AUDITORY-VISUAL SPEECH ANALYSIS: IN SEARCH OF A THEORY

Christian Kroos

MARCS Auditory Laboratories, University of Western Sydney, Australia
c.kroos@uws.edu.au

ABSTRACT

In the last decade auditory-visual speech analysis has benefited greatly from advances in face motion measurement technology. Devices and systems have become widespread, more versatile, easier to use and cheaper. Statistical methods to handle multichannel data returned by the face motion measurements are readily available. However, no comprehensive theory or, minimally common framework to guide auditory-visual speech analysis has emerged. In this paper it is proposed that Articulatory Phonology [3] developed by Browman and Goldstein for auditory-articulatory speech production is capable of filling the gap. Benefits and problems are discussed.

Keywords: Auditory-visual speech, face motion analysis, articulatory phonology

1. INTRODUCTION

When in the early 1960s Charles Hockett attempted to determine the universal "design features" of human language [9], he listed as the first feature: *Vocal-auditory channel*. Research on sign languages, and an increased awareness of a sociocultural bias in favour of spoken language in the past, made it clear that this definition needed to be revised in contrast to the 12 or 15 other features that have retained their validity to the present day. As sign languages have presumably been around for at least as long as vocal languages, and some researchers have even argued that they were the precursors of vocal languages (e.g., [1]), they should have been easily observed and studied, in principle. However, the visual aspects of spoken language are much more difficult to identify as a genuine part of human speech processing. Though "lip reading" abilities of the hearing impaired or deaf were well-known and acknowledged, it is the tight link to auditory speech that remained unnoticed for a long time. That is, speech reading was classified as a special ability that developed if the acoustic signal was permanently unavailable.

Besides the greater difficulty in observing the role that visual speech plays in the complex speech production/perception process there is another crucial difference to the study of other forms of visual communication, like sign language: visual speech can arguably neither be studied nor understood without the reference to auditory speech, that is, without reference to speech vocal tract behaviour or, according to other accounts, the acoustic signal. For the most part visual and auditory speech are intimately linked through their common origin in vocal tract behaviour. Figure 1 shows the three-way relationship among vocal tract behavior, face motion and the acoustic output. This paper deals with what appears to be the primary dichotomy in the development of auditory-visual speech analysis in recent years:

- on the one hand, there have been impressive advances concerning face motion (and also articulatory) measurement techniques,
- on the other hand, the field is plagued by the lack of any comprehensive theoretical foundation or, minimally, at least a common framework to guide auditory-visual speech analysis.

As a consequence, the results from the research resemble more a patchwork than a gradually advancing research undertaking.

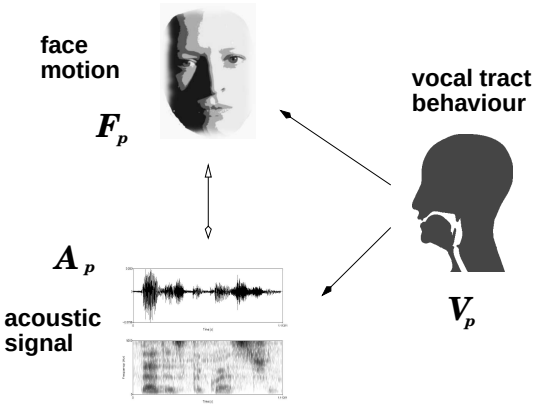
2. ADVANCES IN MEASUREMENT OF FACE MOTION

The rapid development and commercial production of ever more powerful integrated circuits, in general, and CCD and CMOS imaging chips, in particular, has changed the landscape in motion tracking. Hardware components for optical "mocap" systems became affordable as they were mass-produced for the consumer market, and the necessary computationally intense processing of the raw data became manageable as general purpose and specialized DSP chips became cheaper. A brief and by no means comprehensive review is given in the following.

