AN ANALYSIS OF PERCEPTUAL DEPENDENCIES IN AUDIOVISUAL SPEECH PERCEPTION: A NEW APPROACH

Nicholas Altieri, Noah Silbert, and James T. Townsend
Indiana University, Nicholas Altieri: naltieri@indiana.edu, Noah Silbert: nosilber@indiana.edu, and James T. Townsend: jtownsen@indiana.edu

Abstract

Ecological speech signals consist of both auditory and visual information. Determining whether the dimensions of perception, including the auditory and visual components of speech, are combined independently (e.g., Garner & Morton, 1969) is an important problem in cognitive psychology. To investigate whether perceptual dependencies occur in congruent and incongruent audiovisual speech, we implemented the statistical methodology of General Recognition Theory (GRT; Ashby & Townsend, 1986), a multidimensional extension of signal detection theory. We carried out an identification experiment where the auditorily and visually articulated syllables /be/ and /ge/ were combined in a 2 x 2 design to yield four stimulus categories: (A_V) /be_be/, /be_ge/, /ge_be/, and /ge_ge/. The stimuli /be_ge/ and /ge_be/ elicit the classic McGurk fusions of de and bge respectively. Results obtained from model fitting suggest that the auditory and visual components of speech are generally perceived independently, although dependencies can arise with the presentation of incongruent stimuli. Marginal d's and decision criteria also differ as a function of stimulus level.

Auditory and visual speech signals consist of multiple dimensions. Simple auditory signals such as sinusoidal waves vary on dimensions such as frequency and duration, while more complex signals, including speech sounds, vary on other more complex dimensions. Visual speech signals, obtained by lip-reading, also consist of multiple dimensions. Information obtained from visual speech signals provides useful supplementary information to the auditory signal, containing salient information about place of articulation (Summerfield, 1987).

An important issue in cognitive psychology concerns how perceptual dimensions are combined during the stages of information processing (Garner & Morton, 1969; Ashby & Townsend, 1986). A problem relevant to multisensory integration in speech perception is whether the perceptual effect evoked by the auditory component of the speech stimulus is perceived independently of the effect evoked by the visual component. Informally speaking, *perceptual independence* (PI) holds when the perception of each stimulus component *A* and *V* in stimulus *AV* are not contingent upon the perception of the other stimulus component (see Ashby & Townsend, 1986). To offer a more rigorous definition, PI holds when probability of perceiving stimulus *AV* is equal to the product of the probabilities of perceiving *A* and *V* separately (Garner & Morton, 1969).

As Ashby and Townsend (1986) illustrated, the dimensions of perception are not directly observable and therefore powerful mathematically driven techniques are needed to properly investigate claims regarding inter-dimensional dependencies. The authors developed a framework known as General Recognition Theory (GRT)—a multidimensional extension of signal detection theory (Green & Swets, 1966). It is assumed in GRT that the presentation of a multidimensional stimulus, for example, auditory /b/ with visual /b/, elicits a random perceptual effect in multidimensional space. Over time, perceptual effects aggregate according to probability distributions in perceptual space, and this space is partitioned by decision bounds. In its most general form, GRT does not make parametric assumptions.

The issue of dependencies and interactions between the auditory and visual components of the speech signal has implications for current models of audiovisual integration. Braida's (1991) Pre-labeling Model of Integration (PRE), a signal detection approach for consonant identification, predicts the absence of correlation between the various dimensions in the audiovisual perceptual distribution. Garnering support in favor or against such assumptions has significant practical and theoretical implications.

