COORDINATION, CONCURRENCY, AND SYNCHRONY IN COMMUNICATION

1. INTRODUCTION

In this talk I consider some of the challenges that face our attempts to observe and interpret human communicative behavior. These include comprehending the complexity of the signals produced in association with the events of interest and determining how they should be processed and measured. For example, signals may occur simultaneously within and across multiple channels and modalities, at multiple physical locations, and with the potential for signal correspondence at multiple levels of spatial and temporal coordination -- that is, patterns within patterns. Understanding how events are detected, processed and interpreted by a perceiver is equally complex and involves event structures specific both to perception and production. Ideally, the analyses of production and perception should inform one another, though this seldom happens in practice.

Determining what events to measure is limited by technology and the often premature use of theory to characterize the relevant events. Theory-driven research achieves results, but relies heavily on previously defined structures, and rarely accommodates event structures that are emergent in context-specific behavior. In this talk, I argue that emergent events are fundamental to communication, which is inherently context-dependent and ephemeral. I further argue that emergence can be anticipated and indeed must be in order to understand how our most highly-skilled behaviors -- spoken communication and music -- are organized in space and time.

1.1. Observing events in communicative behavior

Broadly speaking I want to convey to you a conceptual framework and rudimentary methodology for observing human communicative behavior. Some questions to bear in mind are:

• How to identify and validate events within spatial and temporal continua that span multiple channels and modalities? The different regions of the body that can be measured such as the head, face and torso are examples of channels. Modalities might be sensory, differentiating what can be seen or heard, or the neuromotor systems that generate what can be seen or heard.

• How to relate the production and perception of events in a common system of description? I treat production and perception as fundamentally connected, but necessarily different and deserving of different domains of description. This is in contrast to the popular trend of the past 15 years or so to try to make one domain mirror the other, so that the descriptive system developed for one domain, production

* Texto apresentado em conferência ministrada por Eric Vatilkotis no dia 30 de setembro de 2009, como participação no Programa Cátedras IEAT/FUNDEP.
for example, accounts for the other. That is, there should be a unified description of production and perception, but not one in which the two components are a simple transform of each other.

• How to optimize the effects of scientific observation (observer bias). Observer bias is inescapable and has been made more problematic by empirical traditions that embrace the myth that the world can be examined objectively. Preserving this myth causes us to attempt to eliminate biasing influences by over-simplifying both our observational methods and the phenomena being observed. Experiment designs that contrast just a few conditions tested on unrealistically simple forms must then be extensively interpreted to fit reality. This can introduces biases worse than ones originally being avoided and often contravene the logic of the discipline. I think we do better to accept the limitations on empirical objectivity, incorporating bias and putting it to work for us, rather than wasting energy trying to eliminate it.

• How to equivocate or co-define events and behavior? This is a huge problem, at least for me, in part because we use these terms casually to mean many different things. Our recognition of behavior is generally describable only in terms of events whose boundaries in space and time can be recognized. A conundrum, however, is that we are linguistically and cognitively conditioned to depend on labels for identifying relevant aspects of behavior. Not only does the labeling process distort the reality of the events they denote, but it begs the question “how do we discover new events for which we do not already have labels?”

1.2. Talk structure

This talk has three components. In the first, I give a conceptual overview in which I use common events in our daily lives to define and contextualize coordination, concurrency and synchrony. By defining and exemplifying these concepts in common terms, rather than within a specific scientific contexts I delve briefly into how these concepts interact with our notions of sequence and causality, and I provide further rationale for differentially describing production and perception. The second part of the talk focuses on a sizable body of research examining auditory-visual speech processing that was carried out at ATR International in Japan with UFMG’s own Hani Yehia and other collaborators such as Kevin Munhall and Takaaki Kuratate. While this work has been quite informative about the perception and production of multimodal speech behavior and has helped create a broader interest in investigating intermodal dependencies, static techniques were used to compute correspondences across the various measurement domains. That is, the temporal organization of the observed behaviors was ignored. This sets up the third and final section of the talk whose focus is on the use of dynamic analyses to examine cross-modal correspondences in language and music and how they pattern in time. It is my firm belief that it is this patterning of behavior that forms the basis for coordination and is a prerequisite for communication to occur.
2. CONCEPTS DEFINED AND CONTEXTUALIZED IN DAILY LIFE

2.1. Coordination, concurrency, and synchrony

Three fundamental concepts are used to describe time-varying events. These are my own working definitions and are not necessarily agreed upon by others.

**Coordination:** All events are coordinated internally and/or externally with the environment, and within the environment perhaps coordination with other specific events or entities, such as other people.

**Concurrency:** Events that co-occur are concurrent, but are not necessarily coordinated. Alternatively, events may be coordinated without being concurrent (e.g., sequences).

**Synchrony:** This overused term usually implies a temporally rigid and immediate coordination between events that are concurrent and coordinated at a fixed phase (preferably at zero degrees offset). Synchrony has more reality as an ideal form similar to one of Euclid’s geometric forms which are perfect in concept, but imperfect in their realization.

2.2. Concepts defined and contextualized in daily life

In order to avoid over-complicating the definitions of concurrency, synchrony and coordination, I turn to examples of events in everyday life whose timing, either absolute or relative to other events, is important and may be controlled or scheduled in some way. It is useful, but not entirely essential to the discussion, if these sample events repeat themselves with some regularity.

Think about things you do daily between waking and arriving at work, such as getting dressed, eating breakfast, traveling to the University. I use this example because reaching UFMG from downtown Belo Horizonte is an adventure requiring careful coordination and planning.

How do you think about these events? Are they isolated; are they independent; or are they just unconnected items juxtaposed in a list by grammatical accident? Do you order events in a sequence where one event has to happen after another; or do you think of them in some other way? Chances are you think of these events as sequenced according to regularly recurring causal criteria. That events recur is essential to pattern recognition, and I think it is fair to say that we readily ascribe causality to observed patterns, but what form does such causality take in sequences? Is it external to the sequence or integral the serial progression of elements in the sequence? So, let us think about what causes or constrains a series of events. Also, we should think about the possible difference between the linguistic act of naming events – a series of event labels – and the actual events themselves. Such a difference is problematic not only in our daily lives, but for our science as well.
2.3. Sequences and serial order

Sequences are series of events whose particular order is either the accidental occurrence of unrelated, independent events or is determined by a set of interrelated causes specific to the series. For events to be independent, their causes must be independent. A sequence may be composed of such events, as shown on the left side of Figure 1. In general, we have difficulty accepting that events that occur in close temporal juxtaposition can be independent, and we exert considerable effort developing causal accounts for sequence order. For events to be serially ordered in the sense that I intend, the sequence of events or their causes must condition other causes and/or events in the sequence, as shown on the right side of Figure 1. In one case events in a sequence are coordinated, in the other there is no coordination.

2.4. Daily activities are not serially ordered

Now, let’s reconsider whether or not our sequence of early-morning activities is serially ordered. The organization and dependencies between getting dressed (D), eating (E), and traveling to work (T) are shown in Figure 2. The are a number of arguments against these activities being serially ordered:

- The relative order of (some) activities can be changed without consequence; that is, their order is arbitrary. For example, I can eat before getting dressed, or travel to work before eating (but not before getting dressed).

- Some activities can be concurrent, overlapping in time. I can eat and get dressed simultaneously. A little more recklessly, I can eat and travel simultaneously, but am unlikely to attempt to dress while traveling to work.

- Some activities are not essential to the sequence and can be omitted from the series. Travel to work is mandatory; being dressed to travel is also mandatory; therefore these two events are sequenced; while it may be unwise to do so over the long term, eating can be omitted.
2.5. Practical limitations of cognitive/linguistic scope

Despite evidence that events in our daily lives are not serially ordered – at least, not in any simple way – we persist in perceiving them that way. A serious examination of why this occurs is beyond the scope of this talk, but part of the reason has to do with our tendency to infer causality generically as a placeholder for more precise knowledge of how events are interrelated. For one thing, there are practical cognitive and linguistics limitations on our ability to conceive and label the complex of factors, internal and external to the organism producing or perceiving these events, that define a sequence of events. Consider:

- The inability to dress while driving is an internal physical constraint. To attempt to put on one’s shoes while trying to steer a vehicle, control its speed, and monitor potential hazards simply cannot be done.

- The need to dress before traveling is an external social constraint against public or, in this case, semi-public nakedness.

