Published in final edited form as:
J Exp Psychol Hum Percept Perform. 2011 August ; 37(4): 1193–1209. doi:10.1037/a0023100.

Psychophysics of the McGurk and Other Audiovisual Speech Integration Effects

Jintao Jiang and **Lynne E. Bernstein**
Division of Communication and Auditory Neuroscience, House Ear Institute, Los Angeles, California 90057, USA

Abstract

When the auditory and visual components of spoken audiovisual nonsense syllables are mismatched, perceivers produce four different types of perceptual responses, auditory correct, visual correct, fusion (the so-called *McGurk effect*), and combination (i.e., two consonants are reported). Here, quantitative measures were developed to account for the distribution of types of perceptual responses to 384 different stimuli from four talkers. The measures included mutual information, the presented acoustic signal versus the acoustic signal recorded with the presented video, and the correlation between the presented acoustic and video stimuli. In Experiment 1, open-set perceptual responses were obtained for acoustic /bA/ or /lA/ dubbed to video /bA, dA, gA, vA, zA, lA, wA, ΔA/. The talker, the video syllable, and the acoustic syllable significantly influenced the type of response. In Experiment 2, the best predictors of response category proportions were a subset of the physical stimulus measures, with the variance accounted for in the perceptual response category proportions between 17% and 52%. That audiovisual stimulus relationships can account for response distributions supports the possibility that internal representations are based on modality-specific stimulus relationships.

Keywords

audiovisual speech perception; congruent and incongruent; quantitative stimulus measures; factor analysis

When perceivers can see as well as hear talkers, speech perception is audiovisual (AV) (e.g., Massaro, 1987; McGurk & MacDonald, 1976; Sumby & Pollack, 1954). Many experiments have been carried out with mismatched AV stimuli in an effort to understand the perceptual integration of auditory and visual components of AV speech stimuli. In such experiment, four main perceptual effects are obtained: The fusion effect or *McGurk effect* (McGurk & MacDonald, 1976) is said to have occurred when, for example, a video /gA/ and an acoustic /bA/ stimulus are presented together, and perceivers report hearing /dA/, a different syllable. That is, the perceptual response is different from the response to either of the uni-sensory stimuli. Numerous studies have shown fusion effects (Green & Kuhl, 1989; Green, Kuhl, Meltzoff, & Stevens, 1991; MacDonald & McGurk, 1978; Manuel, Repp, Studdert-Kennedy, & Liberman, 1983; Massaro, 1987; Sekiyama & Tohkura, 1991; Summerfield &

Correspondence Author: Jintao Jiang, Ph.D., Speech and Hearing Sciences Department, 801 22nd St NW Rome 550, George Washington University, Washington, DC 20052, USA, jjiang@gwu.edu, Phone: 1-202-994-7414.

McGrath, 1984). The combination effect occurs when both the acoustic and video stimuli are perceived sequentially (McGurk & MacDonald, 1976). For example, pairing acoustic /gi/ with a video /bi/ can result in the perceived combination /bgi/ (Green & Norrix, 1997). Mismatched stimuli can also result in either auditory correct or visual correct responses.

We hypothesized that the different types of perceptual effects are a consequence of a perceptual processing system that has learned the normal relationships between auditory and visual stimulus components (Bernstein, Lu, & Jiang, 2008). When that relationship is disrupted by mismatching the stimulus components, systematic changes occur in perceptual responses. However, in order to test this hypothesis, the systematic physical stimulus relationships between components need to be characterized. Stimulus characterizations typically have been perceptual and not quantitative (e.g., Braida, 1991; Grant & Braida, 1991; Massaro, 1998). That is, stimuli are typically described in terms of phoneme identities of component auditory and visual stimuli, and in terms of the match or mismatch between the components. Gross physical stimulus characteristics, such as duration, acoustic amplitude, and video image resolution, are among the few quantitative measures that are commonly reported, although these characteristics are not usually directly relevant to the hypothesis being tested. Even quantitative models of AV perceptual integration have mostly, to our knowledge, used perceptual responses, such as, auditory-only, visual-only, and AV phoneme confusion matrices, as input to the models (Braida, 1991; Grant & Braida, 1991; Massaro, 1998). Stimulus onset asynchrony has been one of the few quantitatively measured independent variables in AV speech research (Massaro, Cohen, & Smeele, 1996; McGrath & Summerfield, 1985; van Wassenhove, Grant, & Poeppel, 2007), but internal features of the stimulus components are typically not quantified (c.f., Munhall, Gribble, Sacco, & Ward, 1996).

An influential theoretical paper by Summerfield (1987) posited that the core of a comprehensive account of AV speech integration would be a perceptually relevant metric of AV speech stimuli. Summerfield proposed as possible metrics, (1) auditory and visual phonetic features (the only qualitative stimulus characterization), (2) the acoustic filter function of the vocal tract, (3) vectors representing the magnitudes of independent acoustical and optical parameters, (4) successive static vocal-tract configurations, and (5) time-varying kinematic patterns. Summerfield also reasoned that these metrics were mostly inadequate. Integration in terms of phonetic features was inadequate, because the empirical evidence led to *ad hoc* rules for integration, depending on whether the features were visually or auditorally presented. The notion that the acoustic vocal tract transfer function could be derived from viewing a talker was deemed inadequate, because too much of speech production cannot be seen. An alternative suggestion was that a look-up table with canonical acoustic values could be accessed through vision, because perceivers "'know' the audio-visual [sic] structure of phonemes." But this metric was noted to have several disadvantages, among them the arbitrary supposition that evidence from vision is translated into auditory representations. The vector combination of acoustic and optical parameters in terms of acoustic spectra and two-dimensional visual images of the lips and teeth viewed from a single direction were criticized because of, for example, the problem of non-invariance of optical signals with diverse views of the talker. The fifth theoretical metric was based on speech dynamics: "If talkers communicated simply by oscillating their lips at different rates, we should have no difficulty in describing audio-visual [sic] speech perception as the integration of audible and visual evidence of modulation rate" (p. 40). Kinematic descriptors implying speech dynamics were proposed as a modality-free representation, but they were also viewed as problematic. For example, "the mapping [to dynamics] depends on the position of the tongue, and recovering dynamic parameters from natural speech acoustics will be at best computationally complex, and at worst intractable." Although all of the metrics were viewed as problematic, the notion of a common metric or representation to

explain AV speech perception been extensively discussed in the literature (e.g., Bernstein, Auer, & Moore, 2004; Fowler, 2004; Green, et al., 1991; Massaro, 1998; Rosenblum, 2008).

The fact that AV stimuli are rarely quantified for perceptual experiments is not because the stimuli have never been quantified. Studies have been carried out to quantify speech signal relationships between measures of the auditory and visual stimulus components (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009; Craig, van Lieshout, & Wong, 2008; Jiang, Alwan, Keating, Auer, & Bernstein, 2002; Yamamoto, Nakamura, & Shikano, 1998; Yehia, Rubin, & Vatikiotis-Bateson, 1998). What has generally not been done is to relate such measures directly to perceptual responses (c.f., Bernstein, et al., 2008).

The Current Study

We hypothesized that speech perceivers have learned quantifiable relationships between auditory and visual speech stimulus components in natural speech stimuli (Bernstein, et al., 2008). Quantities that can be shown to be systematically present in the relationship between stimulus components are candidates for what has been implicitly learned by a perceiver with normal hearing and vision. Here, we examined quantitative relationships between the auditory and visual speech components of mismatched AV stimuli, and we attempted to account for the corresponding relative proportions of occurrence of the four frequently reported categories of perceptual responses (i.e., fusion, combination, auditory correct, and visual correct). We first report on an open-set perceptual identification experiment in which many matched and mismatched AV consonant-vowel (CV) stimuli were presented (Experiment 1); and we then report on an experiment in which quantitative measures of AV speech relationships were obtained, and their perceptual relevance to the results in the first experiment was evaluated (Experiment 2).

The stimuli selected for this study were a large number of physically different auditory and visual stimuli that could challenge the AV signal measures to account for the four response categories. Stimuli were generated through dubbing an acoustic syllable (/bA/ or /lA/) to each video syllable in a set of eight syllables whose consonants spanned different places and manners of articulation (/bA, dA, gA, vA, zA, lA, wA, ΔA/) for each of four talkers, according to three different temporal alignment methods. There were two tokens for each acoustic and video syllable pair, leading to 384 unique AV syllable pairs (8 videos x 2 acoustic x 4 talkers x 3 alignments x 2 pairings), for which open-set consonant identifications were obtained.

Experiment 1

A large set of mismatched and matched AV nonsense CV syllables was presented in an open-set identification task. Participants reported what phonemes they heard, which were categorized into the four frequently reported categories of perceptual responses (i.e., fusion, combination, auditory correct, and visual correct). Results showed that proportions of the perceptual response categories varied reliably with the acoustic stimulus token, the video token, the talker, and the specific AV stimulus pairing.

Method

Participants—Participants were ten adults (ages between 19 and 29 years, mean age 22 years; five females) with self-reported normal hearing, American English as a native language, and normal or corrected-to-normal vision (with visual acuity of 20/30 or better screened with a standard Snellen chart). Participants gave informed consent, and the study was approved by an Institutional Review Board at the Saint Vincent Medical Center in Los Angeles, California.

Speech materials—Stimuli were extracted from a large speech database (Bernstein, Auer, Chaney, Alwan, & Keating, 2000; Jiang, Auer, Alwan, Keating, & Bernstein, 2007). Four talkers (two females) with American English as a native language and significant differences in visual intelligibility (Jiang, et al., 2002) produced the stimuli. A visual intelligibility score was computed on the four talkers. Talker M2 was the most intelligible, with a score of 8.6, M1 and F2 were intermediate, with scores respectively of 3.6 and 6.6, and F1 was the least, with a score of 1.0. Each talker contributed two tokens of eight voiced CV syllables with varied place and manner of articulation, /bA, dA, gA, vA, zA, lA, wA, ΔA/. The two tokens were labeled '1' and '2'.

The talkers were recorded looking directly into a camera and a teleprompter, and their faces filled the picture. Lighting was from both sides and slightly below the talker's head. A production quality camera (Sony DXC-D30 digital) and video recorder (Sony UVW 1800), a directional Sennheiser microphone, and an amplifier were used to make the videorecordings. A separate DAT recorder was used to obtain acoustic signals (sampling frequency of 44.1 kHz), which were used for acoustic measures. Optical recordings obtained simultaneously with the videorecordings used a three-camera, three-dimensional optical motion capture system (Qualisys MCU120/240 Hz CCD Imager), which digitally recorded the positions of passive retro-reflectors during infrared flashes at a rate of 120 Hz (see Figure 1). All of the recorded data streams were synchronized (Jiang, et al., 2002).

