Making communicative performance relevant. A commentary

ERIC VATIKIOTIS-BATESON

University of British Columbia

1. Introduction

I have been asked to comment on three papers whose common link is perhaps best described as expanding the scope of linguistically relevant analysis to include factors critical to spoken communication, but previously regarded as extraneous to understanding linguistic behavior and organization. The papers by Borràs-Comes and Prieto and by Swerts in this issue focus on visual behaviors – *gestures* – that occur while speaking and that are of equal or greater importance in conveying certain types of meaning during spoken interaction than the acoustics. The approaches of the two papers are quite different: Borràs-Comes and Prieto attempt a nuts and bolts demonstration, using careful manipulation of multisensory stimuli, that perceivers can rely more heavily on visual information from the face than on the acoustic information for parsing the difference between statements and short *echo* questions verifying something previously stated. Swerts, on the other hand, demonstrates the efficacy of expressive gestures for perceiving meaning using stimuli extracted from more natural spoken interactions. The third paper, by Walker and Hay (2011), shows that performance factors, indicative for example of a speaker's age, may be critical to understanding the organization of the lexicon. While Walker and Hay's work explores only acoustic performance parameters, the same ingenious methodology can easily be extended to audio-visual performance.

It is not my purpose in what follows to criticize these three papers, nor particularly to say how wonderful they are – that's not my style anyway. Rather, I concentrate more on the conceptual and empirical paths that they individually and, to some extent, collectively suggest for future research. As was beautifully demonstrated by the breadth of interpretation engendered by the word *gesture* in the thematic description of the Twelfth Conference on Laboratory Phonology where these papers were first presented, empirical studies of speech are largely free, for the moment, of the fetters of theoretical and even meta-theoretical (methodological) dogma. I hope to make clear in what follows that the research in each of these three papers represents a gentle departure from tradition, but points the way to new questions in what for Laboratory Phonology is largely uncharted territory. The

JLP 3 (2012), 211–219
DOI 10.1515/lp-2012-0012

1868-6346/12/03–211

remainder of this commentary consists of discussions of each paper separately, followed by a more general discussion.

2. Correlates of social awareness in the visual prosody of growing children (Swerts)

As can be inferred from his paper title (repeated above as the section title), Swerts uses expressive visual behavior to assess the development of social awareness in children. Three studies are presented that combine controlled elicitations of spontaneous interactions with perceptual evaluations indicative of how they are being interpreted. The basic hypothesis is that as children become more socially aware, they gain communicative control over their expression of feelings and emotions such as 1) their level of uncertainty, 2) how positive or negative they feel about something, and 3) whether or not they are being truthful or deceitful. Thus, developing social awareness is indexed by increasing communicative sophistication as inferred from perceiver ratings of uncertainty, truthfulness, and so on.

Although there is no question that this communicative control will be played out in both audible and visual modalities, Swerts' concentration on visual prosody (exclusively in two of his three studies) makes good sense. Politicians, for example, have long known that it is easier to delivers statements of questionable veracity to a radio than to a television audience (*personal communication*, anonymous U.S. presidential press aide 1972); conveniently, political announcements made in the morning are still likely to be heard (e.g., in the car radio) and accommodated before they are viewed on the evening news. Experimental evidence that audiovisual sensitivity to deception is more acute than audio-only sensitivity (Ekman 1985, Hollien 2002) bears this out, as do the results of Swerts' third study. Thus, the first two experiments, examining *uncertainty* and *positive-negative* affect, focus on the visual aspects of expression.

2.1. *Evaluating uncertainty*

The first study reported by Swerts examines the match up between one's own confidence or certainty about something (*feeling of knowing* – FOK) and the judgment of others (*feeling of another's knowing* – FOAK). The expressions produced by children and adults while providing short (one word) answers were judged by both children and adults. Swerts does not tell us explicitly whether the stimuli were visual-only or audiovisual; the evaluation could be made with either type of stimulus presentation. Not surprising perhaps is the finding that adult evaluations of certainty (FOAK) match the original FOK values better for both adults and children. The child judges gave more neutral judgments (see Swerts, Table 1), but appeared to line up a little better with the adults when judging their

age-peers. Also, the adult evaluations are more extreme than those of children suggesting they can make more accurate assessments of their certainty than can children.

As someone always anxious to seek explanations for perception results in the stimulus source, my first inclination with results like these is to analyze the expressions themselves to see whether there is some aspect of the production that adults use to signal certainty, such as less variability or the presence of some distinguishing feature (caricature). I would also want to analyze the acoustics and do cross-domain correlation analysis with the visible behavior of the face and head (á la Barbosa et al. 2012; Yehia et al. 2002). If there is no measurable difference between the production patterns of children and adults, this would suggest that we just get better at interpreting our behavior. There might always be an age-peer advantage in making judgments of certainty, applicable to both children and adults and predicted by the much greater amount of time adults and school-age children spend with their peers rather than with one another.