- The decision to eat or not has consequences outside the span of time being considered. For example, no breakfast in the morning may necessitate an early lunch, which does not fall within the sequence of morning events. The causal connection between these two events requires a longer, explicit time line.

With effort, we can recognize and enumerate some of the causes and constraints on our behavior, but typically we are able to do so only within highly localized and egocentric bounds, that do not accurately reflect the sphere of conditioning factors on our behavior. For example, we are notoriously bad at conceptualizing and describing how the concurrent behavior of others is coordinated with and conditions our own behavior. Yet, in the context of planning my departure for UFMG in the morning, the timing of when to leave the house is conditioned by an assessment of how best to avoid the rest of the commuting population in order to minimize the time spent in traffic congestion. That is, my account of what conditions my behavior and the conditions themselves are so different as to be almost unrelated. Our understanding is, even of simple, recurring daily activities, is hampered cognitively and linguistically in the way that temporal phenomena are invariably stripped of their dynamical
and relational properties and enumerated sequentially. Language specifically, and cognition in general, occur in time and are by necessity so ephemeral that they do not readily support descriptions of concurrent structures. Serial order and, therefore, causality are immanent in the objectification of events that occurs when they are listed in sequence, regardless of their actual interrelatedness.

2.6. Perceiving and producing patterned behavior

Despite linguistic objectification, behavioral events are inherently dynamic and are invariably patterned. Patterns are defined by recurring structures that are spatially and temporally similar. This is essential for pattern detection and comprehension.

Although similar, the perception and production of patterned behavior are not isomorphic. This is not a radical idea and many people would readily support the concept as fundamental. However, it is rare in the research domain for people to put this idea into operation. Typically, researchers concentrate exclusively on either production or perception; and, when they do talk about them together – usually in an effort to provide a unified account, the tendency is to apply the same descriptive frame to both. I, too, am interested in a unified account of production and perception, because I believe neither production nor perception can be properly investigated or understood in isolation. However, we need to be careful not to enforce a parsimonious account of two domains that are subtly, yet I believe fundamentally, different.

In particular, the role of production is to generate behavior that can effect some goal, be repeated when successful, and be modified when unsuccessful. If the production involves interaction of the producers with the physical environment, such as picking up a rock, then the producer must also be able to perceive the event in such a way that refinements of technique or variations can be implemented. In the case of communicative behavior involving other perceivers, then the production and perception of the producer must be coupled with those of others.

Given that production and perception involve different physiological structures, it is likely that event structures within each will differ. For example, speech production involves the coordinated action of neuromotor sub-systems controlling the respiratory and upper airways structures, at least two separate incursions on the postural control system (by way of respiration and speech correlated head motion), and the intricate coupling of the laryngeal and supralaryngeal vocal tract structures responsible for the production and shaping of period sounds, respectively. During production, the auditory system of the producer is also active monitoring and possibly contributing online to the coordination and control of production. By contrast, even if the same sub-systems are somewhat active during perception – exemplified by traces of neuromotor activity in the laryngeal and supralaryngeal structures of the perceiver – the behavioral coordinates cannot be the same as for production.

If the basic coordinate systems for specifying events in space and time differ for
production and perception, then there may be flexibility in the linkage between the two domains. Such flexibility has important implications for categorization and other cognitive processes that must somehow normalize event variability within and across domains. For example, consider how perceivers can identify the same speech sound in the unique acoustic signals produced by many speakers. Similarly, on the production side, a speaker never produces the same acoustic signal twice, but can come close enough to believe that the same speech pattern is being repeated. Inaccuracies in both perception and production are essential for normalization to occur. More importantly, perception and production use these discrepancies, or error signals, to adjust and fine-tune event normalization dynamically without depending on exact reproduction of fixed signals (e.g., Sandell & Chronopoulos, 1997).

These concepts are not fully developed, and will not be directly addressed in the research presentation that follows. However, they are consistent with the approach to observing, measuring, and attempting to understand the organization of communicative behavior that my colleagues and I have developed over the past 15 years.

3. STATIC ANALYSES OF CROSS-MODAL CORRESPONDENCE IN SPEECH

The second part of this talk provides an overview of a sizable body of work that my colleagues and I carried out over a 10 year period beginning in the mid-1990’s. It is especially relevant in the context of a lecture to the UFMG community, since Prof. Hani Yehia, a prominent member of the Engineering Faculty here was a crucial contributor to this endeavor, initially while a post-doctoral fellow in my laboratory group at ATR International in Japan and later after his return to Brazil. Furthermore, as presented in the third section of the talk, our examination of multi-modal correspondences in spoken communication have continued to evolve from primarily static to dynamic analyses thanks to the efforts of one Prof. Yehia’s first PhD students here at UFMG, Dr. Adriano Vilela Barbosa. Thus, the success of this endeavor owes a great deal to UFMG.

3.1. Auditory-visual speech processing

For about 15 years now, interest in the production and perception of multimodal speech has developed rapidly. Starting with a NATO Advanced Study Institute held in Bonas, France, in 1995 (Stork & Hennecke, 1996), a multi-disciplinary group of psychologists, engineers, linguists, clinicians, and artists has converged on a self-styled area of study now known as Auditory-Visual Speech Processing (AVSP), which addresses the production, perception, artificial synthesis and recognition of multimodal speech, as well as associated brain functions.

3.1.1. Basic claims about auditory-visual speech

In the earliest days of this endeavor, there was one basic claim being made about speech perception.
Claim 1 – speech perception is multisensory.

To support this claim there were two fundamental observations:

• Being able to see a speaker’s face improves intelligibility. This was first shown experimentally by Sumby and Pollack (1954) in an experiment using a live presentation of a talker whose voice was partially masked by noise. When he faced the listener, word recognition improved compared to when his face could not be seen.

• The integration of auditory and visual signals is not voluntary. Harry McGurk discovered through an accident of stimulus construction that when a speaker’s audible production of the syllable /ba/ is accompanied by visible /ga/, most listeners hear /da/ or /ða/ (McGurk & MacDonald, 1976).

These findings contributed to a growing interest in how the auditory and visual modalities are integrated neuro-cognitively (Massaro, 1987; Massaro, Venezky, & Taylor, 1979; Summerfield, 1979). For example, is modal integration early or late in sensory processing? Is one modality functionally tied to the other (Robert-Ribes, Piquemal, Schwartz, & Escudier, 1996)? And so on. When combined with psychometric results, neurological investigation of both the timing (through electroencephalography – EEG) and location (using functional Magnetic Resonance Imaging – fMRI) of auditory and visual processing attempted to provide clear answers to these questions (e.g., Callan, Jones, Munhall, Callan, et al., 2004; e.g., Callan, et al., 2002; e.g., Callan, Jones, Munhall, Kroos, et al., 2004).

Beginning with the work of Quentin Summerfield and colleagues (Summerfield, 1987; Summerfield & McGrath, 1984), it became increasingly clear that the relative contribution of auditory and visual information to speech perception varied with task conditions, sometimes providing redundant and other times complementary support for perception (Grant & Seitz, 2000). This led me and my colleagues to attempt to examine the degree of correspondence between audible and visible components of the the production and to determine their role in perception. The results of that effort are reviewed in the second section of this talk and can be summarized for speech production as:

Claim-2: Speech production is largely amodal (redundantly multimodal)

• Audible speech results from sound-source excitation of a time-varying resonator – the vocal tract (VT).

• Since the VT is bounded externally by the face, motions of the lips and jaw, and indirectly the tongue necessarily deform the face while shaping the acoustics.

An unforeseen consequence of including physiological measures of both muscle activity and resulting motion behavior, was
Claim-3: Audible and visible components have one neuromotor source

For example, the face is deformed actively for expressive gestures and passively as part of acoustic speech production.

3.1.2. General methodology

Early on in our efforts to examine the behavioral measures of the muscles, motions, and acoustics associated with speech production, we recognized that in order to determine any linguistic relevance for production results, we would have to validate them perceptually. The easiest way we could see to do this was to use our measured data to generate talking heads stimuli for perceptual evaluation. By definition, the synthesized stimuli would consist entirely of the data we had analyzed – unlike a video recording, which contains all manner of information about which we have no precise knowledge. Because the synthesis process would itself be prone to introducing error by distorting the measured data, it was essential to validate each step of the synthesis process kinematically. The resulting talking head stimuli could then be used to assess the connection between speech perception and production, and ultimately be extended to brain function and other performance domains such as music. So far as I know, this was the first time such care was taken to insure that the characteristics of the stimulus source were known and incorporated in perceptual evaluation of multimodal speech.