Application of GRT to Audiovisual Speech Perception

As with unidimensional signal detection theory, the GRT assumes that observers obtain information from a stimulus and use the perceived value to determine the nature of the stimulus. GRT has the potential to characterize and assess both perceptual and decisional dependencies that exist between the stimulus dimensions. First, stimulus components A and V are *Perceptually Separable* if the perceptual effect of the auditory component, for example, does not depend on the level of the visual component. Secondly, *Decisional Separability* holds when the decision bounds in each dimension are parallel to the x and y axes (i.e., decisions about one dimension do not depend on the other dimension). Finally, *Perceptual Independence* of dimensions within a single stimulus holds in components A and V of stimulus A_iV_j when the perceptual effects of A and V are stochastically independent. Figure 1 shows a pictorial example of violations of Perceptual Separability, Decisional Separability, and Perceptual Independence within the context of a 2 x 2 complete factorial design (described below)

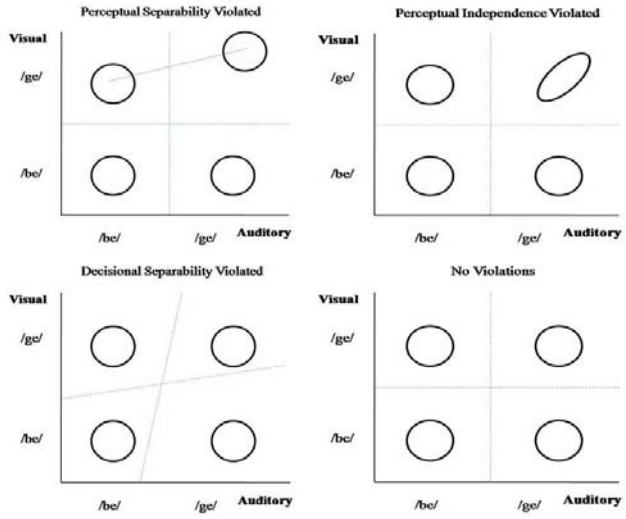

Figure 1: The upper left panel shows a violation of perceptual separability with the participant reporting V [g] more often when it is combined with A /g/ than with A /b/. The panel on the bottom left shows response regions changing size due to failures of decisional separability. The panel on the upper rights shows a failure of perceptual independence in the category A /g/ + V [g]. Finally, the bottom right hand portion of the figure shows the case involving the absence of any violation.

Two non-parametric tests can be carried out using the data from identification-confusion matrices: *sampling independence* (SI) and *marginal response invariance* (MRI). Both tests have proven useful for assessing perceptual and decisional dependencies (Ashby & Townsend, 1986; Thomas, 2001). First, *sampling independence* in stimulus A_iV_j holds if and only if

$$P(a_k v_l \mid A_i V_j) = [P(a_k v_1 \mid A_i V_j) + P(a_k v_2 \mid A_i V_j)] \bullet [P(a_1 v_l \mid A_i V_j) + P(a_2 v_l \mid A_i V_j)] \quad (1)$$

$P(a_k v_l \mid A_i V_j)$ denotes the probability of responding $a_k v_l$ given the presentation of stimulus $A_i V_j$. SI holds if the probability of reporting that the auditory and visual stimuli are at levels *k* and *l*, respectively is equal to the marginal probability of reporting the auditory stimulus at level *k* times the marginal probability of reporting the visual stimulus at level *l*.

Ashby and Townsend (1986) proved that if both perceptual independence and decisional separability hold, SI is implied. That is, if SI fails, then either perceptual independence or decisional separability fails (or both). MRI holds when the probability of correctly identifying one stimulus, say the auditory, does not depend of the level of the other stimulus (i.e., the visual) (see Ashby & Townsend, 1986; Thomas, 2001). MRI holds for component A_i, $i = 1, 2$, if equation 2 holds:

$$P(a_i v_1 \mid A_i V_1) + P(a_i v_2 \mid A_i V_1) = P(a_i v_1 \mid A_i V_2) + P(a_i v_2 \mid A_i V_2) \quad (2)$$

Similarly for component V_j, $j = 1, 2$:

$$P(a_1 v_j \mid A_1 V_j) + P(a_2 v_j \mid A_1 V_j) = P(a_1 v_j \mid A_2 V_j) + P(a_2 v_j \mid A_2 V_j) \quad (3)$$

The sum of the probabilities on the left hand side of (2), for example, denotes the proportion of correct responses to the auditory stimulus when the visual component of the stimulus is at level 1, while the right hand side denotes the probability of a correct response in the auditory channel when the visual component of the stimulus is at level 2 (equation (3) shows the analogous case for the visual channel).