AV stimulus alignment—To generate the stimuli, an alignment between the acoustic and video stimulus components had to be established. In the literature, mismatched AV speech signals are typically aligned on consonant onsets (e.g., acoustic bursts) (Grant, Greenberg, Poeppel, & van Wassenhove, 2004; Massaro, et al., 1996; Munhall, et al., 1996). In the current study, consonants of different durations were used (e.g., acoustic /bA/ versus video /vA/), and knowledge about how to align them was not available. Here, three different temporal alignments were generated for each AV token in order to investigate alignment effects. The alignments involved first manually labeling acoustic consonant and vowel onsets for each speech token (see Figure 2).

One of the three alignments was based on a minimum distance measure. It was obtained by first computing the acoustic 16^{th} order acoustic line spectral pairs (LSPs) (Craig, et al., 2008; Jiang, et al., 2002; Sugamura & Itakura, 1986; Yehia, et al., 1998) of the presented acoustic and the associated *phantom* (the acoustic token recorded with the presented video) acoustic LSPs, and then obtaining the minimum Euclidean distance between the two sets of LSPs, as one stimulus was slid temporally against the other.

For the second type of alignment, the presented and the phantom acoustic signals were aligned at the two consonant onset points (the dubbed video was in its natural temporal relationship with the phantom). For the third type of alignment, the presented acoustic and the phantom acoustic signals were aligned at the vowel onset point.

Every AV stimulus—including congruent/bA/ and /lA/—resulted from dubbing a video $Consonant_2$-/A/ to an acoustic /b_1A/ and /l_1A/ and a video $Consonant_1$-/A/ to acoustic /b_2A/ and /l_2A/. But signals were never dubbed across talkers. In total, there were 384 stimuli (8 video x 2 acoustic x 2 pairings x 3 alignments x 4 talkers).

Stimulus preparation—The original BETACAM video recordings were digitized using an ACCOM real-time digital disk recorder. The sequence of uncompressed video frames for each stimulus were cropped to have an image size of 740 × 480 pixels and were built into an AVI (Audio Video Interleave) file that was compressed using the LIGOS LSX MPEG-Compressor (all **I** frames; frame rate, 29.97 Hz; and bitrate, 7700 Kbits/sec). These video

clips were concatenated to create a single large video file that was authored to a DVD using the SONIC ReelDVD. The corresponding acoustic tokens (48 kHz) were normalized (based on average root-mean-square levels derived from *A*-weighted spectra). All of the audio files were concatenated into a single long file using custom software that ensured frame-locked audio of 8008 samples per 5 video frames. The resulting DVD was a single sequential program chain, which is required by a Pioneer DVD player to allow frame-based searching and random access. In addition, a 1-minute, 1-kHz tone was included on the DVD for the purpose of sound level calibration.

Procedure—Instruction on the open-set consonant identification task and a practice set of 16 trials were given on Day 1. The instructions were to *watch* and *listen to* the talkers, and then identify the consonant or consonants that were *heard*. Participants were shown orthographic representations for the 23 English consonants, /y, w, r, l, m, n, p, t, k, b, d, g, h, T, Δ, s, z, f, v, Σ, Z, tΣ, dZ/, for which sh = /Σ/, zh = /Z/, ch = /tΣ/, j = /dZ/, dh = /Δ/, and th = /T/. The two consonants (/Δ/ and /T/) have identical spelling in English. Responses could include consonant clusters.

A computer program presented each of the AV stimuli and recorded the participant's open-set response. Following each stimulus, a black frame was displayed on the video monitor, and an input box was displayed on the PC monitor. Responses were entered using a computer keyboard followed by pushing the "ENTER" key to obtain the next token. No feedback was given at any time. Participants were instructed to report any mistyping during breaks. Stimuli were presented using a Pioneer DVD player (V7400) and were displayed on a 14" high quality SONY Trinitron PVM monitor at a distance of about one meter. Acoustic stimuli were presented over calibrated TDH-49 headphones at a level of 65 dB SPL that was checked before and after each session.

The acoustic /bA/ and /lA/ tokens in the current study were distinctly different, and their differences might have drawn attention to the acoustic stimuli whenever they were presented in the same stimulus blocks, potentially biasing responses. Therefore, half of the participants received the stimuli blocked by the acoustic /bA/ and /lA/ (blocked design), and the others received the stimuli pseudo-randomly mixed (mixed design). In the mixed design, the 384 tokens were blocked by talkers, and each block comprised 96 tokens for which the audio was /bA/ and /lA/. Each of the blocks, took approximately 10 minutes to complete. In the blocked design, the 384 tokens were first blocked by talkers and then by audio (/bA/ or /lA/). Each block comprised 48 AV tokens from one talker with either audio /bA/ or /lA/, and completion took approximately 5 minutes. Talker order was assigned randomly in each session. In the blocked design, the order of the blocks was randomized within each talker. Within each block, the tokens were randomly ordered. A five-minute break was given between blocks. Over ten sessions, participants contributed ten open-set responses for each stimulus token.

At the completion of the experiment, there were several reports of detecting mismatched stimuli. The mismatches were $A_{bA}V_{wA}$ (four participants), $A_{bA}V_{fA}$ (one participant), $A_{lA}V_{mA}$ (one), $A_{bA}V_{\Pi A}$ (one), $A_{lA}V_{wA}$ (one), $A_{lA}V_{nA}$, and $A_{mA}V_{blA}$ (one). One participant rarely noticed any mismatch, and another participant noticed the mismatches but could not give an example.

Results and Discussion

The open-set responses were tallied, and Figure 3A summarizes separately for each talker the phoneme response proportions for the CVs with audio /bA/, $A_{bA}V$ (99.6% of responses), and audio /lA/, $A_{lA}V$ (99.7% of responses). The figure shows that the response proportions varied across the two audio phonemes, with the largest proportions of responses to $A_{bA}V$

stimuli being /b, T, v, d, Δ/ and the largest proportions of responses to $A_{lA}V$ stimuli being /l, bl, vl/. Of the 23 consonants of English, six (/Σ, Z, tΣ, dZ, y, h/) were never given as responses. $A_{bA}V$ responses mostly comprised a single consonant. $A_{lA}V$ responses were frequently /lA/, and those that were not were mostly combinations (e.g., "bl"). These combination responses were not symmetric, for example, "blA" but not "lbA" was obtained for $A_{lA}V_{bA}$.

Figure 3B shows the consonant identification response category proportions separately for each AV stimulus type and talker. This figure shows that distributions varied across the different types of mismatched stimuli. For example, $A_{bA}V_{dA}$ and $A_{bA}V_{gA}$ resulted in many /Δ/ responses. But $A_{bA}V_{zA}$ resulted in many /d/ responses. In addition, response proportions varied across talkers. The proportion of fusion responses to $A_{bA}V_{gA}$ was similar to what has been previously reported (McGurk & MacDonald, 1976).

Figure 4A–B shows the pooled data as proportions of the response categories, *auditory correct* (e.g., the response to $A_{bA}V_{wA}$ was /bA/), *visual correct* (e.g., the response to $A_{bA}V_{vA}$ was /vA/), *combination* of two or more consonants (e.g., the response to $A_{lA}V_{bA}$ was /blA/), and *fusion* (e.g., the responses to $A_{bA}V_{gA}$ and $A_{bA}V_{zA}$ were /dA/ and /ΔA/, respectively). The data were pooled across the three alignments, two AV pairings, and ten participants. Each response contributed only one count. That is, correct congruent stimuli were scored as auditory correct only and not also as visual correct.

Figures 4A–B show that congruent stimuli were mostly identified correctly. In general, the stimuli with acoustic /bA/ were more susceptible to visual influences than those with acoustic /lA/. The overall $A_{bA}V$ response category proportions were auditory correct .25, visual correct .19, combination .01, and fusion .55. The $A_{lA}V$ response category proportions were auditory correct .57, visual correct .03, combination .34, and fusion .06. Notably, the fusion response was not the majority response across the entire response set. Among the combination responses, 59% were the combination of the presented auditory and visual consonants, 40% were the combination of the auditory consonant and a consonant that was not in the visual token, and 1% were the combination of a consonant that was not in the auditory token and a consonant that was in the visual consonant.

Analyses were carried out to determine whether the experimental design could be reduced in complexity by pooling data. Eight repeated measures analyses of variance were carried out separately for $A_{bA}V$ and $A_{lA}V$, and each response category (i.e., auditory correct, visual correct, combination, and fusion). Arcsine-transformed data were used to stabilize variances. The within-subjects factors were video consonant (8), talker (4), AV pairing (2), and alignment method (3); and the between-subjects factor was blocking (blocked versus mixed stimulus presentation). The *F*-tests for the main and between-subjects effects are listed in Table 1. Combination is not listed in Table 1 for $A_{bA}V$, as only about 1% of the responses were combinations. Visual correct (3% of responses) and fusion (6% of responses) are not listed in Table 1 for $A_{lA}V$ for the same reason.

The main effects that were reliable were the visual stimulus phoneme identity, the individual talker, and token pairing. The alignment and blocking main effects, and most (24 out of 25) of the two-way interactions involving them were not reliable (Bonferroni corrected $p > .05$), thus, permitting pooling across blocking and alignment. These results are consistent with the literature, which reports that relatively large onset asynchronies (±267 ms) do not disrupt AV speech perception (Massaro, et al., 1996; Munhall, et al., 1996). The absence of a reliable blocking effect is also consistent with previous findings that AV speech effects persist across diverse stimulus manipulations (Green, et al., 1991; Massaro, 1987; Walker, Bruce, & O'Malley, 1995).

After pooling across alignment and blocking, repeated measures analyses of variance models with all main effects and interactions were carried out for response categories that had adequate numbers of responses. The within-subjects factors were video consonant (8), talker (4), and pairing (2), and again arcsine-transformed data were used. Table 2 shows that the higher-order interactions were reliable (Bonferroni corrected $p < .05$), suggesting that responses were highly stimulus-specific, involving the specific talker, video consonant, and acoustic syllable. The Video x Talker interactions were expected, because the four talkers differed in their visual intelligibility. The effect of pairing and its interactions with other factors were significant for all of the *F*-tests involving acoustic /lA/. This was attributable to the two acoustic /lA/ tokens of one female talker, which were quite different in terms of their consonant duration and amplitude. Overall, the interactions show that all of the remaining design factors had to be retained for Experiment 2, presenting a challenging data set for establishing psychophysical relationships between response categories and stimulus measures.