Research procedure

1. Measure the multilinear correspondence between audible and visible components of speech production.
2. Develop and validate (kinematically and perceptually) talking heads animated by measured production data.
3. Apply methodology to assess the linkages between the production and perception of spoken communication.

This has been an extensive undertaking involving a large set of collaborators from around the world who either worked with me at ATR in Japan or subsequently in Canada. Collaborators include Adriano Barbosa (Brazil, Canada, Japan), the late Christian Benoît (France), Daniel Callan (Japan), Hugo de Paula (Brazil, Canada, Japan), Takaaki Kuratate (Australia, Germany, Japan), Kevin Munhall (Canada, Japan), Philip Rubin (USA), Mark Tiede (Japan, USA), Sumio Yano (Japan), and Hani Yehia (Brazil, Japan)

3.1.3. Cross-modal correspondence in speech production

The first step in the research procedure outlined above is to assess the correspondences
between audible and visible components of speech production. The analysis of inter- and cross-modal correspondences consisted of physical measures of the acoustics, motions, and muscle activity recorded during multimodal speech production. It was not possible to record all four components shown in Figure 3 simultaneously. Specifically, muscle EMG, and acoustics were recorded along with 2D vocal tract measures and compared to a separate recording of the EMG, acoustics, and 3D face data for the same speech materials. This was purely a limitation of the electromagnetic (EM) system used to measure vocal tract data (Perkell, et al., 1992). There is now an EM system produced by Northern Digital, Inc. that integrates with 3D position measurement and does not obstruct video recording of the face.

In our analysis of these production measures, Hani Yehia and I took the approach that physically tractable multilinear analyses would be the best place to begin examining the structure of audiovisual speech. At the time, this was a novel approach as most researchers were attempting complex nonlinear analyses in which the boundaries between parameterization and empirical measures were often necessarily unclear. This is exemplified in the next section.

3.2. Computational model of speech motor control

As a way of demonstrating how important it is sometimes to back off and simplify an approach to a problem in order to make progress, I provide here a brief overview of the earlier intense effort to devise a computational model of speech motor control that immediately preceded my work with Yehia. This work, shown schematically in Figure 4, used the same production data and was carried out with Mitsuo Kawato and Makoto Hirayama at ATR (Hirayama, Vatikiotis-Bateson, Honda, Koike, & Kawato, 1993; Hirayama, Vatikiotis-Bateson, Kawato, & Jordan, 1992; Wada, Koike, Vatikiotis-Bateson,
3.2.1. Overview

In the early 1990’s, we believed physical measures for speech production of the sort depicted in Figure 3 could parameterize a workable model of speech motor control that would satisfy Marr’s (1982) constraints for computational modeling of biological systems. In particular, it was believed that low dimensional statistical models made of artificial neural networks could capture the conversion of discrete intentions to continuous neuromotor activation by estimating articulator motion from muscle activations (forward dynamics) followed by estimation of the speech acoustics from the vocal tract shape (forward kinematics). Key to the conceptualization was the insight of Kawato and colleagues (Kawato, Maeda, Uno, & Suzuki, 1990; Uno, Kawato, & Suzuki, 1989) that the nervous system’s task of determining trajectory paths to achieve specific targets could be made tractable; in particular, by severely limiting possible solutions using an objective constraint on smoothness and via point targets corresponding to the cognitive sequence of discrete elements, in this case phonemes, specifying the intended utterance (Kawato, 1992).

Back propagation networks were tasked with learning the force-acceleration relation needed to generate trajectories that would pass through the via points corresponding to phoneme sequences for utterances such as *When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow*. Performance parameters were estimated for speaking style – casual, precise, fast, etc. – and smoothness was constrained by a fifth order spline minimizing jerk – the fifth derivative of position. Similar networks were used to learn the relation between 2D position of the lips, jaw, and tongue and correlation parameters (PARCOR) of the acoustics (Kitawaki & Itakura, 1978).
Figure 4. Shown at left is the schematic overview of the computational model of speech motor control developed in collaboration with Hirayama and Kawato in the early 1990’s. Elements of the system depicted by boxes represent measured or known components (muscle EMG was equivocated with motor commands). Elements depicted by ovals were “learned” by neural network training and then used to generate muscle, articulator motion/ vocal tract shape, and acoustic output signals. At right is shown a representative Magnetic Resonance Image (MRI) in the mid-sagittal plane with measurement points (somewhat stylized in location) for locations on the tongue, lips, and jaw.

3.2.2. Computational model of speech motor control – Limitations

While our model for generating appropriate muscle forces for shaping the VT through time from an input string of phoneme via points and a smoothness constraint was quite elegant (Hirayama, Vatikiotis-Bateson, & Kawato, 1994), the results were frustrating. The following table summarizes in bullet form some of the most serious problems encountered with network training.

<table>
<thead>
<tr>
<th>ANN training</th>
<th>Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>• phoneme via points</td>
<td>• utterance-specific setting</td>
</tr>
<tr>
<td>• muscle EMG signals</td>
<td>• sparse sampling (9 of 200 muscles); EMG does not indicate muscle force</td>
</tr>
<tr>
<td>• vocal tract articulator acceleration,</td>
<td>• acceleration = noisy second temporal derivative of position =&gt; no zeros</td>
</tr>
<tr>
<td>• vocal tract acoustics</td>
<td>• PARCOR parameters not dynamic</td>
</tr>
</tbody>
</table>

Overall, our results showed that contrary to expectation, the neural network was learning largely linear relations between muscle EMG and vocal tract articulation. This led to underestimation of articulator position change. Furthermore, during recurrent estimation for trajectory formation, the estimation error had cumulative effects that prevented the model from converging on a singular solution for all but the simplest of utterances. In the end, I concluded that our measured data were too sparse and noisy to provide an adequate test of such a model architecture and that any number of models could work just as badly (Vatikiotis-Bateson, Munhall, Hirayama, Lee, & Terzopoulos, 1996). However, these lessons were important and set the stage for the highly successful analyses of the same data using multilinear analysis.

3.3. Back to multilinear correlation analysis

3.3.1. Computing the correspondence within and across signal domains

The multilinear approach to assessing physiological speech data began with the forward and inverse relations between muscle EMG and articulator motion and between articulator motion and the spectral (PARCOR) parameters of the speech acoustics (Vatikiotis-Bateson & Yehia, 1996; H. Yehia, Tiede, Vatikiotis-Bateson, & Itakura, 1996). Minimum error estimation of articulator motion from muscle EMG was possible to a certain extent on an utterance-by-utterance basis, but no reasonable inversion from motion to muscles was attained. Other colleagues were attempting to estimate
muscle activation patterns from kinematic measures (e.g., Pitermann & Munhall, 2001), but with limited success and only for the simplest utterances.

This exercise was valuable, because it clarified the need to match the dimensionality of the data sets we were attempting to compare across measurement domains. That is, EMG signals for eight muscles could not be used to estimate 51 dimensions of 3D face motion (51 = 3 coordinate axes of motion × 17 marker locations on the face and lips), because the few to many calculation would be mathematically ill-posed. Calculating the forward kinematic correspondence between articulator motion and 10 spectral acoustic parameters constituted a many to few computation, which was not ill-posed, but not ideal either. We turned therefore to the problem of reducing the dimensionality of data within each measurement domain. Using principal component analysis (PCA), which decomposes variance orthogonally in a stepwise fashion, we quickly discovered that the dimensionality was roughly the same for each of the vocal tract, face, and acoustic measurement domains. In representing the spectral acoustics, we replaced PARCOR parameters with line spectrum pairs (LSP), which display a more robust correspondence to the time-varying vocal tract shape (Sugamura & Itakura, 1986).

The results for face, vocal tract (VT), and acoustics (LSP) are shown for two speakers of English and Japanese in the upper panels of Figure 5. For each signal type, regardless of speaker or language, fewer than 10 components are needed to account for roughly 99.5% of the variance.

3.3.2. Summary of cross-domain analysis and synthesis

To wrap up the overview of cross-domain correspondence analysis and subsequent synthesis of face motion and speech acoustics, our studies in the late 1990’s were based on correspondences between time-varying signals computed synchronically for each time sample. Although natural speech utterances were used to greatest extent possible, even running speech, these analyses were static consisting of globally averaged correspondence. Linear correlations were sufficient to do reasonably good forward-inverse estimations between VT and face and unilateral estimation of face from acoustics. We were gratified to find correspondences between acoustic LSP and VT motion that made sense physiologically; in particular, VT correspondence was highest in the 1500-2500 Hz range which indicates good coherence with the resonances (F2 and, to an extent, F3) in front cavity of the vocal tract.