2.2. *Evaluating positive and negative affect*

Using video clips showing the reaction of a child to winning or losing a card game, adults were asked simply to identify a win or loss. Stimuli were made from video of children of two age groups: 8 and 12 year olds. Taking the high accuracy of win-loss identification of the 8 year olds as a reference, the interesting result is that by age 12 Dutch kids have become more adept at masking their emotional response, especially to winning. Unlike expressions of certainty, which may entail more of a second-order, 'certainty about certainty' than a systematic change of behavior, this masking changes the produced affect, and should be easily measurable both visually and acoustically (but not with this stimulus set since the reaction is not necessarily audible).

As Swerts suggests, the scope of this study should be expanded to include additional age groups: at least one between the 8 and 12 year old groups tested and two more age levels above 12 to better understand the socialized masking of emotion that occurs. Interesting possibilities to pursue are whether the masking involves damping the affective behavior – such as general reduction of gestural amplitude – or a more complex overlay of a new type of signal that either diffuses the affect or distracts perceiver attention. Whichever form the masking takes, it likely has both visible and audible consequences.

2.3. *Evaluating deception*

Of the three experimental tasks, this was by far the most difficult for generating clean stimuli as it required children to produce 'cooperative' truthful and deceptive utterances, confounded by other emotions such as fear of the 'dragon' to whom they had to lie and despite the fact that they were talking to cartoon characters. The

child participants in this study were also younger, 5–6 (4.10–6.4) years, than the 7–8 year olds of the other two studies. Adults had to judge the veracity of a child's response to questions such as *Where is the prince?* (Answer: *in the castle*) when presented either visually, acoustically, or audio-visually.

The near chance judgment of audio-only stimuli is no surprise, but neither of the visual conditions produced very strong results. Rather, the 5–6 year old age group appears to be transitional. When divided into older and younger halves (either side of the mean age), deceptions are easier to detect in the slightly older children. Again, analysis of the production data would help determine whether or not younger children are in fact producing discernible differences in their utterances. I suspect that how children conceptualize lying at this age is a more important predictor. Telling a lie and concealing the fact of it are not the same thing. The game context may contribute to the somewhat fuzzy results by emphasizing the importance of telling the lie well, which cannot be appreciated if it is concealed. That is, the game provides a context where children are supposed to tell a lie (presumably not the norm for them). If they recognize this as a structural feature of the game, then their need to demonstrate competence at lying, which indicates their game skill, will counter-indicate the experimental aim of assessing their ability to conceal lies.

3. 'Seeing tunes.' The role of visual gestures in tune interpretation (Borràs-Comes & Prieto)

In their paper, Borràs-Comes and Prieto compare the roles of visual and acoustic signals in distinguishing contrastive focus statements and one-word, echo questions through perceptual evaluation of synthetic and natural stimuli. Their main goal is to demonstrate that perceivers depend more on visual than acoustic information for both identification and processing (reaction time). In the first of their two studies, an 11-step continuum is synthesized for the "pitch" frequencies of the accented syllable of the Catalan word, *pe**ti**ta,* with recorded focus and echo question exemplars at either end of the continuum. These 11 acoustic pitch steps were then spliced together with a video clip showing contrastive focus production and a video showing echo question production. As shown by their Figure 6, the visual information dominated perceiver identification of the audio-visual utterances.

Reaction time predictably increases as the acoustic and visual signals become more incongruent, as shown in their Figure 7. However, where visual information dominated the identification scores, the reaction time result is less clear and the two factor ANOVA performed on video (2 levels) and acoustics (11 levels) is too imbalanced to be entirely trustworthy. Perhaps to level the field a bit, a second study was run with a square (6 × 6) matrix of acoustic and visual signals, synthesized to generate equidistant steps between the two pairs of original signals for

contrastive focus and echo question. This study replicated the identification results of the first, again suggesting a dominant role for the visual component of the stimulus (see Figure 8).

The reaction time results are predictably more noisy for the second study than the first. Perceivers are being given the same number of conditions to identify, but they must deal with more variability in the visual domain. The reduction of the number of acoustic increments undoubtedly has less effect (11 to 6, or *too many to many*) than the increase from 2 to 6 visual increments. Reaction times are on the whole faster in the second study, presumably because the same subjects were used in both studies and the second study came second. This compression makes it more difficult to demonstrate a difference between acoustic and visual contributions to reaction time, but the ANOVA again shows no main effect for the acoustics. The interaction between acoustic and visual stimuli is more complicated than that of the first study, because the effects on reaction time are not just a matter of the congruency of information in two modalities. In the second study, the audiovisual stimuli compiled from the mid-range of the two continua probably lack clear information in either modality. Audio-only and visual-only evaluations of the component stimulus continua would help here.