The simplest GRT experimental protocol in which perceptual independence, perceptual separability, and decisional separability can be simultaneously assessed uses a complete factorial identification paradigm in which each of the experimental stimuli is constructed by combining 2 levels on each of 2 dimensions (see Ashby & Townsend, 1986). A complete factorial experiment using auditory and visual speech stimuli can be readily implemented by combining two auditory consonants, /b/ and /g/, with the same two visual visemes, producing the stimulus set {*A* /b/ *V* /b/; *A* /b/ *V* /g/; *A* /g/ *V* /b/; and *A* /g/ *V* /g/}. This experimental set-up has the advantage of using both incongruent McGurk stimuli (i.e., *A* /b/ and *V* /g/) which produces the fusion /d/ in virtually all adult listeners (McGurk & MacDonald, 1976) as well as congruent audiovisual stimuli, allowing the experimenter to investigate the occurrence of perceptual dependencies in 'normal' stimuli and McGurk stimuli.

Methods

Participants

Eight participants with normal or corrected vision were paid ten dollars for their participation.

Stimuli

Four audiovisual stimuli were created by factorially combining the syllables /be/ and /ge/: /be_be/, /be_ge/, /ge_be/, and /ge_/ge/. The audio, visual, and audiovisual files were edited and the auditory and visual components were combined using Final Cut Pro HD version 4.5. The audio files were sampled at a rate of 48 kHz at a size of 16 bits. The auditory components were mixed in white noise in order to reduce the auditory signal-to-noise ratio to -10db. The brightness of the video files was reduced 90 steps using the brightness video filter

in Final Cut Pro. The duration of the auditory and visual stimuli was approximately 180 milliseconds.

Design and Procedure

Participants were seated 14 to 18 inches in front of a Macintosh computer equipped with Beyer Dynamic-100 headphones. Each trial began with a fixation cross (+) appearing in the center of the monitor for 500ms followed immediately by one of the four possible stimuli. The timer began at the stimulus onset after the fixation cross disappeared from the screen. Participants were instructed to respond as quickly and accurately as possible by pressing the button that corresponded to the presented stimulus (i.e., respond by pressing "de" if /be_ge/ was presented and "be" if /be_be/ was presented). Participants were required to perceive perceptual fusions when the auditory and visual components of the stimulus were incongruent (i.e., "de" if /be_ge/ was presented, and "bge" if /ge_be/ was presented). The congruent stimuli simply required a "be" or "ge" response. There was a 750-millisecond delay between trials.

The experiment consisted of 400 trials with 100 trials presented from each of the four stimulus categories. Participants were run for 2 blocks consisting of 200 trials with a break scheduled between each block. Participants also received 50 practice trials at the onset of each experimental session that were not included in the subsequent data analysis. The experimental session lasted approximately 40-60 minutes.

Results and Discussion

The total proportion correct for /be_be/ was .91, for /be_ge/ it was .84, for /ge_be/ it was .77, and for /ge_ge/ it was .84. Posterior distributions of identification-confusion probabilities were used to assess MRI and SI. MRI is assessed via the Δ statistic, defined as the difference between the left and right sides of equation 2 (or 3). SI is assessed via the π statistic, defined as the difference between the left and right sides of equation 1. The solid lines indicate the (marginal) 99% highest probability density regions of the posterior distributions of Δ (Figure 2) and π (Figure 3) statistics. The dotted lines in Figures 2 and 3 indicate the location of zero (i.e., the presence of MRI or SI). All panels (except the bottom right) indicate Δ (Figure 2) or π (Figure 3) values for individual subjects; the bottom right panels indicate group level values. Simulation results indicate that statistical power is similar to frequentist analyses.