Neural networks, trained on individual principal components of motion, were used to estimate

- face motion from acoustics accounting for more than 95% of the variance;
- acoustic LSP from face motion, which resulted in intelligible speech synthesis.

Compared to the multilinear estimations, the dimensionality of the nonlinear computation remained small while improving the estimation. However, the nonlinear
computations were severely limited to small sets of training utterances and sometimes even to single sentential utterances. We believe this failure to generalize beyond individual utterances indexes the effective limits of static correlation methods in which the scope of any global average does not extend beyond several linguistic phrases (Vatikiotis-Bateson & Yehia, 2002).

Figure 5. Plots showing small number of orthogonal components needed to reconstruct domain specific data (top) and to estimate the components of one domain from those of another (bottom) for a speaker of English (left) and Japanese (right).

Demonstration that data measured in different signal domains are similar in their variance structure, while interesting in its own right, meant that we could eliminate the possibility of ill-posed one-to-many and awkward many-to-one computations across measurement domain.

Using square matrices containing the small numbers of eigenvalues provided by PCA, singular value decomposition (SVD) was used to compute the cross-domain correspondence between signal types. As shown for Face and VT components in the lower panels of Figure 5, along with their within-domain counterparts (from the upper panels), even smaller numbers of orthogonal eigenvectors (<6) are needed to estimate one domain from the other. It is clear from the figure that the extent of cross-domain estimation is less than within-domain estimation. Furthermore, it is direction
dependent; estimation of Face components from VT components \( (r \approx 0.9) \) is better than that of VT from Face \( (r \approx 0.8) \). This asymmetry is similar to that found previously for the correspondence between muscle EMG and VT data, and perhaps for the same reason which we cautiously name causality. I say “cautiously” because we assume that muscle activation is prior to VT motion and that VT motion is prior to the correlated components of Face motion, because the latter is simply the external result of shaping the vocal tract.

There is a second possible explanation, which has greater generality because it also accounts for the relatively poor estimation of speech acoustics from either VT or Face data using these multilinear techniques (H. C. Yehia, Kuratate, & Vatikiotis-Bateson, 1999). Being able to reduce the dimensionality for each measurement domain to roughly the same number of components, does not equivocate differences in the relative contribution of each component to the overall variance. For example, as careful inspection of Figure 5 reveals, fewer components are needed to account for 95% of the variance than either the VT or the acoustic LSPs. Over the years we have seen time and again for a range of speakers of many languages that most of the face motion is accounted for by three components corresponding to jaw height, lip shape, and a much weaker parameter corresponding to bilateral asymmetry in face motion. Another example is that relatively small contributions to variance can have large effects on the robustness of cross-domain correspondence. This is what we found when attempting to synthesis intelligible acoustics from a nonlinear combination of VT and Face components; it was only after including components accounting for less than 0.5% of the variance that intelligible speech could be synthesized from motion (H. C. Yehia, Kuratate, & Vatikiotis-Bateson, 2002). We have never attempted to calculate the information content for these measurement domains, but we suspect that such calculation would reveal that both muscle EMG and speech acoustics contain more signal information than our motion measures of either the vocal tract or the face.

3.3.2. Summary of cross-domain analysis and synthesis

To wrap up the overview of cross-domain correspondence analysis and subsequent synthesis of face motion and speech acoustics, our studies in the late 1990’s were based on correspondences between time-varying signals computed synchronically for each time sample. Although natural speech utterances were used to greatest extent possible, even running speech, these analyses were static consisting of globally averaged correspondence. Linear correlations were sufficient to do reasonably good forward-inverse estimations between VT and face and unilateral estimation of face from acoustics. We were gratified to find correspondences between acoustic LSP and VT motion that made sense physiologically; in particular, VT correspondence was highest in the 1500-2500 Hz range which indicates good coherence with the resonances (F2 and, to an extent, F3) in front cavity of the vocal tract. Neural networks, trained on individual principal components of motion, were used to estimate
• face motion from acoustics accounting for more than 95% of the variance;
• acoustic LSP from face motion, which resulted in intelligible speech synthesis.

Compared to the multilinear estimations, the dimensionality of the nonlinear computation remained small while improving the estimation. However, the nonlinear computations were severely limited to small sets of training utterances and sometimes even to single sentential utterances. We believe this failure to generalize beyond individual utterances indexes the effective limits of static correlation methods in which the scope of any global average does not extend beyond several linguistic phrases (Vatikiotis-Bateson & Yehia, 2002).

On the positive side, these results do support our conclusion that the strong correspondences between motions of the vocal tract and face, and the speech acoustics and the face are the inevitable consequence of an integrated neurophysiology and biomechanics in which time-varying configuration of the vocal tract determine both the output acoustics and deformations of the face.

Figure 6 provides a snapshot reference to a movie created by Dr. Kiyoshi Honda of ATR International. The purpose of this movie in the original talk presentation was to help the audience appreciate the complexity of the vocal tract production system under discussion. That our static analyses capture enough of the structural and functional coordination of the vocal tract to synthesize intelligible speech acoustics indicates, we believe, that speech is relatively sparsely specified by the anatomy and physiology in both space and time.

**Figure 6.** Still frame from MRI movie of the vocal tract showing the spatial and temporal complexity of speech production – courtesy of Kiyoshi Honda, MD (ATR International, Japan).

### 3.4. Amodal vs. non-amodal correspondence

A more complex coordination in which audible and visible events emerged when we discovered that rigid body (6D) motion of the head is highly correlated with the frequency (F0) of vocal fold vibration (H. C.Yehia, et al., 2002). In this case, the information conveyed by the two modalities is not amodal.

Control of the vocal folds and of head motion do not have the same neuromotor source and, indeed, do not appear to have the same perceptual consequences (see below – Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004).

Adding the correspondence between head motion and F0 allowed us to more completely estimate speech acoustics from purely visible behavior with spectral acoustics and some aspects of RMS amplitude coming from face motion and F0 and
some more of the RMS amplitude coming from head motion (H. C. Yehia, et al., 2002).

3.5. Connecting production with perception: Talking Head animation

Analytic demonstration of amodal correspondences between auditory, visual, and articulatory events during speech production does not mean that they are used in perception. There may be other aspects of the behavior, not measured by us, or at least not analyzed that drive perception. Therefore, the relevance of these correspondences for speech perception must be demonstrated. But how?

The simplest way we could think to do this was to use the same amodal production data to generate talking head stimuli whose linguistic relevance can then be validated perceptually. At the time, there were no data-driven talking heads, a study of perceiver eye movement behavior during auditory-visual perception gave us reasons to be optimistic. First was the primary result of that study that linguistically relevant information is not restricted to the mouth region but is distributed everywhere on the lower face (Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998). We interpreted this finding to mean that the visible speech information could not be very detailed spatially, but hypothesized that the temporal information could still be fine-grained, which would compensate for the sparse spatial detail.

Subsequent studies showed that both the spatial and the temporal information effect visual enhancement of speech intelligibility at low resolution (dePaula, et al., 2006; Munhall, Jozan, Kroos, & Vatikiotis-Bateson, 2004).

3.5.1. Overview of Talking Head synthesis and animation

Figure 7 provides the schema for the talking head animation system driven by measured data that was devised principally by Takaaki Kuratate with assistance from Hani Yehia and me. Our goal was to generate talking heads that from the same production data we had used to establish the correspondences between audible and visible speech without introducing errors or linguistically-relevant information extraneous to the original production data.

Since we wanted these heads to be three-dimensional, we did need to collect static structural data for speakers’ heads as well as a set of facial postures that would encompass the range of facial deformations observable during speech. An example of the static structural data, consisting of a high density mesh and a video texture map, is given in the upper left of Figure 7. After experimenting with different postural configurations and set sizes, we settled on nine postures for each speaker (Vatikiotis-Bateson, Kuratate, Kamachi, & Yehia, 1999). Each mesh in the set was calibrated and reduced to roughly 500 nodes, as exemplified in the upper right of Figure 7. From the set of reduced meshes, a mean face mesh and deformation coefficients for each node were computed (center of Figure 7). Time-varying position data collected with OPTOTRAK (Northern Digital, Inc) provided high-precision 3D positions for 17 markers,
shown at the lower left of the figure.