Summary—Congruent and incongruent AV speech stimuli were constructed by dubbing video syllables, whose consonants spanned different places and manners of articulation (i.e., /bA, dA, gA, vA, zA, lA, wA, ΔA/) onto two different acoustic syllables (i.e., /bA/ or /lA/), according to three different alignments (i.e., consonant-onset, vowel-onset, and minimum acoustic-to-phantom distance) and two different token pairings, for four different talkers. Two designs were used, audio blocked or audio mixed. Analyses showed that the audio blocking (blocked versus mixed) and alignment factors were not reliable. But the rate at which auditory correct, visual correct, fusion, and combination responses were produced varied reliably with the acoustic token, the video token, the talker, and the AV pairing. The $A_{bA}V$ stimuli elicited more fusion responses but fewer combination responses, compared to $A_{lA}V$. The heterogeneity of the responses was such that the possibility of establishing reliable associations with physical stimulus measures could be considered a genuine challenge.

Experiment 2

The purpose of Experiment 2 was to determine whether quantified relationships between the auditory and visual components of AV speech stimuli could account for the corresponding relative frequencies of occurrence of the four response categories (i.e., fusion, combination, auditory correct, visual correct) that were in Experiment 1. A frequently cited observation about the relationship between acoustic and video speech stimuli is that the two are perceptually complementary (Binnie, Montgomery, & Jackson, 1974; Breeuwer & Plomp, 1985; Grant & Braida, 1991), possibly implying that correlations between physical acoustic and optical signals should be small. In fact, signal properties of AV speech afford high levels of correlation for congruent AV stimuli (Craig, et al., 2008; Jiang, et al., 2002; Yamamoto, et al., 1998; Yehia, et al., 1998).

Quantitative methods used to demonstrate correlation have included least-squares linear estimation (Kailath, Sayed, & Hassibi, 2000), Hidden Markov Models (Rabiner, 1989), and mutual information models (Nock, Iyengar, & Neti, 2002). For Experiment 2, we applied four types of correlation measures to each stimulus based on the previous work of Jiang et al. (2002), in which acoustic features were shown to predict optical signals and *vice versa*. Two mutual information measures were also generated for each stimulus. The mutual information measures were derived from the same general domain as the probability model discussed by Massaro (1987, 1999). Two additional measures were generated that used only acoustic signals. These were referred to as *phantom* measures. One was the minimum distance used in Experiment 1 to align stimuli, and the other was a measure of the relative durations of the presented acoustic versus the acoustic signal recorded with the video. The

idea behind using these measures was that the fairly tight correlation between congruent auditory and visual components implies that the acoustic phantom signal can stand in as a proxy for the stimulus that is actually presented (i.e., the video signal). Then the relationship between the audio and video can be computed using the same acoustic features.

All of the stimulus measures required defining acoustic and optical signal features. Research on acoustic phonetics carried out over more than six decades has resulted in many well-established acoustic phonetic quantities that are associated directly with auditory speech perceptual effects (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Nearey, 1997; Stevens, 1998), as well as, with physiological effects obtained with event-related potentials (ERPs) (e.g., Callan, Callan, Honda, & Masaki, 2000; Hosokawa, et al., 2002; Papanicolaou, et al., 2003; Simos, et al., 1998; Trébuchon-Da Fonseca, Giraud, Badier, Chauvel, & Liégeois-Chauvel, 2005), and with the blood-oxygen-level-dependent (BOLD) signal obtained with functional magnetic resonance imaging (fMRI) (e.g., Blumstein, Myers, & Rissman, 2005; Hutchison, Blumstein, & Myers, 2008). Here, acoustic line spectral pairs (LSPs) (Craig, et al., 2008; Jiang, et al., 2002; Sugamura & Itakura, 1986; Yehia, et al., 1998) were used to represent the auditory stimuli. The LSPs represent the vocal tract resonances, which are perceptually relevant.

A compact, highly researched, perceptually relevant representation of visual speech analogous to the LSPs has not, however, been established (Jiang, et al., 2007; Munhall & Vatikiotis-Bateson, 1998). Visual speech stimulus characterization has frequently been limited to the type of recording equipment that was used, the gender of the talker and his/her native language, phoneme identity, and such global characteristics, as stimulus token duration (Bernstein, in press-b; Massaro, et al., 1996; Rouger, Fraysse, Deguine, & Barone, 2008; Sekiyama, 1997). A few studies have been carried out on the frequency components of visible speech (e.g., C. S. Campbell & Massaro, 1997; Munhall, Kroos, Jozan, & Vatikiotis-Bateson, 2004). But an optical/visual phonetic description of speech, comparable to that of acoustic/auditory phonetics is not to our knowledge available (Bernstein, in press-b).

Jiang et al. (2007) showed that the perceptual dissimilarities among visual speech stimuli can be significantly accounted for by using the perceptually weighted dissimilarities between three-dimensional motions of points on the talking face. Here, the visual stimuli were represented by the three-dimensional motions of retro-reflectors glued to the talkers' faces (see Figure 1). The recording system does not require image processing as would be needed with two-dimensional video images.

Conceptual descriptions of the measures

Four measures were computed using least-squares linear estimation (Craig, et al., 2008; Jiang, et al., 2002). They are referred to as *correspondence* measures. Two used only acoustic and optical data, and two used additional independent and analogous measures that incorporated mid-sagittal magnetometry of the tongue. The magnetometry data were the two-dimensional motions of pellets glued to mid-sagittal tongue locations. The magnetometer data were included to partially represent visible tongue movement when the mouth is open.

Conceptually, the correspondence measures are computed first by transforming one type of measure (e.g., acoustic) to the other type of measure (e.g., optical) when the stimuli are congruent, and then obtaining a correlation between relative trajectories of the transformed signal (e.g., optical) and of those of the signal that was actually presented (e.g., optical). In this study, the correlations were computed for both matched (congruent) and mismatched (incongruent) stimuli, and both types always involved dubbing across recorded tokens.

Correspondence for acoustic-to-optical and optical-to-acoustic measures were computed as well as for the similar pair of correspondence measures that incorporated the magnetometry data set for each specific talker.

Two other measures were based on the acoustic signals that were actually presented in the perceptual experiment (Experiment 1) and the acoustic signals, which were never presented but had been recorded with the presented video, that is, the *phantom* acoustic signals. The first measure, phantom-to-acoustic duration ratio, was the duration discrepancy between the acoustic stimulus that was presented and the one implied by the video stimulus. The second measure was the minimum distance from Experiment 1.

Mutual information between acoustic and optical speech signals was the third type of measure. Conceptually, mutual information is a measure of shared data structure information. Mutual information was computed for the acoustic and optical signals, and for the acoustic, optical, and mid-sagittal tongue magnetometry signals.

The analyses in Experiment 2 included factor analysis (Kim & Mueller, 1978) to investigate the structure of the data obtained across the eight physical stimulus measures. Correlation and regression analyses were used to investigate directly the reliability and usefulness of the physical measures to account for the perceptual category response proportions.

Method

Signals: High-quality acoustic DAT recordings obtained simultaneously with the video recordings were down-sampled to 14.7 kHz and were then divided into frames of 24 ms, at a frame rate of 120 Hz, for the computation of 16^{th} order LSPs (Sugamura & Itakura, 1986). Although a frame window of 100 ms is preferred for sentences (Craig, et al., 2008), 24-ms frames were used here due to stimulus brevity. (For simplicity, the temporal derivatives of the LSPs were not used.) Three-dimensional optical data were analyzed after head motion compensation, using the two retro-reflectors on the eyebrows and one retro-reflector on the nose ridge (see Jiang, et al., 2002). Compensation removed head motion from the speech motion measurements.

Previous studies showed that tongue motion was correlated with face motion and speech acoustics (Jiang, et al., 2002; Yehia, et al., 1998). Deaf lipreaders report informally that they use glimpses of the tongue inside the mouth. Here, tongue motion was incorporated into three of the eight speech measures. The syllables had been recorded using audio, video, and three-dimensional optical recording systems, and also, independently by the same talkers, using three pellets on the tongue to record mid-sagittal motion magnetometry (Carstens system) (Jiang, et al., 2002). The video recorded simultaneously with the magnetometer signals was not useable for stimulus presentation, because the magnetometry system was in the images (Jiang, et al., 2002). However, the magnetometry data were co-registered with the simultaneously recorded three-dimensional optical data and scaled, and the three data streams (tongue motion, face motion, and acoustics) were synchronized and were processed to have the same feature frame rate (120 frame/second) (see Bernstein, et al., 2000; Jiang, et al., 2002). In the independent magnetometry data set, there were four tokens per syllable.

Experiment 1 showed that the AV stimulus alignment was not a significant factor, so the measurements for Experiment 2 were made only on the consonant-onset-aligned AV tokens. The main consideration was that mismatched AV speech signals are typically aligned on consonant onsets (Grant, et al., 2004; Massaro, et al., 1996; Munhall, et al., 1996), and that correspondence measures are well-defined with the consonant-onset based alignment (Craig, et al., 2008; Jiang, et al., 2002; Sugamura & Itakura, 1986; Yehia, et al., 1998). The signal segmentations from Experiment 1 were used. The initiation point for signal analyses was set

30 ms prior to the manually located acoustic consonant onsets, and analyses were applied for a 280-ms window equivalent to 34 optical frames, at a frame rate of 120 Hz (see Jiang, et al., 2007). Similarly, the independent data set with tongue motion was labeled and segmented for each token. The magnetometry and optical data were first combined to have 57 channels (17 × 3 optical + 3 × 2 magnetometry). Then the segmented four tokens for each syllable type were concatenated to have 136 frames (34 × 4).

