Target Practice on Talking Faces

Adriano Vilela Barbosa*, Andreas Daffertshofer†, Eric Vatikiotis-Bateson‡

* CEFALA, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
† Psychology Department, Vrija Universitat, Amsterdam, The Netherlands
‡ Linguistics Department, University of British Columbia, Vancouver, Canada
§ Vision Dynamics Department, ATR Human Information Science Labs, Kyoto, Japan
(adriano.vilela@gmx.net, marlow@fbw.vu.nl, bateson@atr.jp)

Abstract

A method is described for video-based tracking and analysis of talking heads. Although dots on the face and head are tracked now, clean faces will be tracked once a satisfactory optical flow technique is found. The analysis consists of decomposing the head and face motion components into a common mode and the residual, thus eliminating the need for complex marker tracking and head correction of face motion.

1. Introduction

The analysis and synthesis of talking faces has developed rapidly in recent years. In the entertainment industry, talking faces are being synthesized at near video quality using various combinations of art and algorithm. Unfortunately, cosmetically acceptable animated faces do not necessarily convey sufficient linguistic information to be communicatively acceptable. By contrast, in some research labs (e.g., ATR in Japan, Queens University in Canada, the ICP in France), the communicative realism of talking faces has been more important than their cosmetic quality [1]. Indeed, despite obvious departures from video realism such as lack of important features (e.g., eyes, tongue, skin features), talking faces whose communicative realism has been validated can be watched face-on almost indefinitely (e.g., Massaro and Cohen’s Baldi [2]; Kuratate and colleagues’ data-driven faces [3]).

In addition to communicative realism, we have insisted that the linguistically informative talking faces we use in our research be generated from time-varying measures of multimodal speech production. Realistic output from measured input together provide an experimentally controllable link between the realms of the production and perception of multimodal speech that enables us to determine what in the stimulus source engenders specific aspects of perceiver responses. An example of the experimental manipulation of animation parameters is the recent demonstration by members of our research team that natural head motion augments the intelligibility of talking face animations which either lack or contain incorrect head motion [1].

2. Head and face motion

That head and face motion correspond to such different aspects of speech production yet both contribute, perhaps seamlessly, to multimodal speech processing is non-trivial and raises questions that this paper attempts to clarify and address. In earlier work at ATR, head motion and fundamental frequency (F0) were found to be highly correlated, while face motion was found to be strongly related to the spectral acoustics [4]. The distinctiveness of the head and face motion in their relations with vocal source and vocal tract filter, respectively, suggest that perceivers may simultaneously process quite different temporal characteristics: those corresponding to head motion and intonation/prosody; and those corresponding to the phonetic identification of much smaller contrastive segments (e.g., phonemes, gestures, syllables). Considered from the perspective of the viewing listener, facial motions of the mouth, the chin, and cheek areas correspond to phonetic details while motions of the head provide information about the overall temporal structure of the speech utterance as well as showing a fine-grained temporal correspondence with F0.

On the other hand, the distinct analytic mappings observed in multimodal production between visible motions of the head and face and the source (head) and filter (face) components of the acoustics do not imply that perceivers necessarily recover a similarly componential visual contribution to speech. It is an open question whether or not perceivers integrate head and face components in this way; and if they do, we need to determine how they do it. This latter question is particularly interesting when we consider that head motions are generally larger than facial motions and make a substantial contribution to the overall motion of the facial features. That is, perceivers see facial motions comprised of both head and facial components.

It may be that the analytic distinctiveness of production correlates is not mirrored in the perceptual processing. Rather, the increased visual enhancement to speech
intelligibility that is observed when head motion is added to animated talking faces, may be simply a boost in the visual gain. This interpretation is consistent with the finding that perception of acoustically degraded speech is enhanced even when the visual signal itself is severely degraded via spatial and/or temporal filtering [5, 6]. That is, the combination of facial and head motions sums to a robust visual correlate of the speech event. This occurs despite the apparent sub-threshold amplitudes of face motion – the primary source of the visual gain. This is shown by the finding that head motion alone provides no visual enhancement to speech intelligibility – e.g., when applied to static faces and abstract face-like objects (work in progress at ATR).

3. Head and face motion measures

With the preceding discussion as background, we now turn to the discussion of a new method for reliable recovery of head and face motion during speech from video sequences of talking heads. In this system, motions of the face and head are separated into motions specific to orientation and position of the entire head and motions corresponding to speech and other facial gestures through de-correlation of two principal modalities. Our final goal is to recover linguistically relevant motions of the head and face from video sequences of clean faces (i.e., no markers).