**Figure 7.** Talking head synthesis and animation from structural (upper left) and time-varying performance (lower left) measures. Static 360 degree scans of the head for nine facial postures generated high density 3D meshes and video texture maps. Mesh complexity was reduced from approximately 300,000 nodes to roughly 500 nodes. A mean face mesh and deformation coefficients were computed. Nodes were designated to coincide with the 17 3D position markers (dark points in meshes shown in center and at upper right), and linear estimation was used to calculate the deformation effect of the 17 marker nodes on the remaining 480 or so mesh nodes. Time-varying marker data was then applied frame by frame (at 60 Hz) to the generic mesh (center) resulting in a frame sequence of adapted face meshes and texture maps (lower right). If desired, recorded position and orientation of the head could then be added to the sequence of animated sequences. Finally, the original synchronized sound recording was added to the movie.

The mean mesh was adapted so that each marker would coincide with a mesh node (blackened node points in the central image). Linear estimation coefficients for the remaining 480 nodes were used to set the value of each mesh node for each marker configuration as it changed through time (at 60 Hz). The time-varying mesh was then covered with the texture map from original static recording (upper left) and output as a time-varying talking head, shown at the lower right of the figure (for details, see Kuratate, Vatikiotis-Bateson, & Yehia, 2005; Kuratate, Yehia, & Vatikiotis-Bateson, 1998).

Rigid body head motion data were used to manipulate the orientation and position of the synthesized faces. Adding head motion and the original sound resulted in talking head movies for sentential utterances (see snapshot in Figure 8) that were completely natural in their gross physical characteristics, spatiotemporal behavior, and synchronization of sound and picture.

**Figure 8.** Snapshot of talking head animation using static and dynamic 3D measures.
After validating that the kinematics values input to the talking head synthesis and animation system are not distorted at output, the animations must be validated perceptually. Demonstration that synthesized talking head utterances, such as the one depicted in Figure 8, evoke the same pattern of results as would be obtained using naturally recorded video would constitute perceptual validation. Such validation was obtained in two paradigms. The first was the replication of Sumby and Pollack’s early results (1954) that intelligibility of acoustics presented in masking noise is enhanced when listeners can see the speaker’s face. This was carried out by Jeffrey Jones while at ATR (Jones, unpublished, 2002). The critical conditions in Figure 9 are the audiovisual speech in noise (AV & Noise) and audio only in noise (AO & Noise) conditions. Intelligibility is reliably greater in the multimodal (AV) condition.

![Figure 9. Sumby & Pollack’s (1954) demonstration of visual enhancement of acoustically degraded speech replicated using synthesized talking head sentences.](image)

The second perceptual validation of the talking heads was included in a larger study aimed demonstrating the value of synthetic multimodal stimuli whose source parameters are measured production data and are under full experimenter control (Munhall, Jones, et al., 2004). The results are based on the same 20 Japanese sentence animations used above. Figure 10 shows the effects on word and syllable (hiragana) recovery for various configurations of auditoryvisual stimuli. That all combinations of visual and auditory components are more intelligible than auditory only speech presented in noise is strong evidence that the animations are perceptually valid. That kinematic distortion results in lower intelligibility than either the natural face and head motion or no head conditions suggests that the visual enhancement intelligibility is not merely the result of adding a correlated visual channel. The most interesting finding however, is that the presence of head motion has a profound effect on intelligibility. Why this should be so is not immediately obvious, as the head motion does not convey information about the linguistic content. My hypothesis is that the head motion helps perceivers align themselves to the event structure of the stimulus. This is another form of the coordination outlined at the outset of this talk, but at the level of entrainment between independent entities.
Having established that the talking head animations convey perceptually detectable linguistic information, we can cautiously assert that the audible and visible components of speech production identified and examined by Yehia and colleagues are relevant to multisensory speech perception. Further support for this claim comes from the fact that intelligible acoustics can be synthesized from purely visible motion of the head and face.

Additionally, that the head, whose primary movement role is arguably postural, appears to play a crucial role in multisensory speech perception, suggests that systems of diverse neurophysiology, anatomy, and function can contribute to multisensory event processing. This has obvious extensions to more distal motion systems such as hand gestures and body posture as well as local, but distinct, event systems such as those involved with expression of motion.

There is now a large body of research, too large to review here, that has explored many aspects of these linkages in the perceptual domain for diverse purposes including machine recognition, discourse analysis, alternate linguistic systems such as sign language, and so on. Within my own sphere of research collaborations the persistence of the auditory-visual linkage has been examined for

- **degraded stimulus resolution**: we demonstrate that visual enhancement of speech intelligibility is achieved at low levels of spatial (Munhall, Jozan, et al., 2004) and temporal frequency (dePaula, et al., 2006),

- **production and perception characteristics of Lombard speech** (Lombard, 1911), *speech recorded in noise* (Lombard speech) has different effects on multisensory perception than speech recorded in quiet, but presented in noise (Vatikiotis-Bateson,
Barbosa, Chow, Tan, & Yehia, 2007),

- talker identity conveyed by one modality (visual or auditory) induces identification in the other (Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003), depending not only on consistency of speaking style, not content (Lander, Hill, Kamachi, & Vatikiotis-Bateson, 2007),

- brain function benchmarks for multi-sensory processing of multimodal speech that trivially indicate linguistically-relevant mirror neuron circuitry (Rizzolatti & Arbib, 1998) for auditory-visual speech processing (Callan, et al., 2002), and more importantly suggest that visual enhancement of speech intelligibility may be primarily due to increasing the gain on auditory processing (Callan, Jones, Munhall, Callan, et al., 2004).

4. DYNAMIC ANALYSIS OF CROSS-MODAL CORRESPONDENCE

We now arrive at the final section of the talk in which our approach to dynamic correspondence analysis of time-varying production data is presented. In making the transition from static to dynamic analysis, I do not want to dismiss the value of the earlier static analyses. Even though these analyses required only one space-time constraint – synchrony across data channels and measurement domains for each time sample; and even though correspondences were computed as singular global averages for the span of an utterance, we learned a great deal about the organization of multimodal speech behavior. In particular, we discovered important features of speech production within each measurement domain – the vocal tract, face, head, acoustics, and even muscle EMG – can be characterized with a small number of dimensional degrees-of-freedom, be they independent muscles, orthogonal movement components (PCA), or spectral acoustic parameters (LSP). Similarly, coordination across measurement domains can be characterized by small numbers of orthogonal components (SVD). Using the same production data, we were able to generate linguistically valid talking heads that have allowed us to examine multisensory speech perception in terms of multimodal speech production.

Despite this progress, we still learned nothing really about the dependency of the observed correspondences on the temporal organization of the speech events that give rise to the observed coordination. How much better would it be if correspondence analysis could show how speech and other communicative behaviors are patterned and coordinated in time? In what follows, I present our method for computing sample-by-sample correspondence between signals. I argue that the resulting time-varying correspondence is indicative of two important facts about the organization of biological behavior. First, time-varying fluctuations in measurable correspondence are inherent to non-pathological biological behavior; and second, our ultimate assessment of spatiotemporal coordination must be in terms not of the degree of correspondence, but in the patterning created by the fluctuation of correspondence through time. That is, the moment of correspondence between two signals does not occur at constant phase, and it is these shifts in phase that may index the relevant pattern of coordination.
between two event streams.

### 4.1. Instantaneous correlation analysis

We have created a filter that computes instantaneous correlation between any two signals (Adriano Vilela Barbosa, Yehia, & Vatikiotis-Bateson, submitted). The filter’s basic form is

$$\rho(k) = \frac{S_{xy}(k)}{\sqrt{S_{xx}(k)S_{yy}(k)}}$$

It has two modes:

- **a unidirectional mode that looks at samples on or before the filter point**

  $$S_{xy}(k) = \sum_{l=0}^{\infty} c e^{-\eta l} x(k-l) y(k-l)$$

- **and a bidirectional mode that looks both before and after the filter point**

  $$S_{xy}(k) = \sum_{l=-\infty}^{\infty} c e^{-\eta|l|} x(k-l) y(k-l)$$

Cutoff frequency of the two filters is related by $\eta_2 = 1.5067 \eta_1$ within the range $0 < \eta < 1$.