MRI holds for auditory /b/ for all subjects except S4 and S6; both failures are due to a greater probability of responding 'auditory /b/' when presented with visual /b/ than when presented with visual /g/. MRI holds for auditory /g/ for all subjects except S2 and S3; both failures are due to a greater probability of responding 'auditory /g/' when presented with visual /g/ than when presented with visual /b/. MRI holds for visual /b/ for all subjects except S5 and S7; both failures are due to a greater probability of responding 'visual /b/' when presented with auditory /b/ than when presented with auditory /g/. Finally, MRI holds for visual /g/ for all subjects. MRI only fails at the group level for auditory /g/. The fact that MRI holds for at least one level on each dimension for every subject suggests that decisional separability holds across the board for both dimensions; each failure of MRI can be accounted for by failure of perceptual separability such that salience varies on one dimension across levels of the other dimension.

If decisional separability holds, as suggested by the analyses of MRI, then any failures of SI can be interpreted as failures of perceptual independence. There are only a small number of such failures here. SI fails for A /b/ V /b/ stimulus for S4, for the A /b/ V /g/ stimulus for S3 and S8. However, in a number other cases, zero is very near the limit of the

99% highest probability density region of the π statistic, indicating a near failure of SI. In each case, SI fails due to a greater proportion of the relevant perceptual distribution falling in the correct response region than in the adjacent partially correct regions (relative to a case in which perceptual independence holds).

Marginal Response Invariance

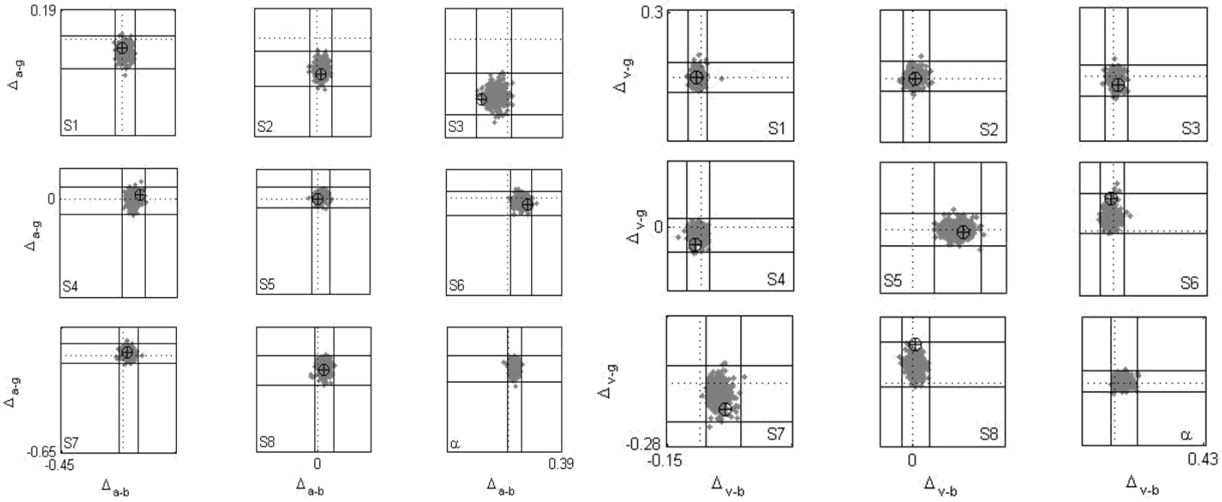

Figure 2: Posterior distributions (gray dots) and observed values of Δ, assessing MRI.