In this system, face and head motion are represented by pixel trajectories derived from optical flow analysis (OFA) of image sequences and subsequent dimension reduction using principal component analysis (PCA). The multilinear decomposition afforded by PCA provides a principal mode of motion common to the head and a residual containing all other motion and noise. At present, the clean face motion recovery is not yet performing well. This is due in part to the difficulty of extracting optical flow data that are both accurate and still manageable computationally. In principal, however, additional modes can be identified – e.g., to distinguish emotive and speech gestures – using PCA and/or independent component analysis (ICA). In what follows, a brief description of the analysis components is provided and then the procedure is demonstrated for data in which blue dots placed on the speaker’s face and head were tracked.

Briefly, optical flow is the vector calculation of image differences in position and orientation from one image to the next in a sequence [7]. The pixel-based changes in chrominance, luminance, etc. from one video frame (field) to the next correspond to motion of the object (e.g., a talking face) being recorded and the interaction
between the changed orientation and position of object surfaces and the lighting source(s). Although OFA has been around for a long time now, previously we used an iterative wavelet based approach [8] because OFA did not provide reliable tracking when motion changed direction. Both wavelet and optical flow methods are computationally expensive at high resolution; however, optical flow is non-iterative and therefore less expensive, and it affords easier identification of regions of interest (ROI) that can greatly reduce the size of the calculation.

Methodologically, a physically non-invasive technique such as this will greatly expand the applicability and naturalness of speech-related motion capture in laboratory, clinical, and field settings. The technique also affords investigation of the issues raised in the previous section. Specifically, by analyzing head and face motion as a single behavior composed of multiple computable modes, we hope to improve our understanding of how perceivers are able to retrieve linguistically relevant information from degraded visual signals.

3.1. Two-dimensional motion recovery from dots

The method used to recover the 2-D position of dots placed on the face was developed by Barbosa [9]. To find the dots shown in Fig. 1 requires (i) defining rectangular search regions for each marker and (ii) finding the marker within the search region. Search regions are defined by building a sequential series of triangles connecting two known (reference) dots with a 3rd unknown dot, beginning with two base markers (black markers in Fig. 2) whose positions are already determined to be maximally stable. The location of the unknown dot is found by first finding the center of its rectangular search region relative to the two reference dots, and then finding the position of the dot with the search region. Dot search is done using the linear combination of the two chrominance channels (YCbCr format). Chrominance values within the search region are summed for small rectangular samples. The computed sums for each rectangular sample and its distance from the center are linearly combined to make a cost function; dot location is determined by the sample that minimizes the cost function – see Fig. 3. The recovered motion paths for the dot tracking are shown in Fig. 4.

3.2. Decomposition of head and face motion

Face and head motion are separated by a sequence of steps beginning with PCA of the computed changes in dot location from one frame to the next. The resulting eigenvalues are then grouped by sign (+/-) of the position change horizontally and vertically. Same sign components are then summed for the horizontal and vertical dimension to recover the common mode of motion – i.e., for the head. The residual contains everything else including motion of the face and any error from the head motion extraction.

An example of this analysis is shown in Fig. 6. The data analyzed were high quality (Sony Digital Betacam) video recordings of spontaneous monologue and scripted
sentence utterances produced in a dialect of French spoken in western Canada (Alberta). For this subject, PCA decomposition into one primary mode and the residual generated head motion measures with 1/2 pixel (roughly .5 mm) accuracy. The video-based measures were validated kinematically by comparing the positions of OPTOTRAK irleds (infrared LEDs) placed on the head (not shown in Fig. 6) as estimated with the dot tracking procedure and the OPTOTRAK’s measurement of 3D position projected to the image plane. Fig. 5 compares horizontal and vertical motion computed from video and OPTOTRAK. We conclude from this that the residual is composed primarily of face motion independent of the head and constitutes a simple but effective way to separate different components of speech-correlated motion.

4. Discussion

The video-based motion recovery system will be used to track clean faces as soon as a robust optical flow method is implemented. Our hope is that the decoupled head and face motion for clean faces will still be as accurate as the current method, which has been sufficient for mapping face and head motion to acoustics [9]. In addition, we hope to be able to decouple additional facial modes – in particular, linguistic and non-linguistic facial expressions. Finally, we expect the recovered motion values to be sufficient to generate communicatively realistic animated faces for use in perception studies.

5. Acknowledgements

This project was supported by a TAO grant to ATR HIS Labs. Thanks are also due to Katsunori Shimohara and Sumio Yano of ATR for their continued encouragement.

6. References