![Figure 11. Computing instantaneous correlation. Top panel: the signals $x(t) = \sin(\pi t)$ (thick red line) and $y(t) = \cos(\pi t)$ (thin black line). Middle and bottom panels: Instantaneous correlation ($\rho$) using the uni-directional and bi-directional methods for two values of $\eta$ – 0.3 (thick red line) and 0.5 (thin black line).](Image)

The behavior of the two filters is demonstrated in Figure 11 for two sine waves 90 degrees out of phase with each other. Two cutoff frequencies are used. Higher cutoff

---

2. In the original talk, the unidirectional and bidirectional modes were termed **causal** and **non-causal**, respectively, which while consistent with formal terminology have undesirable implications for the type of system being modeled.
frequencies are determined by smaller values of $\eta$. The uni-directional filter introduces a slight rightward shift that increases at higher values of $\eta$, because there is a larger number of prior samples that have non-negligible effects on the instantaneous correlation as the filter moves forward through time.

4.2. Computing and visualizing instantaneous correlation in 2D

Computing instantaneous correlation between two signals, as shown in Figure 11 above, is a big improvement over our previous static method that computed average correlation over a signal span. Technically, it is possible to compute correlation between signals at any temporal offset and, if the signals are well-behaved the way these sine waves are, then a reasonable solution can be obtained so long as you can identify the phase lag between the signals before applying the algorithm. In fact, biological signals are never this well behaved; the correspondence between signals will often involve a temporal offset (guaranteed when dealing with serial behavior) and is always subject to temporal fluctuations. For example, the motion of my lips to produce a /b/ precedes the moment the acoustics deliver the sound we hear as /b/ by 160-180 ms (Abry, Lallouache, & Cathiard, 1996). Since temporal fluctuations are rampant in both music and language, we added temporal offset as the second dimension of a 2D correlation map.

Figure 12 shows the instantaneous cross-correlation between the same two sine waves for temporal offsets 1 sec before and after the zero-offset point. The visualization shows the same alternation of positive and negative correlation depicted in Figure 11 at zero-offset (white solid line) in this figure. The dashed white line at -0.5 sec offset indicates the continuous correlation ($r=1$) that occurs when two sine waves of identical frequency are coincident. Similarly, a perfect negative correlation would be obtained at 0.5 sec offset that occurs when two identical sine waves are 180 degrees out of phase.
Turning now to less well-behaved case of biological behavior, Figure 13 shows the behavior of the two filter types at two cutoff frequencies when applied to two measures of visible head and face motion during speech production. The two measurement techniques were 1) marker tracking of a sparsely distributed set of 23 blue paper dots (Adriano Vilela Barbosa & Vatikiotis-Bateson, 2006), and 2) video-based recovery of pixel velocity using optical flow analysis (OFA) summed over a region of interest (ROI) that included the head and face (Adriano V. Barbosa, Yehia, & Vatikiotis-Bateson, 2008; Horn & Schunk, 1981).

As can be seen in the upper panel, the time series of dot-marker velocity and the summed optical flow amplitudes are quite similar. The OFA results are not quite as sensitive as marker tracking at low amplitudes, but is more sensitive when there is more motion. Still, if the two measurement techniques are both reliable, there should be a high positive correspondence between the two signals at zero offset. This is what we observe, especially for the smaller value of $\eta$, which is less sensitive to small local differences between the two signals. Since the kinematic accuracy of the dot-tracking method has already been verified, the strong correspondence between the two signals overall suggests that head and face motion are being reliably recovered from the video image sequence.

The differential effects of $\eta$ are shown more clearly in the 2D representation given in Figure 14. Here there is not only a strong correspondence between the two signals at zero offset, but also stability in the iterative pattern of the utterance. That the motion of one syllable is highly correlated with syllables preceding and following it is clearly represented in the bottom and middle panels by the ladders of alternating positive and negative correlation.

Finally, comparison of the three panels suggests that visualization of the 2D correlation map tells different aspects of the story depending on the $\eta$-values chosen.
Figure 14. Instantaneous correlation between two measurements of the same head and face motion of the two signals. Instantaneous correlation of the two head and face motion signals shown in Figure 13 is plotted as a function of time and temporal offset for three values of η (0.1, 0.3, 0.5) and a ±0.3 sec range of temporal offsets. All three η values show the strong positive correspondence between the two signals at zero offset.

4.3. Language as performance: Speech and gesture

Not so many years ago, it would have been heretical within the formal domain of linguistics to insist on the importance of analyzing performance to understand language structure. Through the years there have been a few voices insisting that language and gesture should be examined together (e.g., McNeill, 1992). Dwight Bolinger (1975, Ch. 2) went so far as to claim that language is gesture. Combining traditional field linguistic transcription and analytic techniques with correspondence analysis of multi-camera recordings, we have begun to examine language as performance in context of indigenous languages of North America and Africa. In what follows, I describe some of our work with Plains Cree, a language spoken in the Western Provinces (Alberta, Manitoba, Saskatchewan) of Canada, and Shona which is the principal language spoken in Zimbabwe. Although brief, these descriptions usefully describe the potential usefulness of our techniques in assessing coordination across diverse physical, as well as signal domains.

4.3.1. Speech and gesture in Plains Cree

In work with colleagues Rose-Marie Déchaîne and Joseph Deschamps, we have begun to examine how linguistic structure and physical expression are coordinated during various types of language performance in Plains Cree. Using data from multi-camera recordings of interviews and conversations with Cree elders (see Figure 15), our instantaneous correlation and linguistic analyses have shown that syntactic phrasing is coordinated with and at least co-dependent on gestural phrasing. We have also found multiple instances of gestural anaphora, where an elaborate gestural sequence such as bringing the extended hand up to the head while talking about
thinking is subsequently indicated by a different and much smaller gesture. Finally, we have found examples of semantic collusion where visible gestures of hands and face and spoken words are both needed to convey certain meanings, particularly in the treatment of concepts and systems alien to earlier Cree culture and language (Déchaine, Vliet, Cardinal, Barbosa, & Vatikiotis-Bateson, 2008).

Figure 15. Sample of multicamera recording of Plains Cree. White boxes define Regions of Interest for calculating summed motion of the head-face (left) and hands (right) using optical flow analysis.

Figure 16 shows samples of the signals that were extracted from the multi-camera video recording of the Cree speaker. These include the summed motions over regions of interest defined for the head and face and for the hands (see Figure 15), and the acoustic amplitude (RMS). Needless to say, it is difficult to discern much correspondence between the signals by simply looking at the time-series in Figure 16. In fact, there is not the kind of persistent high correlation that was seen in Figure 13 for the comparison of the two face motion measurement techniques. However, when the 2D correlation map is computed, what correspondence does exist is readily apparent, as shown in Figure 17.

Figure 16. Measured signals used in correspondence analyses shown in Figure 17. From top: audio waveform for the Cree speaker; acoustic amplitude (RMS); summed face+head motion; and summed hand motion. Motion was summed for the regions of interest shown in Figure 15.
Our main interest in analyzing these data was to determine the correspondence through time between the speech acoustic signal, the head and face motion that our previous analyses had shown were globally correlated with the acoustics, and motion of the hands which we know to be involved in communicatively-relevant gesturing. What is presented here is preliminary analysis intended only to show the potential of the 2D correlation method in making such assessments; more detailed linguistic-gestural analyses are underway. The time course of the speech acoustics visible motions of the face, potentially separate coordination of speech acoustics with orofacial and head motion, on the one hand, and hand motion, on the other.

The correlations shown in Figure 17 show that, at least for this stretch of speech, the head + face and hands are correlated with RMS amplitude at different times with different temporal (phase) offsets, and to different extents. The face x RMS correlation shows that when there is a positive correspondence, there RMS lags the head and face motion by about 150 msec, which fits well with our expectation that visible indicators of speech articulation, particularly for consonant constrictions precede their acoustic consequences, which are not evidenced in acoustic amplitude until the onset of a following vowel. The stronger correlation occurring late in the sample at a 300 msec offset is the result of an emphatic head movement at the beginning of a new phrase. If the motion data were factored using PCA, we would predict a primary mode corresponding to head motion that would be correlated with RMS amplitude at the larger (300 msec) temporal offset, and a residual that corresponding to lower face motion that would be correlated with the smaller (150 msec) temporal offset.