Phantom measures: Phantom log duration ratio (*PADR*), log(*Pdur/Adur*), used the duration of the presented acoustic consonant (*Adur*), measured from the consonant onset to the vowel onset, and the duration of the phantom consonant, the acoustic signal (*Pdur*) that was recorded with the presented video. The minimum acoustic-to-phantom (*minAP*) distances from Experiment 1 were also used. Given an acoustic Consonant$_1$-/A/ and a video Consonant$_2$-/A/, the distance between the acoustic signal and the phantom acoustic signal was computed as:

$$d(T_{\mathrm{d}})=\left\| LSP_{\mathrm{A}}^{S_1,L_{\mathrm{A}}} - LSP_{\mathrm{pA}}^{S_2+T_{\mathrm{d}},L_{\mathrm{pA}}} \right\|_2, \tag{1}$$

where LSP_{A} and LSP_{pA} were from the acoustic Consonant$_1$-/A/ and the phantom acoustic Consonant$_2$-/A/, respectively; S_1 and S_2 represent the consonant onset points for the acoustic Consonant$_1$-/A/ and phantom acoustic Consonant$_2$-/A/, respectively; L_{A} approximated the consonant duration in the acoustic Consonant$_1$-/A/; and T_{d} represents temporal shifting of the acoustic Consonant$_1$-/A/ across the phantom acoustic Consonant$_2$-/A/. The minimum distance represents the minimum spectral discrepancy between the stimulus that was presented and the one recorded with the video stimulus.

Correspondence measures: The following correspondence measures were computed: LSPs to optical recordings (*LSP2O*); optical recordings to LSPs (*O2LSP*); LSPs to optical and magnetometer recordings (*LSP2OM*); and optical and magnetometer recordings to LSPs (*OM2LSP*). The acoustic /l$_1$A/ and video /v$_2$A/ are used here for illustration of how these measures were computed. The computation for *LSP2O* is described in detail. The other measures were carried out in analogous fashion. The method follows from Jiang et al. (2002).

Let $\boldsymbol{O}_{\mathrm{v2A}}$ and $\boldsymbol{LSP}_{\mathrm{l1A}}$ represent the three-dimensional optical motion measures for /v$_2$A/ and the acoustic LSP measures for /l$_1$A/, respectively. First $\boldsymbol{LSP}_{\mathrm{l1A}}$ is transformed into the optical domain using the transformation matrix $\boldsymbol{W}_{\mathrm{A2V}}$. Here, the acoustic /l$_1$A/ to optical /l$_1$A/ transformation was used. $\boldsymbol{W}_{\mathrm{A2V}}$ was computed using least-squares linear estimation (Kailath, et al., 2000), which in this case is,

$$\boldsymbol{W}_{\mathrm{A2V}}=\operatorname{argmin}\{\|\boldsymbol{O}_{\mathrm{l1a}} - \boldsymbol{W}_{\mathrm{A2V}} \cdot \boldsymbol{LSP}_{\mathrm{l1a}} - \boldsymbol{c}' \cdot \boldsymbol{1}\|\}_2, \tag{2}$$

where $\boldsymbol{c}$ is a constant vector and $\boldsymbol{1}$ is vector with all ones. Then a Pearson correlation coefficient between the transformed $\boldsymbol{W}_{\mathrm{A2V}}\cdot\boldsymbol{LSP}_{\mathrm{l1A}}$ and the measured optical data $\boldsymbol{O}_{\mathrm{v2A}}$ was computed. Those correlations for each stimulus and each type of transformation comprised the correspondence measures. That is, there were 128 correlations (corresponding to the 128 consonant-onset aligned stimuli) for each of the *LSP2O*, *O2LSP*, *LSP2OM*, and *OM2LSP* measures.

Mutual information measures: Mutual information between signals is the extent to which there is shared versus independent data structure information. The mutual information

measure was modified from Nock et al. (2002). Feature vectors derived from the physical signals were considered to be samples from a multivariate Gaussian probability distribution. Before computing mutual information, acoustic, optical, and magnetometry measurements were normalized to have zero mean and unit standard deviation.

For example, the mutual information of optical and acoustic LSP signals, $I(\boldsymbol{O;LSP})$, was empirically computed for each AV stimulus as:

$$I(\boldsymbol{LSP};\boldsymbol{O})=\frac{1}{2}\log\left(\frac{\det(\Sigma_{\mathrm{LSP}})\cdot\det(\Sigma_{\mathrm{O}})}{\det(\Sigma_{\mathrm{LSP,O}})}\right), \tag{3}$$

where Σ_{LSP}, Σ_{O}, and $\Sigma_{\mathrm{LSP,O}}$ denote empirical estimates of covariance matrices for acoustic, optical, and joint AV signals distributions, respectively, and det(·) represents the calculation of the determinant of a square matrix. The determinant of the square matrix Σ_{LSP} can be computed as:

$$\det(\Sigma_{\mathrm{LSP}})=\prod_{i=1}^{N_{\mathrm{LSP}}}\lambda_{\mathrm{LSP}}^{j}, \tag{4}$$

where λ_{LSP}'s represents the eigenvalues with descending amplitudes of the square matrix Σ_{LSP}, and N_{LSP} represents the number of eigenvalues. Therefore, Equation-3 can be re-written as:

$$I(\boldsymbol{LSP};\boldsymbol{O})\approx\frac{1}{2}\left(\sum_{i=1}^{8}\log\left(\lambda_{\mathrm{LSP}}^{i}\right)+\sum_{j=1}^{12}\log\left(\lambda_{\mathrm{O}}^{j}\right)-\sum_{k=1}^{12}\log\left(\lambda_{\mathrm{LSP,O}}^{k}\right)\right). \tag{5}$$

As an approximation, only the first 8 (N_{LSP}), 12 (N_{O}), and 12 ($N_{\mathrm{LSP,O}}$) eigenvalues were retained for Σ_{LSP}, Σ_{O}, and $\Sigma_{\mathrm{LSP,O}}$, respectively. The more joint structure in the covariance between the acoustic and optical signals, the larger the mutual information.

Measures incorporating magnetometry data: For each stimulus token, the following measures were computed with concatenated magnetometry data (*M* for magnetometry): optical to acoustic correspondence (*OM2LSP*), acoustic to optical correspondence (*LSP2OM*), and mutual information (*muinfoM*).

Perceptual measures: Experiment 1 showed that the AV stimulus alignment and blocking were not significant factors. Therefore, response data in Experiment 1 were pooled across the three alignments and ten participants. The pooling resulted in 128 response proportions (8 video x 2 acoustic x 2 pairings x 4 talkers) per category, corresponding to the 128 consonant-onset aligned stimuli.

Factor analysis: Underlying (latent) physical AV stimulus dimensions were sought using exploratory factor analysis based on principal components analysis (PCA) (Kim & Mueller, 1978). No *a priori* relationship was imposed on the measures. The analysis was carried out using 128 consonant-onset aligned AV tokens (8 video x 2 acoustic x 2 pairings x 4 talkers), for which there were eight measurements each. This ratio of tokens to measurements is adequate for factor analysis. Underlying factors were extracted and rotated (using Varimax with Kaiser normalization) to obtain a parsimonious result without loss of information.

Pearson's correlation and linear regression analysis: These analyses related all of the perceptual category scores with all of the valid physical measures from all of the auditory and visual stimuli. The aim was to account for response category effects without using the perceptual identities of the auditory syllables. Therefore, analyses were not performed separately for $A_{bA}V$ and $A_{lA}V$ stimuli.

Pearson product-moment coefficients (Pearson's correlation) were computed between the physical measures and the proportion response category scores for each of the four categories of responses (auditory correct, visual correct, combination, and fusion). The correlations from the across-talker data were Bonferroni corrected. The correlations from the per-talker data were also Bonferroni corrected. However, the latter correction depended on the number of talkers that produced significant correlations for each physical measure and response category pair. The thresholds for significant correlations were set at .59 for 1 talker, .42 for 2 talkers, .32 for 3 talkers, and .23 for 4 talkers.

The same stimulus-response relationships were used to compute stepwise linear regression fits between perceptual and physical measures. Regression analysis was applied to the overall data set, not to the data of individual talkers. Only physical measures involving significant correlations with the response category scores were entered as independent variables in the regression analyses. The order in which these physical measures were entered was based on the magnitude of their correlations (from highest to lowest). The overall variances accounted for and the variances accounted for from each significant physical measure were computed. The computed regression line was overlaid on perceptual-physical scatter plots.

Results and Discussion

Physical measures: Factor analysis was carried out in order to gain understanding of the relationships among the physical measures. Three factors were obtained, accounting for 75% of the total variance in the physical data. Following removal of loadings below .40, all of the rotated component factor loadings were greater than .73 (Table 3). In order to gain insight into the factors, scatter plots were generated using components on latent dimensions versus the physical measures that were loaded on the factor (see Figure 5). Notably, the three factors loaded differentially on the measures that were theoretically and computationally different. The scatter plots were used to assign labels to the factors: Factor 1 was labeled *correspondence*; Factor 2 was labeled *phantom relationships*; and Factor 3 was labeled *mutual information*. The factor analysis confirms that the measures comprising each factor provide independent information.

Figure 5A–D shows a distinct cluster of stimuli high on correspondence and on the physical measure. Examination of this cluster revealed that it comprises congruent stimuli. Because this cluster could overly influence correlations between perceptual response categories and physical measures, analyses reported below were carried out with and without the congruent stimuli. Fortunately, results were similar across analyses, because the goal here was to account for responses to both matched and mismatched stimuli within the same model.

Figures 5A and 5C show also that LSP to optical (*LSP2O*) and LSP to optical and magnetometry (*LSP2OM*) measures were more effective in producing a continuum of values than were the transformations in the opposite direction (i.e., *O2LSP* and *OM2LSP*). This is because the optical measures are intrinsically less informative than the LSP measures (Jiang, et al., 2002). Also, variance is generally lower for correspondence measures that used the magnetometry data, perhaps, because the measures used concatenation of four token segments.

The two measures, *PADR* and *minAP*, that loaded on the phantom relationships factor used the phantom acoustic signal, that is, the one that had been recorded with the presented video but was not presented. Both measures produced a continuum of values with congruent stimuli in the middle of the range (see Figure 5E–F). Stimuli with acoustic /bA/ were high on the phantom relationships factor, and those with acoustic /lA/ were low. This is because the /bA/ acoustic stimulus was generally shorter in duration than the acoustic stimulus associated with the presented video, and *vice versa* for the /lA/ acoustic stimulus. The phantom relationships dimension simultaneously captures spectral and temporal differences across consonants.

The two information measures, *muinfo* and *muinforM*, loaded on the mutual information factor (see Figure 5G–H). Interestingly, this factor and the two measures resulted in a more uniform spread of the stimuli across dimensions than the other measures: This occurred, because the acoustic /bA/ and /lA/ stimuli paired with each video stimulus were located approximately in the same stimulus space, rather than segregating into different parts of the space; as a result, the visual stimuli determined the distribution through the space. That is, the optical data largely account for the distribution of the mutual information measures, and this latent factor appears to provide little information about the relationship across acoustic and video stimulus components. This result is interesting in light of the results further below, showing that the mutual information measures did not account at all for the perceptual response distributions.