Far from being discouraging, the somewhat mushy results of this study raise questions that may be more interesting than deciding the relative contribution of visual and acoustic information. For example, and this relates to Swerts' interest in visual prosody, how much sensitivity to subtle differences in visual patterning can we muster? Do we attune to gestural features such as a particular set of the eyebrows or a timed head excursion, analogous to the popular notion of targets; or is it more a matter of spatial and temporal patterning in which information is streamed over an array of functionally coupled components, whose exact values are never quite the same from one episode to the next?

Regarding the coupling of acoustic and visual signals, the linear morphing techniques for creating the continua were the same in that the same number of equidistant steps were created; but this does not mean that a 1.2 tone increment in fundamental frequency is equivalent to the visual morphing step-size. It is quite possible that the synthetic acoustic and visual stimuli in the middle of the continua were more mismatched than casual inspection revealed. There are now techniques such as the one pioneered by Kuratate and Yehia (Kuratate et al. 2005) for generating kinematically accurate talking heads that would inform the veracity of intermediate steps in synthetic continua.

Of course, reaction time may be problematic as the sole measure of stimulus processing. The results of the second study already suggest there may be multiple effects on reaction time and that the correspondence between processing load and distance from an expected multisensory form is anything but simple. The results may have been confounded by subjects' prior experience with the continua of the first study. This is easily tested, but I think it is actually a good thing to thoroughly familiarize subjects with the task, thereby reducing the volatility of their

reactions and potentially increasing the trustworthiness of the observed differences in performance.

4. Congruence between 'word age' and 'voice age' facilitates lexical access (Walker & Hay)

According to Walker and Hay, not even the lexicon is safe from socio-phonetic influences. I, for one, am relieved that a lifetime of experience with words in context can actually facilitate processing isolated words in a lexical decision (word/non-word) task.

In their study, a moderately young range of perceivers were subjected to lexical and plausible non-lexical items spoken in isolation by a "young" speaker (age 22) and an "older" speaker (age 50). The lexical items were chosen from two corpora recorded about 10 years apart. The earlier corpus, the Intermediate Archive (IA), contained older speakers born 1890–1930; the later, the Canterbury Corpus (CC), comprised speakers born 1930–1984. Subsets of stimuli were identified representing typical older forms in the IA corpus, typical younger forms in the CC, and a group of less determinate age common to both corpora. Rounding out the stimulus set was a subset of closely matching nonsense forms.

Glossing over the enormous amount of work that went into setting this study up and all the various analyses conducted, the study generated two key results. First, listeners had little trouble making the right decision regardless of the congruency of speaker age and the era of the word – error rates were very low across the board, but they were lower when the speaker and word were matched in age. Second, lexical decision was faster (reaction time) when the age of the voice matched the era of the word.

Despite careful bias testing throughout the study, the age range of the listeners was skewed toward the younger end (the oldest listener was 48 years old), and this might explain why error rates for nonsense words, which were twice as high as for real words, were reliably lower for the younger voice. Unfortunately, the possibility that age matching between producer and perceiver adds another layer of the same ilk as that being tested cannot be addressed simply by increasing the age range of listeners. Older listeners show signs of processing lexical items differently, indicative, for example, of changes in hemispheric organization (Rastatter & McGuire, 1992). Another, more testable, at-source age factor might be the age of the person(s) constructing the list of closely matching nonce forms. An elderly person might well have generated phonotactically similar nonsense words differently and/or influenced by other age-related factors than the young person(s) who made the list used here.

Fine-tuning the methodology is (and probably should be) a never-ending chore, but this study opens the door to a vast array of studies aimed at fleshing out the implications that the lexicon may be structured more like a busy dance hall, where

some characters are more actively engaged than others and identifiable by style of dress and behavior, than as a repository for minimally distinct forms waiting to be asked out. These corpora can be mined for other acoustically conveyed differences such as the socioeconomic match between word form and speaker, likely effects of word familiarity on fluency, not to mention a host of visual and audiovisual probes of the link between production and perception. Obviously, corpora such as these collected in New Zealand cannot easily be linked to others collected for other dialects of English elsewhere or other languages, but the techniques used to exercise them can be usefully shared and improved upon with an eye to asking the same questions of any data set.