The hands x RMS correlation, shown in the middle panel of Figure 17, tells quite a...
different story. Although I cannot yet explain every region of positive correspondence, it is clear that there is a relatively continuous weak correlation very close to zero offset. There is no physical constraint preventing such synchrony as there appears to be for visible components of speech articulation and the resulting acoustics. Again, however, there are instances of stronger correspondences, occurring early and late in the sample, where the hands either precede (200 msec) or follow (300 msec) the corresponding period of RMS amplitude.

Finally, the face x hands correlation shows an early correspondence at -50 msec. This equals the difference between the face x RMS and the hands x RMS correlations in this part of the sample, and suggests that for these moments the face, head, and hands are all coordinated. Later in the sample, there is only a weak correspondence at the 300 msec offset, which does not indicate a shared correspondence across the three domains.

4.3.2. Speech and gesture in Shona

Another analysis of language as performance, similar to the one underway for Plains Cree is focuses on story-telling in Shona – a Bantu language spoken in Zimbabwe. Preliminary results (Déchaine, Barbosa, Mudzingwa, & Vatikiotis-Bateson, 2008) suggest that, while there may be a superset of motion patterns that may be recruited during spoken communication, their use is idiosyncratic to specific individuals and shows no language-dependent specification. This does not preclude specific cultural and social factors from determining which gestures are used and how. Figure 18 shows a single frame of a multi-camera recording of a married couple during a session where they each recounted their versions of three stories. Just as the accounts differed between the two speakers, so too did their gestural repertoire and use. As an interesting aside, the man’s gestures were very similar to that of the Plains Cree speaker depicted above. Although differing in culture and language, there may be temperamental factors – nervousness, anxiety, etc. – similar to both male speakers that play a larger role in structuring gesture use than language, culture, or even gender.

Figure 18. Story-telling in Shona (Zimbabwe) during a multi-camera recording.

One obvious analysis to be carried out with multi-talker recordings of this sort is to compute interspeaker coordination as a means of assessing entrainment between two speakers. Such an analysis does not make as much sense with two people as familiar
with each other as a married couple, but with two speakers just meeting for the first

time, there are numerous dimensions of possible coordination in which one

speaker takes on production characteristics of the other – acoustic properties, choice

of words (lexical), phrasing (syntax), and gestural.

4.4. Coordination among (thousands of) individuals at a rock concert

At the outset of this talk, I asserted that coordination is inherent to all biological

behavior and does not require conscious agency. A.T. Winfree’s marvelous book on

biological time is a tour de force of support for these simple truths (1980). I have tried

to show in this talk that linguistic communication depends on coordinated action

within the individual to produce coherent signals for others to perceive. In passing,

I have also suggested that coordination between individuals is also essential during

communication. In addition to coordinating which words are used or how they are

phrased, there are obvious constraints on turn-taking that shape discourse and myriad

other factors that determine its effectiveness.

Language, however, is not the only highly skilled means that humans have for

coordinating – and communicating – with one another. Music has been ubiquitous in

human culture, though recently not necessarily in every human’s experience, and

has served many roles in influencing people’s feelings, moods, and behavior. In this

section, I consider the role of communication in controlling coordination between a

singer and his audience and within the audience itself on a large scale.

Commercially available video footage of the Queen Live Aid 1985 concert held in

Wembley Stadium (Wembley, UK) is being used to assess how the singer, Freddie

Mercury, was so successful at directing an audience of 72,000 to interact with him and

the music in specific and highly coordinated ways.

Figure 19. Bird’s eye view of the 72,000 member audience and stage for the Queen Live Aid 1985 concert held

at Wembley, UK. Note the relatively small monitor screen to the left of the stage and even smaller screen to

the right of the stage.
In what follows, two examples of Mercury’s control of the audience are examined. It is clear that his goal is to engage the audience in the music physically. What is particularly interesting is that he varies the style of engagement from song to song throughout the concert, and within each song he appears to know exactly how long the audience can sustain the high level of coordination after it is attained.

4.4.1. Coordinated voice

In a brief rendition of the Afro-Caribbean song, Day-O, Mercury engages the audience in an astonishing phrase-by-phrase, sequence of call and mimicked response. Unlike the monotonously-repeated motion sequence used in Radio Gaga (see below), the challenge to the audience is to repeat vocal phrases of increasing duration and complex pitch and timing. Figure 20 shows Freddie Mercury calling out to the audience during Day-O. His arm motions indicated both pitch and duration (rhythmic timing in motion).

![Freddie Mercury](image)

**Figure 20.** Freddie Mercury “calling out” a phrase to be imitated by the audience during Day-O.

This is confirmed by instantaneous correlation, shown in Figure 21, between the visible motion (almost entirely Freddie Mercury) and the acoustic RMS amplitude (alternating between singer and audience). Mercury begins a sustained production of “day- o” at 15 sec, using his arm motion both to punctuate the beginning and end of the phrase and to mark the steady rise in pitch during its course. His motions are correlated not only with his own fluctuating acoustic amplitude, but also that of the audience response which begins at 21 sec, indicative of almost the same degree of precision in singer and audience.
Figure 21. Performance measures (top) and 2D correlation (bottom) of body motion (singer) and acoustic RMS (singer and crowd).

Figure 22 shows the auto-correlation of the RMS amplitude across a ±8 sec range. Two interesting features emerge. First is the high degree of temporal coherence of vocalizations even when separated by large amounts of time. Second is the failure of the 2D correlation algorithm to calculate a correspondence during the sustained call of Mercury (15-20 sec) and response of the audience (21-26 sec). The onsets and offsets of singer and audience are highly correlated, but the filter used for this analysis had too low a cutoff frequency to track the rapid fluctuations that might have sustained the correlation estimate.

Figure 22. The 2D correlation algorithm used to compute auto-correlation on RMS amplitude for ±8 sec offset. Note the gap in correspondence during the two sustained vocalizations (15-20 sec; 21-26 sec).

4.4.2. Coordinated motion

During his performance of Radio Gaga, Mercury gets the audience to produce a simple movement sequence in synchrony with his own movements and the rhythm of the song. Figure 23 shows Mercury showing the audience the shape and timing of the movement gesture he wants to see produced. Correlation analysis was carried out on
three parts of the song: early training, middle confirmation, and late finale, as shown in Figures 24-26.

Figure 23. The first two of three segments analyzed consist primarily of movement sequences with verbal instructions occasionally inserted between stanzas of the song. The finale segment is not shown.

Figure 24. Optical flow data and analysis of Radio Gaga, early training phase. Top left: Sample image for analysis showing ROIs for assessing coordination within the audience. Top right: image depicting optical flow analysis where motion amplitude is represented by intensity – large motion white, little or no motion black. Lower panels, top to bottom: time-series plotting summed pixel velocities for the left and right ROIs; 1D instantaneous correlation of two audience ROIs at zero-offset (1D); 2D instantaneous correlation for an offset range of ±0.5 sec.

As can be seen in the 1D and 2D correlation plots in Figure 24, motion of the two sections of audience under analysis is tightly coordinated around zero-offset. Corresponding to the dark band at zero-offset, the 1D instantaneous correlation never dropped below r=.87. Although some of the correspondence may incorporate noise, at least as much signal failed to be captured due to lighting model effects and the low resolution of the commercially distributed digital video after down sampling and digital compression (MPEG-1).
Figure 25. Optical flow data and analysis of Radio Gaga, middle confirmation phase. Top left: Sample image for analysis showing ROIs for assessing coordination within the audience. Top right: image depicting optical flow analysis where motion amplitude is represented by intensity – large motion white, little or no motion black. Lower panels, top to bottom: time-series plotting summed pixel velocities for the left and right ROIs; 1D instantaneous correlation of two audience ROIs at zero-offset (1D); 2D instantaneous correlation for an offset range of ±0.5 sec.

Figure 25 shows the correlation between audience sections to be even stronger about 40 seconds into the song. This can be seen in the higher 1D correlation (r>.93) and the distinct narrow band of correspondence along zero-offset in the 2D correspondence. This level of audience coordination is sustained until at least the two-minute mark in the song and is, I believe, what Mercury intended to achieve through his carefully structured combination of repeated demonstration of the intended gestural pattern and verbal commands.