2.1. Marker-based systems

Marker-based systems can be divided into two groups according to whether they use active or passive markers. Active markers are usually LEDs that emit infrared light pulses at a characteristic

Figure 1: Relations among vocal tract behaviour, face motion, and acoustic signal in auditory-visual speech



frequency for each marker, e.g., *Optotrak* (Northern Digital, Inc), *PhoeniX* (PhoeniX Technologies, Inc). They have the advantage that the sensors are tracked consistently and with high accuracy. On the downside, the number of sensors is limited, the sensors are relatively big and are wired to a device providing power to the LEDs and generating the pulse series with the characteristic frequency (though *PhoeniX* just introduced wireless LEDs). Accordingly researchers who employed these systems in their experiments had to make careful decisions about where to place the available 15-20 markers.

Passive marker systems (e.g., *Qualisys* (Qualisys Medical AB), *Elite* (BTS), *Vicon* (Oxford Metrics)) typically use highly reflective spherical markers. Small markers are available and there are no wires impeding natural face motion, but since single markers cannot be uniquely identified, marker confusion problems can only be resolved *posthoc*.

Both types of system have problems with reflections from sources other than the markers, though passive marker systems are in general more strongly affected. Also, for many tasks marker occlusions can only be avoided by adding more and more camera devices to the system.

An interesting alternative to the commercial systems mentioned above is the "home made" passive marker system developed at ICP (Grenoble, France) that uses beads as markers and tracks them with an appearance model [17].

2.2. Marker-free methods

Most marker-free methods operate on plain video recorded with a single or multiple cameras and the actual tracking happens exclusively on the software side (image processing and computer vision algorithms). However, face motion poses some notoriously difficult-to-handle problems:

- face motion is non-rigid motion; its degrees of freedom cannot be determined and, if approximated, are dependent on the chosen resolution;
- some areas of the face, like the cheeks, do not exhibit strong image gradients;

Accordingly, face motion cannot be conveniently modeled as *rigid body* motion and certain areas provide very little structure on which to base the motion tracking. The second problem can be alleviated by introducing structure artificially. Tracking using structured light [4] uses the projection of a light grid on the participant's face during the recording and the deformation of the regular pattern allows reconstruction of the 3D shape of the face and its changes over time. However, since no flesh points are tracked, the usefulness of the results is limited. In the commercial system *Mova Contour* (Mova) phosphorescent makeup applied to the face creates randomly scattered irregular shaped, but very dense patterns on the facial surface which only show up in fluorescent light. The moving face is then recorded with a camera array and patterns are tracked retrospectively with a very high resolution.

Another way to deal with the problems mentioned above is to circumvent them. Feature tracking methods measure the changes over time of predetermined facial features, e.g., lips, eyebrows, chin outline, nostrils. Note that the choice of features is usually governed by technical consideration regarding the tracking process, i.e., the requirement of a strong image gradient, and not by conclusive knowledge of the working mechanisms of auditory-visual speech processing. Nevertheless, for many tasks, adding, for instance, visual speech parameters to the acoustic feature vector in Automatic Speech Recognition (ASR), the impoverished description of speech face motion gained in this way might be sufficient.

Global video-based face motion tracking methods ideally make no high level assumptions about the areas of the face or the nature of the face motion to be tracked. Roughly speaking two main approaches can be identified:

1. Using dense optical flow (for a review see [2]) and applying additional constraints beyond the smoothness requirements at some stage of the tracking. For instance, in [5] the direct optical flow based tracking was optimized with an underlying muscle model using Kalman filtering.
2. Generating some kind of appearance model and using image registration techniques to find the

motion parameters that would have most likely produced the actual video image. For instance, in [13] a parameterised ellipsoidal mesh was used to model the facial surface as a whole and areas between neighbouring mesh nodes to model patches of it. A two-dimensional normalized cross-correlation served as technique to register the patches from the previous video frame with the current one. A coarse-to-fine strategy employing different mesh resolutions and a wavelet-based multi-resolution analysis on the image side were subsequently used to gradually increase the resolution of the tracking.

3. MEASUREMENTS AND WHAT THEN?

The typical measurement results consist of a time series of ten to several tens of measurement points with high internal correlations (*multi-collinearity*). Obviously a dimensionality reduction would be a preferable first step. And it is here where the problems start. A statistically motivated, all-purpose procedure like *Principal Component Analysis* (see [11] for a review) yields good results usually recovering a high amount of the variance in the data with only a few components [13, 8]. However, the components become increasingly difficult to interpret beyond the first two components which usually can be attributed to jaw movements and lip rounding.