In summary, the physical measures were shown to comprise a three-dimensional space, whose dimensions corresponded with the measures that were theoretically and computationally grouped together. The correspondence and phantom relationships dimensions distributed the stimuli in terms of the relationships between the auditory and visual stimulus components. The mutual information dimension distributed the stimuli in terms of the visual stimulus component, and so, could be considered a different type of dimension than the other two.

Correlation and regression analyses with physical and perceptual measures:
Preliminary results not presented here showed that the latent dimensions from the factor analysis are less effective in accounting for response category proportions than the raw physical measures. Therefore, only the raw physical measures were used for the correlation and regression analyses presented below. Tables 4–6 show the results. Tables 4 and 5 display bivariate correlations and multiple R^2 statistics for the physical stimulus measures versus the response category proportions obtained in Experiment 1. Table 4 shows the results without congruent stimuli, and Table 5 shows the results with congruent stimuli. Bonferroni-corrected significant correlations are indicated with bolding. The variance accounted for, R^2, across measures for each response category were obtained with stepwise entry of factors in terms of magnitude of the bivariate correlations. Table 6 shows the bivariate correlations on a per-talker basis, using all the stimuli, 32 per talker. It also shows in the right part of the table the pattern of significant correlations. As explained in the methods, the significance of the individual correlations was adjusted in relationship to the number of talkers with suprathreshold correlations for each physical measure and response category pair. The signs in the summary section of Table 6 indicate the direction of the correlations, and the numbers in parentheses indicate the number of times the correlation was significant across talkers.

The goal here was to account for response category proportions for all of the stimuli, without using knowledge of whether the stimuli were matched, and without knowing their specific phoneme identities. A unified model of AV speech psychophysics should be able to account for AV stimuli from matched to highly mismatched. Therefore, in what follows, we

focus on the correlation and regression results that included the congruent stimuli (Tables 5 and 6). 1

In Table 5, three of the four correspondence measures (i.e., *LSP2O*, *LSP2OM*, and *OM2LSP*) were negatively correlated with fusion response proportions. The correspondence measures were Pearson correlation coefficients computed between a stimulus component (e.g., auditory) that was transformed via a weights matrix into the other component (e.g., optical) and the optical data for the presented stimulus. The lower the correlation between the transformed and the presented stimulus the more fusion responses were given. Correspondence measures were not reliably correlated with combination response proportions. But all of the correspondence measures were positively correlated with auditory correct response proportions. So when the presented stimulus and the transformed stimulus were similar, auditory correct proportions increased. *LSP2O* correlated negatively with visual correct responses. That is, the lower the correlation between LSPs transformed to optical data and the optical data from the presented stimulus the more visual correct responses were obtained.

The duration difference measure, *PADR*, was significantly correlated with the four perceptual category responses, positively with fusion and visual correct, and negatively with combination and auditory correct. So, as the phantom became longer, relative to the presented audio, fusion and visual correct responses increased. But as it became shorter relative to the presented audio, combination and auditory correct responses increased. The minimum spectral distance, *minAP*, was negatively correlated with the fusion responses. As spectral difference increased between the presented and the phantom audio, fusion responses declined. Mutual information measures were not reliably correlated with any of the four perceptual category responses. The scatter plots for perceptual category scores versus the corresponding significant physical measures were shown for all talkers (Figure 6).

Table 5 shows that the response category variance accounted for, R^2, by the physical measures was .52 for fusion, .20 for combination, .41 for auditory correct, and .17 for visual correct. Experiment 1 showed that there were significant differences between talkers, so being able to show significant results across talkers and all stimuli suggests that the measures generalize. However, pooled data can smooth out important variation. Table 6 shows individual talker correlations and a summary of the correlations that were significant across talkers. There were differences between the results in Tables 5 and 6.

Two correlations were reliable only in the data pooled across talkers. Those were *OM2LSP* with fusion and *LSP2O* with visual correct. The thresholds for significant correlations were .59 for 1 talker, .42 for 2 talkers, .32 for 3 talkers, and .23 for 4 talkers. *OM2LSP* trended towards significant across all talkers, but failed all of the threshold tests. The *LSP2O* also failed the tests for significance, although the trend matched the correlation obtained with pooled data. This measure might have been reduced in its effect, because the visual correct response was relatively infrequent, with the overall $A_{bA}V$ visual correct response proportion being .19, and the overall $A_{lA}V$ visual correct response proportion being .03. Although the number of stimuli entered in each correlation in Table 6 was the same ($N = 32$), the number of responses per category for each stimulus varied, which could account for the reduction in significant physical measures when the data were broken down by talker.

Correlations between combination and *minAP* were significant in individual talker results but not across talkers. This could be due to individual talker differences in speech

[1]Were a larger stimulus set practical, we would have included additional phonemes for the video stimuli. The clustering of congruent stimuli in Figure 5 was considered to be due to the relatively sparse sampling of the AV stimulus space.

production tempo as well as vocal tract resonances, such that this measure needs to be normalized across talkers.

Summarizing the pattern of results, fusion responses increased with low correspondence, with increased phantom audio length, and with a smaller minimum distance between the phantom audio and the presented audio. Combination responses increased as the phantom became shorter relative to the presented audio, and in the individual results, as the minimum distance increased. Auditory correct responses were associated with increased correspondence and a decrease in the phantom duration relative to the presented audio. Visual correct responses were associated with an increase in the duration of the phantom and with a decrease in correspondence. Mutual information was not important for accounting for the four perceptual category responses.

Summary: Eight AV speech measures were computed on a large set of stimuli. A factor analysis was applied to the physical measures to investigate the physical dimensions of the stimulus data. Then, the original physical measures were correlated with the perceptual response category scores. Regression analyses that were carried out across all of the stimuli to predict response category proportions showed that the physical measures accounted for significant proportions of the variance in perceptual response proportions for the individual stimuli. The largest R^2 statistics involved fusion and auditory correct responses. Different patterns of significant correlations were obtained across the four perceptual response categories.

General Discussion

Open-set identification responses were obtained in Experiment 1 in response to matched and mismatched AV nonsense syllables. Although only two auditory syllables and eight video syllables were combined, identification responses used most of the English consonants. In addition, consonant response patterns varied across talkers. When the individual responses were pooled in terms of response categories (i.e., fusion, combination, auditory correct, and visual correct), their proportions varied as a function of the acoustic consonant, the video consonant, and the talker. This variability seemed to us to pose a real challenge for setting up a psychophysical relationship between stimuli and physical stimulus measures. It also readily explains why experiments that use the McGurk/fusion effect might, in the absence of a psychophysical account, require extensive pretesting of stimuli in order to obtain specific experimental effects.

There were two patterns that held strongly across talkers in Experiment 1. They were (1) responses to $A_{bA}V$ were influenced more by the video component of the stimulus than were responses to $A_{lA}V$; and (2) fusion responses were most frequently obtained with $A_{bA}V$, and combination responses were most frequently obtained with $A_{lA}V$. The overall $A_{bA}V$ response category proportions were auditory correct .25, visual correct .19, combination .01, and fusion .55. The $A_{lA}V$ response category proportions were auditory correct .57, visual correct .03, combination .34, and fusion .06. These proportions show that very frequently, perceivers report one or both of the syllables correctly. Although participants in this experiment were monitored for attention to the video screen, future studies with eyetracking are needed to determine whether gaze is systematically related to reports that include correct identifications.

The goal for Experiment 2 was to account for the response category proportions in terms of physical stimulus measures. We hypothesized that speech perceivers have learned the normal relationships between acoustic and video speech syllables. When the normal relationships are violated, we expect that perceptual processing results in systematic effects

(Bernstein, et al., 2008). Because we hypothesized that perceptual effects were due to AV stimulus relationships, all of the measures that were applied in Experiment 2 quantified relationships between auditory and visual stimulus components. Eight physical measures were computed on each stimulus. Factor analysis showed that three significant latent factors could account for the structure of the stimulus measurements, accounting for 75% of their total variance. The three factors loaded differentially on the physical measures that were theoretically and computationally different, and the factors were labeled *correspondence*, *phantom relationships*, and *mutual information*. Thus, this particular set of measures appeared potentially to cover a perceptually valid stimulus space. Results not reported here comparing factor scores with physical measures showed, however, that the raw physical measures were more powerful than the factors in accounting for response category variance; so the raw measures were used in subsequent correlation and regression analyses. Mutual information, however, was found to be not at all useful in accounting for variance in the perceptual measures, demonstrating that measures that capture stimulus structure are not necessarily useful in accounting for perceptual responses.

Most of the correlations between the stimulus measures and the response category proportions for individual stimuli generalized across the talkers, the video consonants, and the acoustic syllables. Analyses were carried out with and without the matched stimuli. We prefer to include the former, because measures should be able to account for stimuli that vary in terms of incongruity, from matched to highly incongruent.

Regression analyses showed that across talkers, 17% to 52% of the variance in response category proportions was accounted for by the physical measures. The largest variance accounted for, 52%, was for the fusion (i.e., McGurk-type) responses. Auditory correct variance accounted for was 41%. Combination and visual correct were, respectively, 20% and 17%. The systematic relationships observed here between the physical stimulus measures and the response categories are consistent with a processing system that uses auditory and visual stimulus component relationships in generating percepts.

Psychophysical measures

Tables 5 and 6 show that within the correspondence and phantom measures, some of the physical measures were more effective than others. Correspondence measures that included magnetometry data were more effective in explaining response variance than those without. Because the magnetometry data were used as a proxy for measures of visible tongue movement, their efficacy supports the expectation that perceivers use visible tongue movement. We would expect that other types of measures of visible tongue movement could be used to improve perceptual response predictions.

Tables 5 and 6 show negative correlations between correspondence measures and fusion response category proportions and positive correlations between correspondence measures and auditory correct response category proportions. The correspondence measure is related to global formant patterns and articulatory movements. Fusion appears to result from more extreme global differences between the acoustic and video stimulus components. The correspondence measures are referred to here as *global* measures, because their values reflect a relatively long portion of the syllable.