5. Final remarks

The three papers discussed here each expand the scope of linguistic analysis to include other experimental domains. Swerts brings to the table a clever methodology for examining development of linguistic communication skills in children, focusing heavily on the communicative importance of visual behavior patterns. Borràs-Comes and Prieto also draw our attention to the visual domain at the level of processing prosodic-semantic links, but with an overarching emphasis on speech being essentially multimodal. Walker and Hay jolt our view of the lexicon by adding new dimensions of experiential texture. Like Swerts, they appear to have a grand plan and have developed an ingenious methodology that can be easily extended to ask a host of new questions about lexical representation.

It is great to see impressive progress being made in areas that challenge the traditional primacy of phonemics, an activity based solely on acoustic perception (or so the old-timers would claim). Visible behaviors were never considered properly linguistic, they were paralinguistic – interesting and maybe important, but basically unanalyzable except insofar as there were acoustic correspondences in the prosody; and American structuralists like Hockett or Twaddell, who began their essays with the exultant reminder that language is crucially social, never took it any further. Dwight Bolinger bucked this trend and that of the new generation of generative linguists, who were happier working from judgments of linguistic competence than recordings of actual behavior, by claiming not just that gesture is important to the study of language, but that *language is gesture* (Bolinger 1975, Chapter 2). Strictly speaking, even linguistic studies of prosody have never been properly integrated with segment-based approaches to phonetics and phonology.

Now, the role of visual behavior in conveying meaning is more obvious. There is a burgeoning body of evidence suggesting that, redundant or not, spoken communication is multimodal and involves processing streams of information in chunks that are not easily reducible to a unit so mysteriously simple as the phoneme. Couple this with the expanded scope of formal studies of meaning that

necessarily go beyond the word and sentence even, and the likelihood that our language is not functionally or structurally codified independently of our experience as language learners and users, and I think it is clear that exciting times are ahead.

Of course, there are still vestiges of our phonemics roots. Without launching into a historical overview of modern phonology, it strikes me as odd that seemingly everything that cannot be subjected to a phonemic analysis is hierarchically aligned as *sub*-phonemic, somehow beneath, rather than outside of or beyond the phonemic domain. Walker and Hay happen to be the only ones to use this term among the three papers discussed here, but I hear it from students and others who, when asked what it means, say "you know, talker characteristics and all the other cool stuff about speech." This may not be the most pressing complaint in the discipline to raise; however, the connotation of *sub*-phonemic strikes me as unfortunate when the voice age effects reported by Walker and Hay most likely entail subtle differences in word intonation, formerly a suprasegmental phenomenon, and differences in vowel quality, which minimally have syllable-sized consequences. These three papers are ample and exciting proof that there is nothing "sub" about the wide range of visual and audio-visual gestures that inform our understanding of emotion, truthfulness, syntactic function of words, or the lexicon. These studies open up a welcome expanse of research questions that acknowledges the intricate and crucial interplay among a host of multimodal performance factors that goes far beyond a Procrustean hierarchy in which the phoneme alone plays a crucial part.

Acknowledgement

I am grateful to Caroline Smith and Ian Maddieson, the organizers of the Twelfth Conference on Laboratory Phonology, for inviting me to participate in the conference and their subsequent patience; and to Laurel Fais for invaluable editing and intellectual support. Supported by grants from SSHRC and NSERC (Canada).

Correspondence e-mail address: Eric.Vatikiotis-Bateson@ubc.ca

References

Barbosa, A. V., Rose-Marie Déchaine, Yehia, H. C., & Erci Vatikiotis-Bateson. 2012. Quantifying time-varying coordination of multimodal speech signals using Correlation Map Analysis. *Journal of the Acoustic Society of America*, 131(3). 2162–2172.

Bolinger, Dwight. 1975. *Aspects of language* (2nd ed.). New York: Harcourt Brace Jovanovich.

Ekman, Paul. 1985. *Telling lies: clues to deceit in the marketplace, politics, and marriage*. New York: W. W. Norton & Company.

Hollien, Harry. 2002. *Forensic voice identification*. San Diego: Academic Press.

Kuratate, Takaaki, Eric Vatikiotis-Bateson, & Hana Camille Yehia. 2005. Estimation and animation of faces using facial motion mapping and a 3D face database. In John G. Clement & Murray K. Marks (eds.), *Computer-graphic facial reconstruction*, 325–346. Burlington, MA: Elsevier Academic Press.

Rastatter, Michael P., & Richard A. McGuire. 1992. Hemispheric processing strategies for lexical decisions among elderly persons. *Perceptual Motor Skill*s 75(3). 1275–1280.

Walker, Abby & Jennifer Hay. 2011. Congruence between 'word age' and 'voice age' facilitates lexical access. *Laboratory Phonology* 2(1). 219–237.

Yehia, Hana Camille, Takaaki Kuratate, & Eric Vatikiotis-Bateson. 2002. Linking facial animation, head motion, and speech acoustics. *Journal of Phonetics* 30(3). 555–568.