In fact, it is not even likely that the higher components pick up physiologically meaningful behaviour: it would only be the case if the measured surface motion could be expressed as a linear combination of the underlying face muscle activity, but the passive "filtering" effect of the different layers of facial tissue and the intricately distributed insertion points of the facial muscles makes this relationship more complex and probably non-linear.

An alternative first step in analysis consists of determining components using information from the other "modalities", e.g., components that maximize the correlation (*Canonical Correlation Analysis*) or joint co-variance (*Partial Least Squares*) between either the articulatory or the acoustic signal and the face motion data. But there are drawbacks here, too: First, simultaneously recorded articulatory and face motion data are not readily available (but see [20]), and second, procedures like this can only capture linear relationships. The latter might not be the severe limitation it seems to be on first sight. Previous research [21] has shown that a high amount of variance can be "explained" with linear models and in many of the remaining cases problems can be linearized. The quotation marks enclosing "explained" are well deserved, since determining meaningful components in face motion via their relation to the other modalities does not explain anything, but simply pushes the explanation problem to a different level, one step higher. All findings remain phenomenological, although now inter-connected between modalities, if they cannot be evaluated relative to a theory of auditory-visual speech production. And such a theory is still missing.

4. ISSUES REGARDING A THEORY OF AUDITORY-VISUAL SPEECH PRODUCTION

4.1. Theoretical foundations (or lack thereof)

All major *speech perception* theories (e.g., [14, 6, 15] had to account for auditory-visual speech, since effects of visual speech, like the increase of intelligibility in noisy environments when watching the speaker's face [19], or the McGurk effect [16], are well-documented and were found to be pronounced and robust.

However, in our view there is only one comprehensive theory on *speech production*: Browman and Goldstein's *Articulatory Phonology* (AP) [3] in combination with Saltzman's *Task-Dynamic Model* (TD) [18]. Very briefly, Articulatory Phonology poses that the units of speech are dynamic articulatory gestures realized through movements of independently acting vocal tract organs. The latter are identified as the lips, the tongue tip, body and dorsum, the velum and the glottis. Gestures are defined as dynamical systems, the behavior of which can be fully described with a characteristic set of parameter values according to the Task-Dynamic Model. Every possible utterance of any language can then be modelled as parallel and serial combinations of these gestures. That is, gestures can be either executed simultaneously or strung together in time. Most of the time, however, both situations arise creating different amounts of gestural overlap. The gestures themselves are discrete and context-independent, their low-level (physical) realisations are the continuous, context-dependent and multi-dimensional patterns of vocal tract behavior found by speech production research. Note that the high-level (phonological) gestures do not need to specify details of the motor execution like context effects, since they are lawfully entailed by the dynamical systems implementing the gestures.

In principle, there is nothing preventing the treatment of visual speech production within the framework of AP/TD. Since the basic units of speech are assumed to be articulatory gestures, the sensory channel through which the information about them

is transmitted is of secondary importance. However, no explicit references to visual speech have so far been made.

4.2. Causal and non-causal functional relationships

If auditory-visual speech can be described within AP/TD, is there then any need to have a separate theory (or branch of it) dedicated to auditory-visual speech? The answer to this question depends on the answer to another question: Are the visible speech gestures exclusively a causal consequence of the vocal tract gestures or are there non-causal functional relationships, too?

- If all face motion providing phonetic information were the mechanical consequence of articulator movements, visual speech could be explained within AP/TD and no separate approach would be necessary (but see Section 4.3 for remaining problems).
- If there are non-causal functional relationships, e.g., between head or eye-brow movements and laryngeal activity, the picture becomes more complicated, since it can not be safely assumed that the related visual gestures behave in the same way as the vocal tract gestures.