A somewhat puzzling result is the relationship between correspondence measures and the *minAP* measures. They were both negatively correlated with fusion response proportions. Why are low values on both measures associated with increased fusion responses? This result can be explained by the fact that although the correspondence measures and the *minAP* measure corresponded to similarities and dissimilarities, respectively, the former focused on the relative temporal dynamics, while the latter focused on the absolute spectra,

which could be similar for a very brief duration. That is, low correspondence measures mean that the relative formant or articulatory trajectories of the presented and the transformed stimulus (in the same modality) are different; and low *minAP* values mean that absolute spectra of the presented audio and the phantom audio are highly similar but primarily for only a brief duration. We investigated whether *minAP* could be excluded from the regression equations. Variance accounted for was diminished when *minAP* was not used. These measures, thus, both appear to be capturing some aspects of how the stimuli are related perceptually to each other. These results suggest that both global and local relationships contribute to the fusion response. The same relationships contribute to auditory correct responses, but in their case, correspondence is high and *minAP* is low.

PADR measures the relative duration of the consonants presented versus the phantom recorded with the presented video. It was the most useful measure in Tables 5 and 6. Why this measure is so successful is a question. One possibility is that duration differences are sensitive to phonetic manner differences across AV stimulus components. In general, continuous consonants such as /v, z, l, w, ∆/ are longer than the stop consonants /b, d, g/. Duration differences also imply local spectral differences, so *PADR* is consistent in being grouped with *minAP* in the factor analysis. *PADR* is likely a proxy for more specific measures of internal stimulus dynamics and spectra.

Implications

Research on AV speech perception has focused intensively on the question of whether the internal mental representation of AV speech is amodal or distributed across modality-specific representations (for reviews see,Bernstein, et al., 2004; R. Campbell, 2008; Rosenblum, 2008). This question can be viewed as one properly addressed with evidence from neural measures that can indicate the participation of brain areas responsible for processing auditory and visual input and for integrating information (e.g., Bernstein, et al., 2008; Hasson, Skipper, Nusbaum, & Small, 2007; Miller & d'Esposito, 2005; Ojanen, et al., 2005; Pekkola, et al., 2005; Ponton, Bernstein, & Auer, 2009; Sams, et al., 1991; Skipper, van Wassenhove, Nusbaum, & Small, 2007). But within the neuroimaging literature there is ongoing debate about whether the interaction of auditory and visual speech stimuli is due to convergence onto a neuronal substrate somewhere in the brain or due to connections across modality-specific representations (Bernstein, in press-a; Bernstein, et al., 2004).

The perceptual results here show that although AV speech percepts are often compellingly unified, sometimes they are not. The combination responses show that both auditory and visual components can be perceived. Auditory correct and visual correct responses also suggest that perception involves independent component processing. But these effects are open to interpretation or question, as participants, for example, could have failed to attend to one or the other component. The results of Experiment 2 do, however, speak to the issue of whether perceptual representations are amodal. The finding that AV stimulus *relationship* can account significantly for response types does support the possibility that internal representations are not amodal, and that perception is based on modality-specific stimulus relationships. If amodal representations are the output of perceptual processing, then they must somehow preserve stimulus relationship information in order to be consistent with results here.

The results of a study by Rosenblum, Miller, and Sanchez (2007), in which previous lipreading of a talker conferred a small but statistically significant benefit later to perceiving the same talker's speech in noise was interpreted by the authors as evidence for an amodal representation of idiolectal information. In the current study, quantitative measures from one talker were not applied to speech of another talker, and indeed, we showed that individual tokens and talkers were significantly different in producing different distributions of the

perceptual response categories. How might the results here speak to the Rosenblum-et-al. results? The efficacy of the *PADR* and *minAP* measures suggests that the presented video does activate auditory representations that are congruent with the talker's visible speech. However, the auditory representation could be due to a distributed network that connects modality-specific representations. Over a lifetime of perceiving AV speech stimuli, perception of visible speech might well activate auditory imagery, particularly, in good lipreaders (lipreading ability in Rosenblum-et-al. was required to be what we judge to be moderate for hearing lipreaders). Lipreading followed by listening to the same talker's speech in noise, within the same test session, could potentially benefit from activation of auditory phonetic imagery or representations induced by visual speech input. Again, however, we suggest here that evidence about the neural representation of speech needs crucially to include direct neural measures.

The finding that correspondence measures were more effective when the acoustic was used to predict the video than *vice versa* could suggest that perceivers depend on the more reliable, more informative stimulus (Welch & Warren, 1986). On the other hand, predictions from acoustics to optical data are generally more accurate, because the former is more informative (Jiang, et al., 2002). Those measures also produce a wider spread in the measurements, leading to higher correlations.

In the current study, there were several limitations: (1) The number of subjects was relatively modest; (2) Analyses of stimulus-response relationships were not carried out on a per-subject basis; (3) A limited number of video nonsense CV syllables was used; and (4) Only auditory /bA/ and /lA/ tokens were used. Most importantly, the physical stimulus measures were based primarily on signal considerations. Alternative approaches could make use of measures related to hypotheses about representations at the level of neural processing (see for example, Chandrasekaran, et al., 2009).

Summary and conclusions

Quantitative characterization of the stimuli has been a major methodological tool for understanding perception and its neural bases. That tool has been mostly absent from research on AV speech perception and its neural bases. Summerfield (1987) posited that the core of a comprehensive account of AV speech integration would be a perceptually relevant metric of AV speech stimuli. The current study sought to establish psychophysical relationships between AV stimulus and response measures. The stimulus measures were all defined as relationships between acoustic and optical signals, and the response measures were the proportions with which different categories of response were given to each of the stimuli. The physical relationships were effective in accounting for the distributions of perceptual response categories across talkers and stimuli. The success of this study suggests that stimulus measures such as these can be useful in furthering understanding of AV speech perception and neural processing. For example, using stimuli calibrated to have a range of correspondence values, neural activations can be sought as a function of those calibrated stimuli (Bernstein, et al., 2008).

Acknowledgments

Jintao Jiang and Lynne E. Bernstein carried out much of this study while in the Division of Communication and Auditory Neuroscience, House Ear Institute, Los Angeles, California. They are now with the Speech and Hearing Sciences Department, George Washington University, Washington, DC. We acknowledge the assistance of Edward T. Auer, Brian Chaney, Laurel Fisher, and John Jordan. Some of the results were presented at a conference, Auditory-Visual Speech Processing 2005 (British Columbia, Canada, July 24–27, 2005). This research was supported in part by awards from the National Institutes of Health (DC006035 and DC008308) (Bernstein, PI). It was also supported by the National Science Foundation. (The views expressed here are those of the authors and do not necessarily represent those of the National Science Foundation.)

References

Bernstein, LE. Multisensory information integration for communication and speech. In: Stein, BE., editor. The New Handbook of Multisensory Integration. Cambridge, MA: MIT Press; (in press-a)

Bernstein, LE. Visual speech perception. In: Bailly, G.; Perrier, P.; Vatiokis-Bateson, E., editors. Audiovisual Speech Processing. Cambridge, England: Cambridge University Press; (in press-b)

Bernstein LE, Auer ET Jr. Chaney B, Alwan A, Keating PA. Development of a facility for simultaneous recordings of acoustic, optical (3-D motion and video), and physiological speech data. Journal of the Acoustical Society of America. 2000; 107(5):2887.

Bernstein, LE.; Auer, ET., Jr.; Moore, JK. Audiovisual speech binding: Convergence or association?. In: Calvert, GA.; Spence, C.; Stein, BE., editors. Handbook of Multisensory Processes. Cambridge, MA: MIT Press; 2004. p. 203-223.

Bernstein LE, Lu Z-L, Jiang J. Quantified acoustic-optical speech signal incongruity identifies cortical sites of audiovisual speech processing. Brain Research. 2008; 1242:172–184. [PubMed: 18495091]

Binnie CA, Montgomery AA, Jackson PL. Auditory and visual contributions to the perception of consonants. Journal of Speech & Hearing Research. 1974; 17(4):619–630. [PubMed: 4444283]

Blumstein SE, Myers EB, Rissman J. The perception of voice onset time: An fMRI investigation of phonetic category structure. Journal of Cognitive Neuroscience. 2005; 17(9):1353–1366. [PubMed: 16197689]

Braida LD. Crossmodal integration in the identification of consonant segments. Quarterly Journal of Experimental Psychology: Human Experimental Psychology. 1991; 43A(3):647–677.

Breeuwer M, Plomp R. Speechreading supplemented with formant-frequency information from voiced speech. Journal of the Acoustical Society of America. 1985; 77(1):314–317. [PubMed: 3973225]

Callan DE, Callan AM, Honda K, Masaki S. Single-sweep EEG analysis of neural processes underlying perception and production of vowels. Cognitive Brain Research. 2000; 10(1–2):173–176. [PubMed: 10978705]

Campbell CS, Massaro DW. Perception of visible speech: Influence of spatial quantization. Perception. 1997; 26(5):627–644. [PubMed: 9488886]

Campbell R. The processing of audio-visual speech: Empirical and neural bases. Philosophical Transactions of the Royal Society of London B: Biological Sciences. 2008; 363:1001–1010.

Chandrasekaran C, Trubanova A, Stillittano S, Caplier A, Ghazanfar AA. The natural statistics of audiovisual speech. PLoS Computational Biology. 2009; 5(7):e1000436. [PubMed: 19609344]

Craig MS, van Lieshout P, Wong W. A linear model of acoustic-to-facial mapping: Model parameters, data set size, and generalization across speakers. Journal of the Acoustical Society of America. 2008; 124(5):3183–3190. [PubMed: 19045802]

Fowler, C. Speech as a supramodal or amodal phenomenon. In: Calvert, G.; Spence, C.; Stein, BE., editors. Handbook of Multisensory Processes. Cambridge, MA: MIT Press; 2004. p. 189-202.

Grant KW, Braida LD. Evaluating the articulation index for auditory-visual input. Journal of the Acoustical Society of America. 1991; 89(6):2952–2960. [PubMed: 1918633]

Grant KW, Greenberg S, Poeppel D, van Wassenhove V. Effects of spectro-temporal asynchrony in auditory and auditory-visual speech processing. Seminars in Hearing. 2004; 25:241–255.

Green KP, Kuhl PK. The role of visual information in the processing of place and manner features in speech perception. Perception & Psychophysics. 1989; 45(1):34–42. [PubMed: 2913568]

Green KP, Kuhl PK, Meltzoff AN, Stevens EB. Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. Perception & Psychophysics. 1991; 50(6):524–536. [PubMed: 1780200]

Green KP, Norrix LW. Acoustic cues to place of articulation and the McGurk effect: The role of release bursts, aspiration, and formant transitions. Journal of Speech, Language, & Hearing Research. 1997; 40(3):646–665.

