mouth motion.

3. Results

We quantified the degree of similarity of the mouth movements when uttering that same sentence on different occasions in two ways. 1) To examine temporal changes, we used Dynamic Time Warping (DTW) [7]. 2) To examine amplitude differences we normalized the durations and measured variation of the motion from the mean.

1) Dynamic time warping (DTW) is a procedure that provides a measure of comparison between series of data points (inherent distance). For example, DTW can expand or compress one time series to resemble another one and by summing the distances of individually aligned elements an inherent distance between the two can be computed (Figure 3).

![Figure 3: An example of the time-series of the contribution of the Mouth guided principal component when uttering “soap can wash most dirt away” for sessions 1 and 4 and the DTW.](image)

2) The second method we used for quantifying utterance stability was to examine differences in the amplitude of the mouth principal component while normalizing for time. This was done using the following procedure:

1) To normalize the duration of all utterances of each sentence to a fixed duration of 100 frames (by linear interpolation).
2) To construct an average for each sentence and each speaker.
3) To calculate the area of a one standard deviation (SD) ribbon about the mean. This latter value was then used as an index of variability.

A summary of these data is shown in Figure 5 for each speaker (same order as in Figure 4) and for each sentence. The red centre curve in each panel shows the mean value (collapsed over the six sessions) for the interpolated frame. The blue ribbon represents 1 SD around this mean.

The within-speaker warping cost data was averaged and compared (paired t-test) with the between-speaker cost; there was a significant difference, t(149) = 27.5, p < 0.000 (two tailed).

2) The second method we used for quantifying utterance stability was to examine differences in the amplitude of the mouth principal component while normalizing for time. This was done using the following procedure:

1) To normalize the duration of all utterances of each sentence to a fixed duration of 100 frames (by linear interpolation).
2) To construct an average for each sentence and each speaker.
3) To calculate the area of a one standard deviation (SD) ribbon about the mean. This latter value was then used as an index of variability.

A summary of these data is shown in Figure 5 for each speaker (same order as in Figure 4) and for each sentence. The red centre curve in each panel shows the mean value (collapsed over the six sessions) for the interpolated frame. The blue ribbon represents 1 SD around this mean.
Figure 5: A representation of the variability in the magnitude of the contribution of the Mouth PC for each speaker and each sentence. The centre red curve represents the mean PC value for that speaker averaged over the six capture sessions. The blue ribbon shows 1 SD around this mean.

Figure 6 shows the mean area of the blue ribbon (Figure 5) for each of the speakers.

Figure 6: Average area of 1 SD around the mean of the contribution of the time normalized Mouth PC for the four speakers

It is clear from Figures 4-6 that the mouth motion data for the non-native speakers reciting English sentences was just as stable as that of the native speaker. That is, even though these speakers have a non-native accent, the articulation of their non-native speech is consistent.

4. Discussion

The question of how consistent people are in the way that they articulate speech (in terms of what can be measured from the visible articulators) is an important one for visual speech based biometrics. Here, in a first approach to this issue, we used an inexpensive 3D camera-and-software-based measure of mouth articulation (parametized by gPCA) to examine the reliability for four speakers across six capture sessions.

We examined two measures of consistency, DTW cost, in which the stability of the contribution of the Mouth PC over time was determined by parametrically contrasting the cost of warping the Mouth PC of each of six utterances per speaker per sentence with each other (within-speaker cost). This data was then compared with the between-speaker warping cost. This showed that an individual was more consistent with themselves than with other speakers.

The second method used involved normalizing for time, and comparing the Mouth PC trajectories against a mean speaker value for each sentence. These data showed that speakers have relatively consistent trajectories across sessions and that the morphology of these differed across the talkers. This difference holds promise that visible speech motion may be useful in discriminating speakers (although this pattern may be somewhat of a product of the different language backgrounds of the speakers – itself an interesting possibility).

The current study only reported mouth motion (something constrained by the utterance), we are currently examining other visible speech related articulations (jaw, eyebrow, rigid head motion) that may allow for more idiosyncratic motion.

5. Acknowledgements

The authors acknowledge the support of career development post-doctoral bursaries to authors four and five and an ARC discovery grant (DP130104447) to the first two authors supporting author three.

6. References