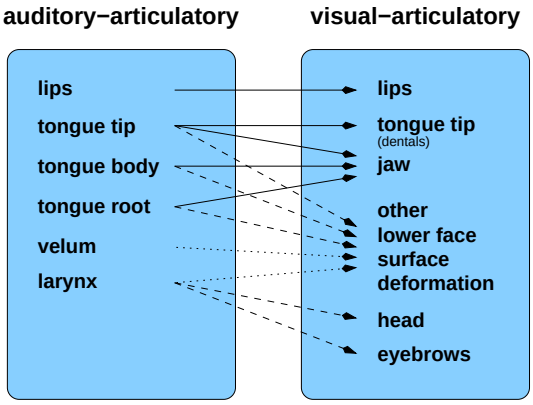
Unfortunately the research to answer the second question has not yet been conducted. In Section 3 we described ways to statistically assess the relationships between vocal tract behavior and face motion. In studies employing the described or similar methods face motion data and articulatory data were not recorded simultaneously with the exception of [12], and in all of them the articulatory data were limited to the mid-sagittal plane, thus unable to capture important characteristics of certain phoneme categories (laterals), missing potential sources of variance of others (e.g., the depth of the groove occurring at tongue back and root for many vowels) and ignoring lateral asymmetries in general. High cross-modal correlation recovering high amounts of variance were found nevertheless. Keeping the usual statistical textbook warnings about inferring causality from correlations in mind and the above shortcomings of the measurement methods it becomes clear that we can currently not decide which relations are causal and which are not with the exception of the trivial case of jaw and lip movements where the domains of visual and acoustic/auditory speech overlap. An ideal tool for investigating and answering this question would be a complete and detailed model of the vocal tract and the face, including bone structure, cartilages, muscles, tendons, etc. Clearly, such a model, be it a virtual computer or a real-world mechanical model, is still far away.

4.3. Articulatory Phonology and visual speech

Even if we assume that all relationships are causal or, equivalently, that existing non-causal functional relationships behave in all relevant aspects as if they were causal, interpreting speech face motion within the framework of AP/DT is still not straight forward. If visual speech were investigated without any knowledge about the acoustic/auditory channel, the two primary articulators would be the lips and the jaw. As reviewed in Section 3, they emerge as dominant essential components, and though lip shape and jaw position are not fully independent the fact that they are recovered by different principal components shows that they are functionally orthogonal, at least in ordinary speech. Lip gestures map directly from the articulatory-auditory to the visual domain: changes in the lip configuration lead to changes in the resonance frequencies of the vocal tract and they are also entirely visible. The jaw, however, it is not an articulatory organ in AP, but plays a subordinate role as part of a *coordinative structure* realizing gestures of the lips and the tongue. Accordingly some of the articulatory gestures must be fully or partially reflected in jaw movements.

Figure 2 visualizes potential mappings. Dashed lines indicate that there is some but not (yet) conclusive evidence for a relationship, dotted lines indicate hypothesized relations that, if they exist, would produce very weak effects.

Figure 2: Mapping of articulatory organs to their visual counterparts



As suggested in the figure, tongue tip, body, and root gestures might to a certain degree be recoverable from jaw position. But none of the three tongue "organs" confine the jaw to a certain position in an absolute sense, which should make it difficult to detect the active organ reliably. In addition, compensatory behaviour involving jaw and tongue (e.g., [7])

should result in visual ambiguity regarding the degree of constriction. Accordingly it can be predicted that those gestures that require the most precise jaw settings in natural unperturbed speech should be the easiest to recover and should be the default in ambiguous cases.

It also can be assumed that there are other minor surface deformations of the lower face that are consequences of tongue gestures. Velum gestures most likely have no visual complement at all, though some very subtle indications for a nasal consonant or more likely its absence (e.g., increased intra-oral air pressure moving the cheeks slightly outwards for stop consonants, but not for nasals) can not be completely ruled out. Correlations between the larynx and F0 have been found, but whether there are any visual complements to the voicing gesture needs to be investigated.

Similar to the articulatory-to-acoustic relationship, yet much more pronounced, the mapping between vocal tract behaviour and face motion appears to be a many-to-one mapping. As shown in [10] for the articulatory-to-acoustic mapping the inverse problem might be solvable in many cases. It would be worthwhile to investigate whether the differences in the many-to-one mappings are in fact to some degree complementary so that they could be used mutually to provide more constraints and help to disambiguate further.