Hasson U, Skipper JI, Nusbaum HC, Small SL. Abstract coding of audiovisual speech: Beyond sensory representation. Neuron. 2007; 56(6):1116–1126. [PubMed: 18093531]

Hosokawa M, Okazaki S, Kawakubo Y, Maekawa H, Ozaki H. Topographic change of ERP due to discrimination of CV syllables with various vowel durations. International Congress Series. 2002; 1232:53–57.

Hutchison ER, Blumstein SE, Myers EB. An event-related fMRI investigation of voice-onset time discrimination. Neuroimage. 2008; 40(1):342–352. [PubMed: 18248740]

Jiang J, Alwan A, Keating PA, Auer ET Jr. Bernstein LE. On the relationship between face movements, tongue movements, and speech acoustics. EURASIP Journal on Applied Signal Processing. 2002; 2002(11):1174–1188.

Jiang J, Auer ET Jr. Alwan A, Keating PA, Bernstein LE. Similarity structure in visual speech perception and optical phonetic signals. Perception & Psychophysics. 2007; 69(7):1070–1083. [PubMed: 18038946]

Kailath, T.; Sayed, AH.; Hassibi, B. Linear Estimation. Upper Saddle River, NJ: Prentice Hall; 2000.

Kim, J-O.; Mueller, CW. Factor Analysis: Statistical Methods and Practical Issues. Beverly Hills, CA: Sage; 1978.

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M. Perception of the speech code. Psychological Review. 1967; 74(6):431–461. [PubMed: 4170865]

MacDonald J, McGurk H. Visual influences on speech perception processes. Perception & Psychophysics. 1978; 24(3):253–257. [PubMed: 704285]

Manuel SY, Repp BH, Studdert-Kennedy M, Liberman AM. Exploring the "McGurk Effect". Journal of the Acoustical Society of America. 1983; 74(S1):S66.

Massaro, DW. Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry. Hillsdale, NJ: Lawrence Erlbaum Associates; 1987.

Massaro, DW. Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. Cambridge, MA: MIT Press; 1998.

Massaro DW. Speechreading: Illusion or window into pattern recognition. Trends in Cognitive Sciences. 1999; 3(8):310–317. [PubMed: 10431185]

Massaro DW, Cohen MM, Smeele PMT. Perception of asynchronous and conflicting visual and auditory speech. Journal of the Acoustical Society of America. 1996; 100(3):1777–1786. [PubMed: 8817903]

McGrath M, Summerfield Q. Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. Journal of the Acoustical Society of America. 1985; 77(2):678–685. [PubMed: 3973239]

McGurk H, MacDonald J. Hearing lips and seeing voices. Nature. 1976; 264(5588):746–748. [PubMed: 1012311]

Miller LM, d'Esposito M. Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. Journal of Neuroscience. 2005; 25(25):5884–5893. [PubMed: 15976077]

Munhall KG, Gribble P, Sacco L, Ward M. Temporal constraints on the McGurk effect. Perception & Psychophysics. 1996; 58(3):351–362. [PubMed: 8935896]

Munhall KG, Kroos C, Jozan G, Vatikiotis-Bateson E. Spatial frequency requirements for audiovisual speech perception. Perception & Psychophysics. 2004; 66(4):574–583. [PubMed: 15311657]

Munhall, KG.; Vatikiotis-Bateson, E. The moving face during speech communication. In: Campbell, R.; Dodd, B.; Burnham, D., editors. Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech. East Sussex, UK: Psychology Press; 1998. p. 123-139.

Nearey TM. Speech perception as pattern recognition. Journal of the Acoustical Society of America. 1997; 101(6):3241–3254. [PubMed: 9193041]

Nock, HJ.; Iyengar, G.; Neti, C. Assessing face and speech consistency for monologue detection in video; Paper presented at the ACM International Conference on Multimedia; Juan-les-Pins, France. 2002.

Ojanen V, Möttönen R, Pekkola J, Jaaskelainen IP, Joensuu R, Autti T, et al. Processing of audiovisual speech in Broca's area. NeuroImage. 2005; 25(2):333–338. [PubMed: 15784412]

Papanicolaou AC, Castillo E, Breier JI, Davis RN, Simos PG, Diehl RL. Differential brain activation patterns during perception of voice and tone onset time series: A MEG study. Neuroimage. 2003; 18(2):448–459. [PubMed: 12595198]

Pekkola J, Ojanen V, Autti T, Jaaskelainen IP, Möttönen R, Tarkianen A, et al. Primary auditory cortex activation by visual speech: An fMRI study at 3T. NeuroReport. 2005; 16(2):125–128. [PubMed: 15671860]

Ponton CW, Bernstein LE, Auer ET Jr. Mismatch negativity with visual-only and audiovisual speech. Brain Topography. 2009; 21:207–215. [PubMed: 19404730]

Rabiner LR. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of The IEEE. 1989; 77(2):257–286.

Rosenblum LD. Speech perception as a multimodal phenomenon. Current Directions in Psychological Science. 2008; 17(6):405–409.

Rosenblum LD, Miller RM, Sanchez K. Lip-read me now, hear me better later: Cross-modal transfer of talker-familiarity effects. Psychological Science. 2007; 18(5):392–396. [PubMed: 17576277]

Rouger J, Fraysse B, Deguine O, Barone P. McGurk effects in cochlear-implanted deaf subjects. Brain Research. 2008; 1188:87–99. [PubMed: 18062941]

Sams M, Aulanko R, Hamalainen M, Hari R, Lounasmaa OV, Lu ST, et al. Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. Neuroscience Letters. 1991; 127(1):141–145. [PubMed: 1881611]

Sekiyama K. Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. Perception & Psychophysics. 1997; 59(1):73–80. [PubMed: 9038409]

Sekiyama K, Tohkura Y. McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. Journal of the Acoustical Society of America. 1991; 90(4):1797–1805. [PubMed: 1960275]

Simos PG, Diehl RL, Breier JI, Molis MR, Zouridakis G, Papanicolaou AC. MEG correlates of categorical perception of a voice onset time continuum in humans. Cognitive Brain Research. 1998; 7(2):215–219. [PubMed: 9774735]

Skipper JI, van Wassenhove V, Nusbaum HC, Small SL. Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. Cerebral Cortex. 2007; 17(10):2387–2399. [PubMed: 17218482]

Stevens, KN. Acoustic Phonetics. Cambridge, MA: MIT Press; 1998.

Sugamura N, Itakura F. Speech analysis and synthesis methods developed at ECL in NTT - from LPC to LSP. Speech Communication. 1986; 5(2):199–215.

Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America. 1954; 26(2):212–215.

Summerfield, Q. Some preliminaries to a comprehensive account of audio-visual speech perception. In: Dodd, B.; Campbell, R., editors. Hearing by Eye: The Psychology of Lip-reading. London: Erlbaum; 1987. p. 3-52.

Summerfield Q, McGrath M. Detection and resolution of audio-visual incompatibility in the perception of vowels. Quarterly Journal of Experimental Psychology A, Human Experimental Psychology. 1984; 36(1):51–74.

Trébuchon-Da Fonseca A, Giraud K, Badier J-M, Chauvel P, Liégeois-Chauvel C. Hemispheric lateralization of voice onset time (VOT) comparison between depth and scalp EEG recordings. NeuroImage. 2005; 27(1):1–14. [PubMed: 15896982]

van Wassenhove V, Grant KW, Poeppel D. Temporal window of integration in auditory-visual speech perception. Neuropsychologia. 2007; 45(3):598–607. [PubMed: 16530232]

Walker S, Bruce V, O'Malley C. Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. Perception & Psychophysics. 1995; 57(8):1124–1133. [PubMed: 8539088]

Welch, RB.; Warren, DH. Intersensory interactions. In: Boff, KR.; Kaufman, L.; Thomas, JP., editors. Handbook of Perception and Human Performance, Volume I: Sensory Processes and Perception. New York, NY: Wiley; 1986. p. 25-21-25-36.

Yamamoto E, Nakamura S, Shikano K. Lip movement synthesis from speech based on Hidden Markov Models. Speech Communication. 1998; 26(1–2):105–115.

Yehia H, Rubin P, Vatikiotis-Bateson E. Quantitative association of vocal-tract and facial behavior. Speech Communication. 1998; 26(1–2):23–43.

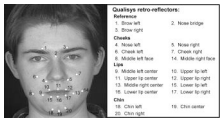

Figure 1.
Placement of optical retro-reflectors. The figures shows one talker with retro-reflectors placed on her face. Placement was similar for the other talkers.