Figure 26. Optical flow data and analysis of Radio Gaga, middle confirmation phase. Top left: Sample image for analysis showing ROIs for assessing coordination within the audience. Top right: image depicting optical flow analysis where motion amplitude is represented by intensity – large motion white, little or no motion black. Lower panels, top to bottom: time-series plotting summed pixel velocities for the left and right ROIs; 1D instantaneous correlation of two audience ROIs at zero-offset (1D); 2D instantaneous correlation for an offset range of ±0.5 sec.
Late in the song the audience appears to begin losing its concentration as indicated in Figure 26 by correspondence values almost identical to those seen at the beginning of the song. Although purely speculative at this point, the rise and eventual fall of coordination could be limited by attentional constraints. These are constraints that Mercury understood and manipulated masterfully in his performances by keeping song length short at about three minutes and by changing the modality of entrainment between concurrent (Radio Gaga) and serial (Day-O) coordination.

4.4.3. Coordination Summary

The preceding excursion was intended merely to convey how various aspects of spatial and temporal coordination during a complex performance can be quantified and visualized. Together, numbers and pictures can help ground behavioral interactions within an audience (Radio Gaga) and between audience and performer (Day-O) conceptually in terms of coordination, concurrency, and synchrony. For reasons of space, the quantitative analysis of the performer-audience was not presented for Radio Gaga. Had I done so, you would have seen a similar progression of coordination for the three analysis segments, with the added nuance that during the middle confirmation segment Freddie Mercury’s motion gestures reduce their precision while taking on symbolic characteristics analogous to anaphora in language or conducting in musical ensemble. For example, once the audience knew what it was supposed to do and had been drilled by Mercury, rather than raising his arm and holding it raised for the full musical beat as he did during the earlier training phase, he needed only to raise his arm at the proper onset and lower it again almost immediately.

This last point connects with an even more crucial factor in the success of Mercury’s concert; namely, its dependence on the tremendous common knowledge of the music by both musicians and audience. Concurrent coordination to the degree observed in these performances cannot be achieved if the audience does not already have extensive knowledge of the music and possibly even prior knowledge of the kind of interactive performance that will occur. That is, the audience’s knowledge makes it possible to anticipate the content from moment to moment; all that is required is a conductor/choir master.

Also missing from the analysis is the coordination between Mercury and the other musicians and, of course, the acoustic correspondence to the motion behavior of the sort shown for language performance in earlier sections. We hope to have these analyses completed soon, but you should by now be able to imagine how these analyses will look and what they might show.

I interpreted the rise and fall of concurrent coordination within the audience as perhaps due to attentional limitations. There is a large literature on attention and performance that does consider issues such as prior knowledge or familiarity, but not in terms of constraints on coordination. This particular case of coordination of a large crowd that can hear the source, but not necessarily see it (i.e., the 40,000 people or so that are too far...
away to clearly discern Mercury’s stage antics), adds another dimension of complexity; clearly some people can watch Mercury’s movements while hearing the music, but others who cannot see Mercury must rely on the audience around them. This means that the attentional domain is perceptually multidimensional as it is also multisensory; people moving to the audible music are not only watching whatever they can see from their locations, but also becoming coordinated by physically bumping into one another.

It is obvious that large numbers of people enjoy coordinating with others. It is a form of connectedness that suggests intimacy without long-term commitment. Indeed, if my interpretation of attentional loading is correct, long-term coordination of this sort would not be sustainable or even desirable given the difficulty. However, there is another aspect of coordination that emerges when watching Mercury direct the actions of such a large audience. The audience to some extent at least must surrender to the control imposed by the music and its leadership. Such surrender of control may be cathartic in short episodes, but why would anyone want to make it a habit? To understand this we must shift our attention to the musicians that make such larger scale coordination possible in the first place.

4.5. Enthusiastic coordination, concurrency, and synchrony: Clarinet trio

At a simple approximation, music appears to be about achieving some form of rapture through extremes of control in time and space (e.g., pitch). Musicians submit to rigorous training in which they develop and hone precise neuromotor and perceptual skills that enable them to reproduce subtle rhythmic patterns and complex melodic sequences. Even more extraordinary to my mind is the ability of musicians to create music collaboratively in which each musician contributes a component that on its own has no resemblance to the musicality of the final product. Put another way, each part follows a rhythmically defined, patterned series of notes that when played alone, may not be pleasant to hear. Yet, when combined as the composer hopes, the individual parts fuse into a beautiful event sequence.

This is exemplified by Professor Mauricio Loureiro and two students of the UFMG music school in their rendition of Mozart’s Divertimento in B flat Major KV. 229 No. 3 for three Basset Horns (1783) played on 2 soprano clarinets and bass clarinet. To demonstrate the points just made, they will play three segments: First, they will play the entire piece which is 2-3 minutes long; Second they will play the final page (Figure 28) of the piece individually one instrument at a time; Finally, they will play the final page together. Figure 27 shows the trio.
As you heard in the performance (or can hear from the recording posted on the IEAT website), when the trio play the piece *ensemble*, it is difficult to isolate the component instruments, even when two are *soprano* and one is *bass*. Then, when the instruments individually play the measures on the final page, only the principal clarinet (Prof. Loureiro) is melodic and reminiscent of the piece; the others appear, at times, to consist of randomly spaced short sequences. Finally, when the last page is played again *ensemble*, the independence of the components heard in the second phase persists simultaneously with the impression of seamless fusion. That is, the correspondence between components exists even when played in isolation because the musicians are capable of such high levels of precise control in reproducing their component parts as specified. At the same time, the musical event of interest is emergent in the fusion of separate pieces.

But what about that precision? As with speech or any other form of behavioral production, no musician plays the same piece twice exactly the same way. Furthermore, it is well-known that if performers follow the prescribed tempo and rhythm too precisely, it gives the music a wooden or mechanical feel. Indeed, what musicians do instead is deviate from the prescribed pattern, ever so slightly perhaps, but in ways that allow them to interact with one another. This appears to be the realm of musical precision and artistry. Currently, my colleagues and I at UBC, in collaboration with Professors Loureiro, Yehia, and their students at UFMG are attempting to systematically examine the spatial and temporal characteristics of musical expression using the measurement and analytic tools of *optical flow analysis* and recurrent estimation of *instantaneous correlation*, described in this talk. My ultimate goal is to find a common structure for communicative expression in skilled behavior be it music, dance, speech, or simply walking down the street.
5. CONCLUSION

I have covered a lot of ground in this talk, most of it fleetingly either in the interest of time or because I do not yet have much to say on specific topics. However, I think I may have said enough about the conceptual and practical challenges that need to be addressed in order to understand the production and perception of communicative behavior to at least pique your interest and lay a coherent foundation for considering these issues. We know that human behavior is coordinated internally and externally. Interaction between and among people, whether intentionally communicative or not, is extraordinarily coordinated in space, time, and cognitive dimensions well beyond the scope of this talk, but minimally we can take heart that the time-varying properties of behavior have a great deal in common at any level of observation. It is just that some assumptions such as the isomorphism between perception and production may need to be reconsidered. It is my sincere hope that some day we can demonstrate how similar patterns of coordination are regardless of the specific behavior or behavioral domain.

I have given substantial weight in this talk to our tools for assessing spatial and temporal coordination. In many ways, they are surely primitive and must eventually be replaced with improved algorithms. However, their simplicity ensures a certain longevity in that they can be applied to any behavior for which two or more signals exist; therefore, they can already be applied to most time-varying problems. Even more primitive and simple, our use of optical flow analysis to recover motion from video will surely be integrated with more sophisticated techniques that entail object identification and tracking in 3D. Especially promising are the active appearance models (AAM) being developed by Iain Matthews, Simon Lucey and colleagues (e.g., Lucey, Wang, Cox,
Sridharan, & Cohn, 2009; Matthews, Cootes, Bangham, Cox, & Harvey, 2002). These models use prior knowledge to increase robust identification over a wide range of transformation. However, our simple optical flow techniques already free us from the constraints of the laboratory environment, enabling us to analyze naturally occurring behavior. Thankfully, most communicative interaction occurs in stable environments and concerns about where the performers are play a minor role. Nevertheless, more measuring naturalistic behavior does raise the issue of how to conduct quantitative analysis of non-repeated events. Uniqueness is anathema to traditional psychometric techniques in which everything hinges on calculating variance and determining whether it is signal or noise.

Finally, a huge amount of work remains to be done to properly assess the linkage between perception and production. I argued in this talk that treating them as non-isomorphic is essential to error-based learning. I think the difference is also essential to stabilizing the perception of stable patterns in fluctuating behavior, which ultimately makes coordination between biological systems possible. But these are but a few dimensions of the problem amongst many.
6. Bibliography


Kuratate, T., Vatikiotis-Bateson, E., & Yehia, H. C. (2005). Estimation and animation...