The major advantage of studying auditory-visual speech within the AP/DT model and using it as a framework for the analysis is that within it articulatory gestures are well defined both on the physical level (surface phenomena) and on the cognitive speech processing level (phonology) in contrast to segmental approaches where, for instance, the complement of a phoneme on the physical level is hard to define.[1]

On the gestural level, predictions about visual gestures can be made and they can guide visual speech analysis. A comparison to the analysis of auditory speech makes the substantial improvement in the "analysis landscape" immediately evident. If acoustic speech analysis in the past would have been more or less limited to seeking statistical regularities in the acoustic signal and correlating it with articulatory measurements, then many of the advances in the field would not have been possible, no matter how carefully experiments on the perception side were designed. Describing and analysing acoustic speech within theoretical frameworks that explained the *underlying mechanisms* of its production process allowed the progression beyond the level of simply aggregating phenomenological findings.

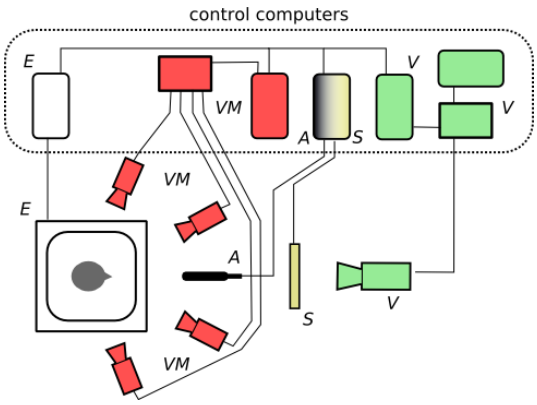
5. CONCLUSION AND OUTLOOK

In the last decade auditory-visual speech analysis has profited massively from advances in face motion tracking and measurement technology. Devices and systems have become widespread, more versatile, easier to use and cheaper. Statistical tools to handle multi-channel data returned by the face motion measurements are readily available. It seems, however, that - casually formulated - it is not clear what we are looking for. While on the application side the field has been able to progress without the need to understand the underlying mechanisms of auditory-visual speech in detail, e.g., in AVASR (Audio-Visual Automatic Speech Recognition), auditory-visual speech analysis critically requires a deeper understanding of speech face motion and how it is controlled in humans in order to advance in the same way.

AP together with TD offers in our view a very promising framework, though there are open questions to be answered first and many problems to be solved along the way. It would bring a very valuable dowry in the marriage, so to speak, a unified gestural approach spanning the entire phonetic-phonological domain that has the potential to explain the difficulties to determine the number of visemes and peculiarities of the McGurk effect. However, to sort out remaining problems speech face motion needs to be studied on the gestural level and hypotheses regarding the relationship between articulatory gestures and face motion need to be tested. Since currently no complete model of the human vocal tract and face exists, the quest for a theory that has been the topic of this paper ironically has to begin (and the paper to end) with a call for more data. The shortage of simultaneous 3D realtime recordings of vocal tract behavior and face motion, and the acoustic signal has already been mentioned. The situation looks even more bleak when taking into account that for any reasonable attempt to understand auditory-visual speech the data of many speakers and many languages/dialects are needed. For a comparison simply imagine phonetics and phonology would have until now comprised only of, say, English, German, Swedish, French, and Japanese.

At MARCS Auditory Laboratories we recently created a new facility to study auditory-visual speech production (and perception) offering the possibility to record data in the three modalities simultaneously. The setup consist of a 3D-EMA system (AG500, *Carstens Medizinelektronik*), a high-resolution face motion tracking system (Vicon system with four MX40 cameras, *Oxford Metrics*), a high-speed high-resolution camera (Phantom v10

Figure 3: Setup of the system to study auditory-visual speech production at MARCS Auditory Laboratories - **E**: Electromagnetic Articulograph; **VM**: Vicon MX40 system, **A**: Audio; **S**: Stimulus presentation; **V**: High-speed Video



with ImageCube high speed mass storage system, *Vision Research*) and three powerful workstations and a data server. Figure 3 shows the setup. To minimise reflection the acrylic glass cube of the EMA system was painted black. The combined equipment can record articulatory and face motion data at 200 Hz with the 4 Mega-Pixel resolution (2352 x 1728 pixels up to 160 fps) of the motion tracking Vicon cameras only slightly reduced by windowing and the color video Phantom camera still on full resolution (2,400 x 1,800 pixel up to 480 fps). Thus we are now in the position to examine auditory-visual speech as the cross-modal process it genuinely is and provide data to test a gestural account of speech face motion.