Sampling Independence

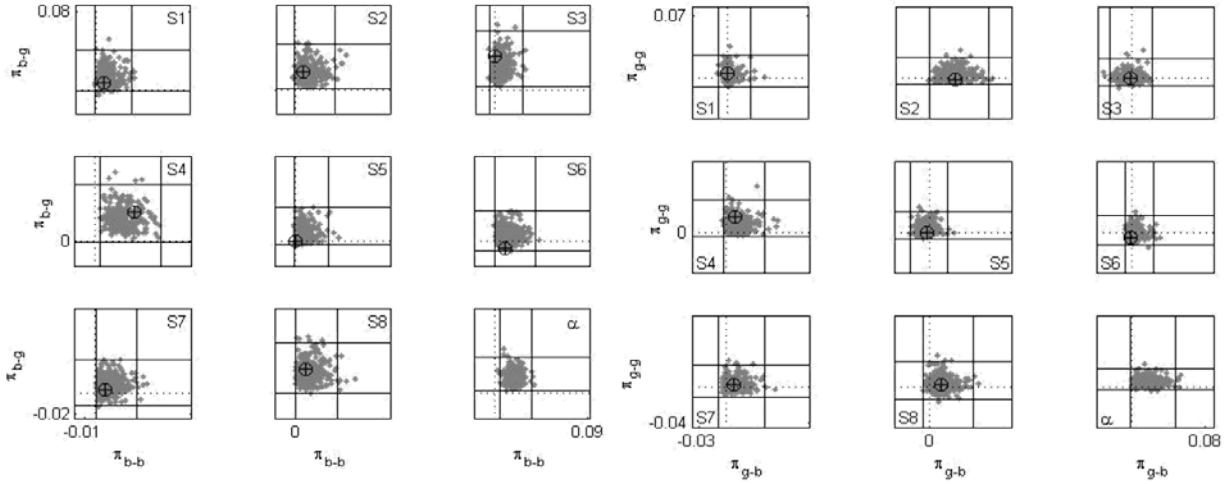

Figure 3: Posterior distributions (gray dots) and observed values of π assessing SI.

Summary and Conclusion

This paper represents a preliminary assessment, using a robust mathematical framework, of how listeners integrate multidimensional speech stimuli. Specifically, we addressed the question of whether congruent and incongruent audiovisual speech stimuli are perceived independently. That is, is a correlation between the perceptual evidence from the auditory and

visual dimensions responsible for producing the McGurk effect? We observed failures of MRI suggesting that salience of one dimension varies across levels of the other dimension in perceptual space. We also found that perceptual independence fails occasionally in audiovisual speech stimuli, though not necessarily for categories involving perceptual fusions. These results have potential implications of models of audiovisual integration, particularly signal detection models (e.g., Braida, 1991) containing parameters that allow for correlation between perceptual dimensions.

Acknowledgements

This study was supported by the National Institute of Health (Grant No. DC-00111) and the NIH Speech Training Grant (No. DC-00012), and by NIMH 057717-07 and AFOSR FA9550-07-1-0078 grants to J.T.T. We wish to acknowledge personnel in James Townsend's and David Pisoni's laboratories for discussion and insightful comments.

References

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93,* 154-179.

Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. *The Quarterly Journal of Experimental Psychology, 43A*(3), 647-677.

Garner, W. R., & Morton, J. (1969). Perceptual independence: Definitions, models, and experimental paradigms. *Psychological Bulletin, 72,* 233-259.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: John Wiley & Sons.

McGurk, H., & MacDonald, J. W. (1976). Hearing lips and seeing voices. *Nature, 264,* 746-748.

Summerfield, Q (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *The Psychology of Lip-Reading* (pp. 3-50). Hillsdale, NJ: LEA.

Thomas, R. D. (2001). Characterizing perceptual interactions. In M.J. Wenger & J.T. Townsend (Eds.), *Computational, Geometric, and Process Perspectives on Facial Recognition* (pp. 193-227). Mahwah, NJ: Lawrence Erlbaum Associates.