The author would like to thank Michael Tyler, Chris Davis, and two anonymous reviewers for many insightful comments.

6. REFERENCES

[1] Arbib, M. A. 2002. The mirror system, imitation, and the evolution of language. In: Dautenhahn, K., Nehaniv, C. L., (eds), *Imitation in animals and artifacts*. Cambridge, Massachusetts: MIT Press 229–280.

[2] Barron, J. L., Fleet, D. J., Beauchemin, S. S. 1994. Performance of optical flow techniques. *International Journal of Computer Vision* 12:1, 43–77.

[3] Browman, C. P., Goldstein, L. 1992. Articulatory phonology: An overview. *Phonetica* 49, 155–180.

[4] Carter, J., Shadle, C., Davies, C. May 1996. On the use of structured light in speech research. *ETRW - 4th Speech Prod. Seminar* Autrans. 229–232.

[5] Essa, I., Pentland, A. July 1997. Coding, analysis, interpretation and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19 (7), 757–763.

[6] Fowler, C. A. 1986. An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics* 14, 3–28.

[7] Geumann, A., Kroos, C., Tillmann, H. G. 1999. Are there compensatory effects in natural speech? *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS)* San Francisco, USA.

[8] Gutierrez-Osuna, R., Kakumanu, P. K., Esposito, A., Garcia, O. N., Bojórquez, A., Castillo, J. L., Rudomín, I. 2005. Speech-driven facial animation with realistic dynamics. *IEEE Transactions on Multimedia* 7(1), 33–42.

[9] Hockett, C. 1963. The problem of universals in language. In: Greenberg, J., (ed), *Universals of Language*. Cambridge, MA: MIT Press.

[10] Hodgen, J., Rubin, P., McDermott, E., Katagiri, S., Goldstein, L. in press. Inverting mappings from smooth paths through R^n to paths through R^m: A technique applied to recovering articulation from acoustics. *Speech Communication*.

[11] Jackson, J. E. 1991. *A user's guide to principal components*. New York: John Wiley & Sons.

[12] Jiang, J., Alwan, A., Bernstein, L. E., Keating, P., Auer, E. 2000. On the correlation between facial movements, tongue movements and speech acoustics. *International Conference on Spoken Language Processing* volume 1 Bejing, China. 42–45.

[13] Kroos, C., Kuratate, T., Vatikiotis-Bateson, E. 2002. Video-based face motion measurement. *Journal of Phonetics (special issue)* 30, 569–590.

[14] Liberman, A. M., Mattingly, I. G. 1985. The motor theory of speech perception revised. *Cognition* 21(1), 1–36.

[15] Massaro, D. W. 1998. *Perceiving Talking Faces. From Speech Perception to a Behavioral Principle*. MIT Press.

[16] McGurk, H., MacDonald, J. 1976. Hearing lips and seeing voices. *Nature* 264, 746–748.

[17] Revèret, L., Bailly, G., Badin, P. 2000. MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. *International Conference on Speech and Language Processing* Beijing, China.

[18] Saltzman, E. L., Munhall, K. G. 1989. A dynamical approach to gestural patterning in speech production. *Ecological Psychology* 1(4), 333–382.

[19] Sumby, W., Pollack, I. 1954. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America* 26, 212–215.

[20] Wrench, A. 2000. A multi-channel/multi-speaker articulatory database for continuous speech. recognition research. *Phonus* (5), 1–13.

[21] Yehia, H. C., Rubin, P. E., Vatikiotis-Bateson, E. 1998. Quantitative association of vocal-tract and facial behavior. *Speech Communication* 26, 23–44.

[1] Strictly speaking, a *viseme* (as the term is currently used) is not defined at all, since there is no visual-only language of which it could be the smallest units conveying a distinction of meaning: only sign languages could have visemes.