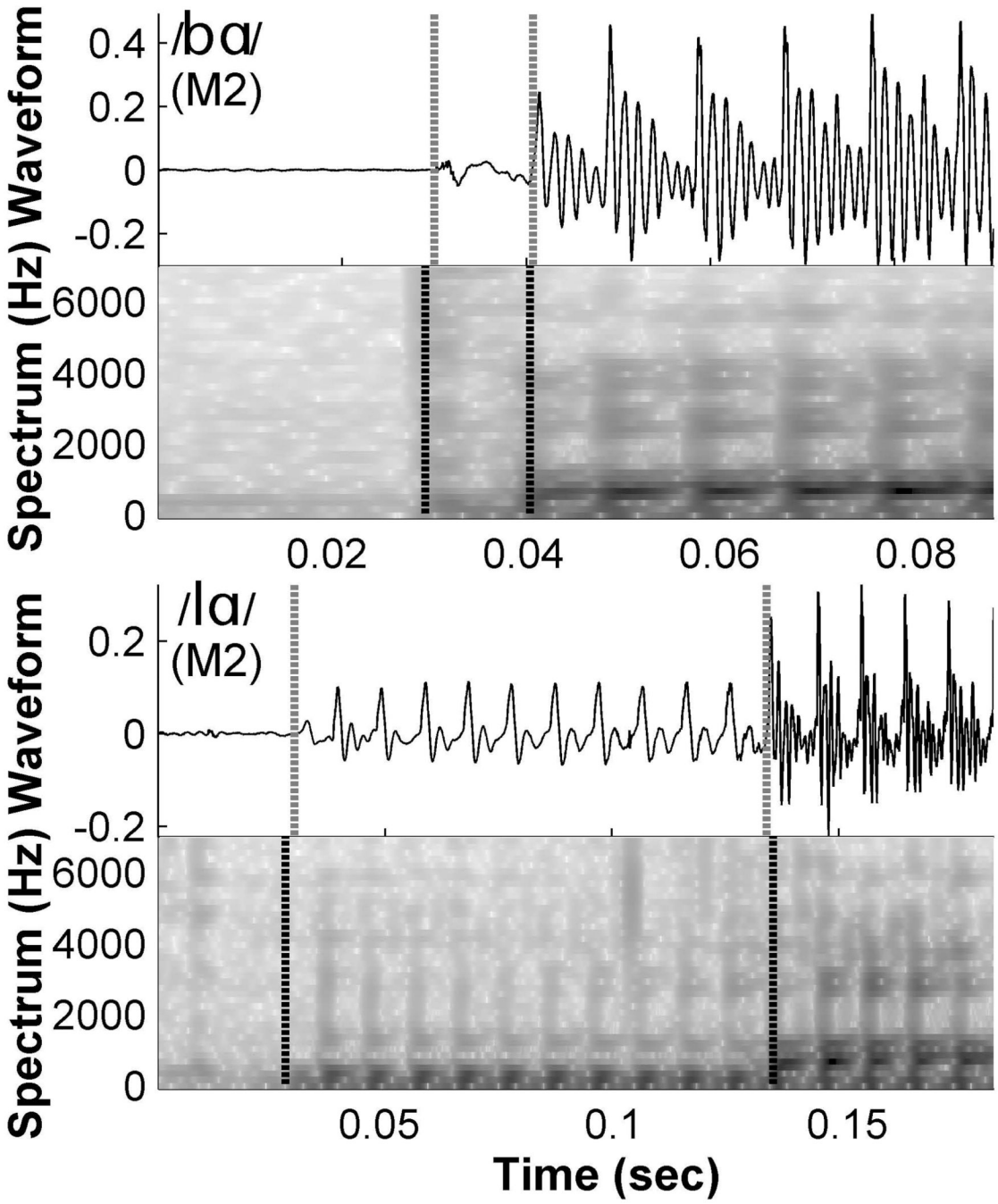

Figure 2.
Labeling of consonant and vowel onsets for a /bA/ (top) and a /lA/ (bottom) token from Talker M2. Waveforms and spectra were used to assist the labeling of consonant and vowel onsets.

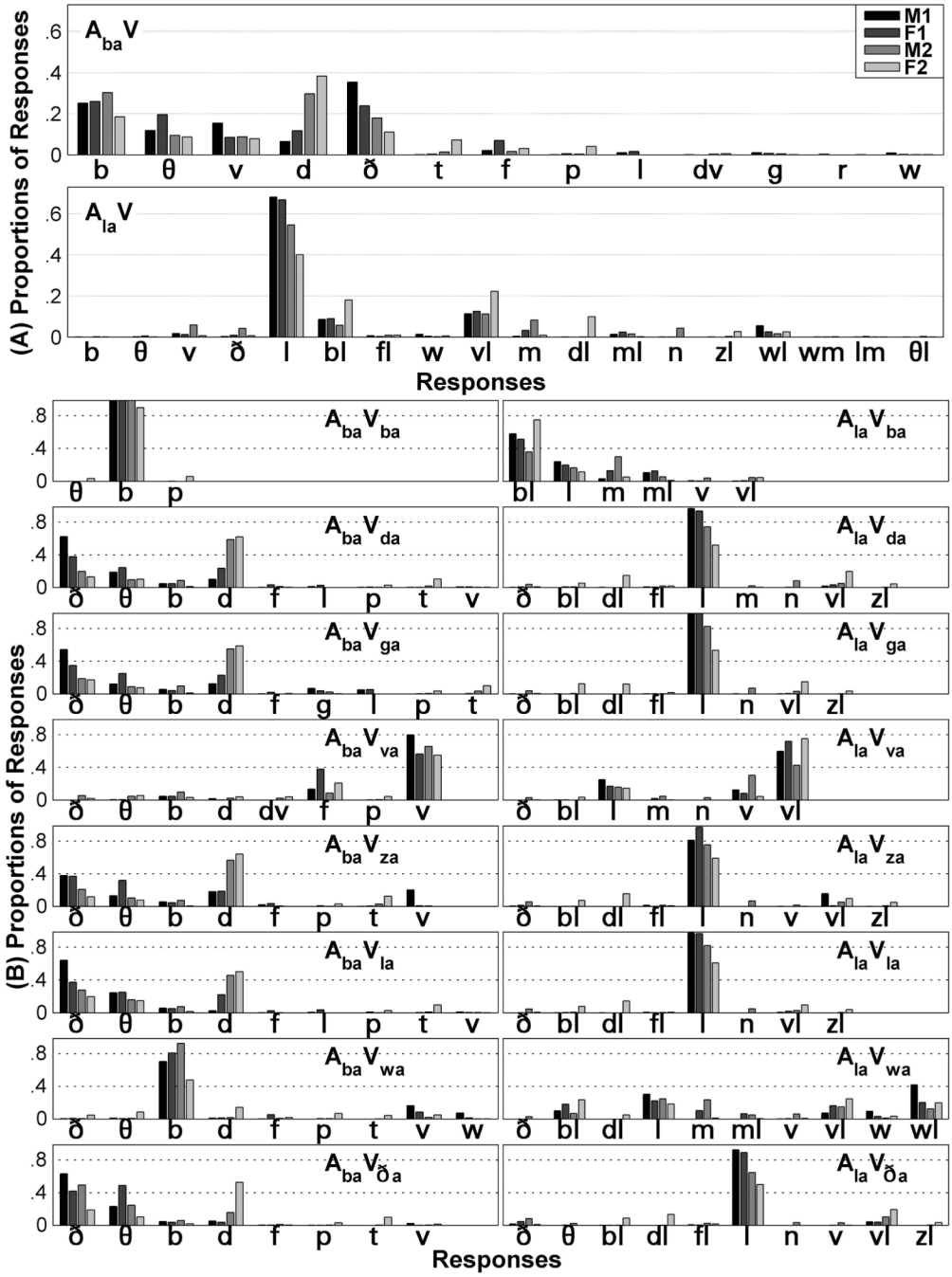

Figure 3.

Proportions of open-set consonant identifications. Proportions in Part A pool across video stimuli for the four talkers (M1, F1, M2, and F2), and proportions in Part B pool across tokens within each audiovisual stimulus type for the four talkers. The top of Part A shows response proportions to $A_{bA}V$, and the bottom of Part A shows response proportions to $A_{lA}V$. The infrequent (fewer than 10) responses are not shown. The bars account for 99.6% of $A_{bA}V$ and 99.7% of $A_{lA}V$ responses. In Part B, consonant identification response proportions omit infrequent responses.

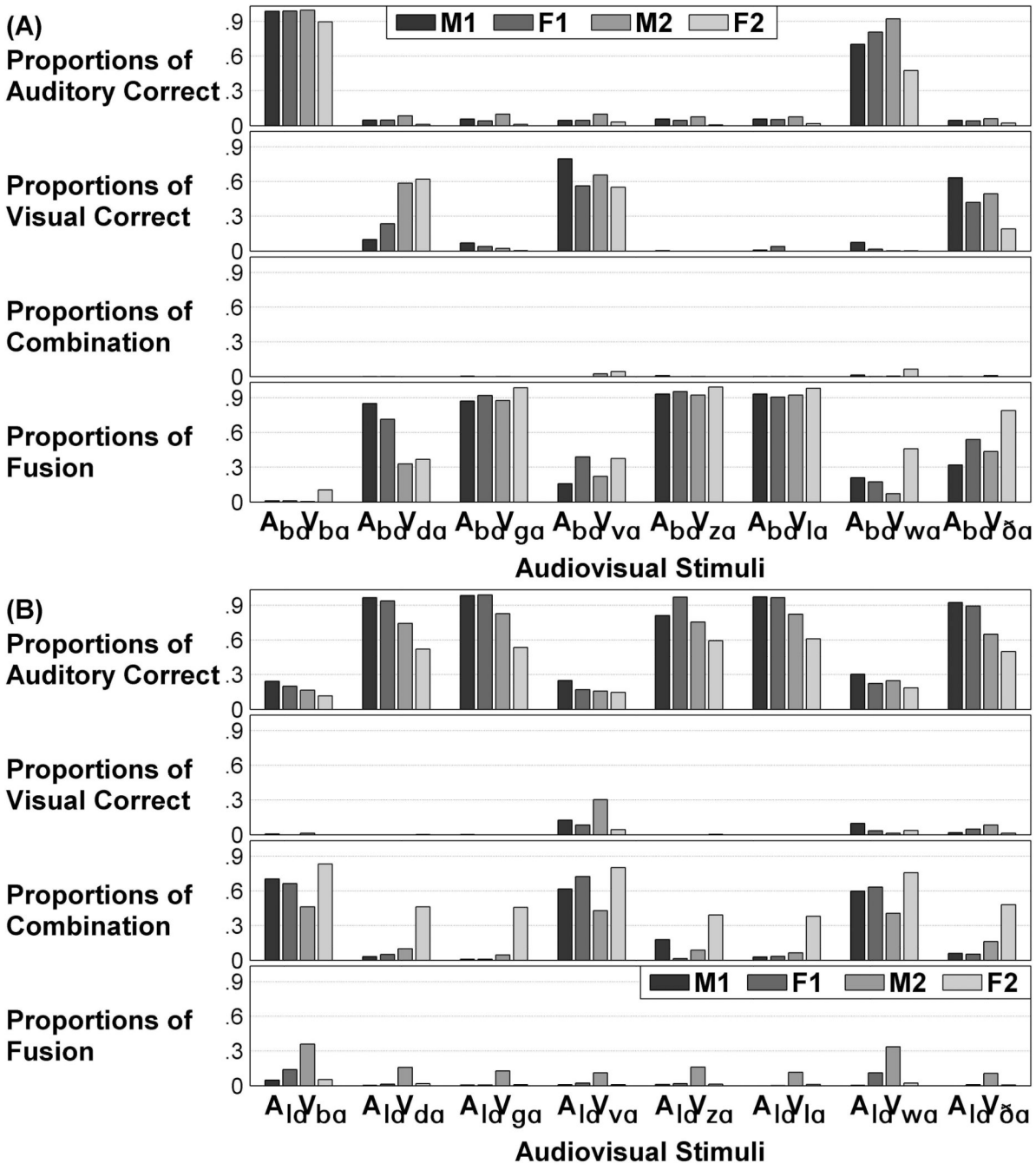

Figure 4.

Categorized responses (auditory correct, visual correct, combination, and fusion) shown as proportions (*y*-axis) for (A) $A_{bA}V$ and (B) $A_{lA}V$, with the four talkers (M1, F1, M2, and F2).

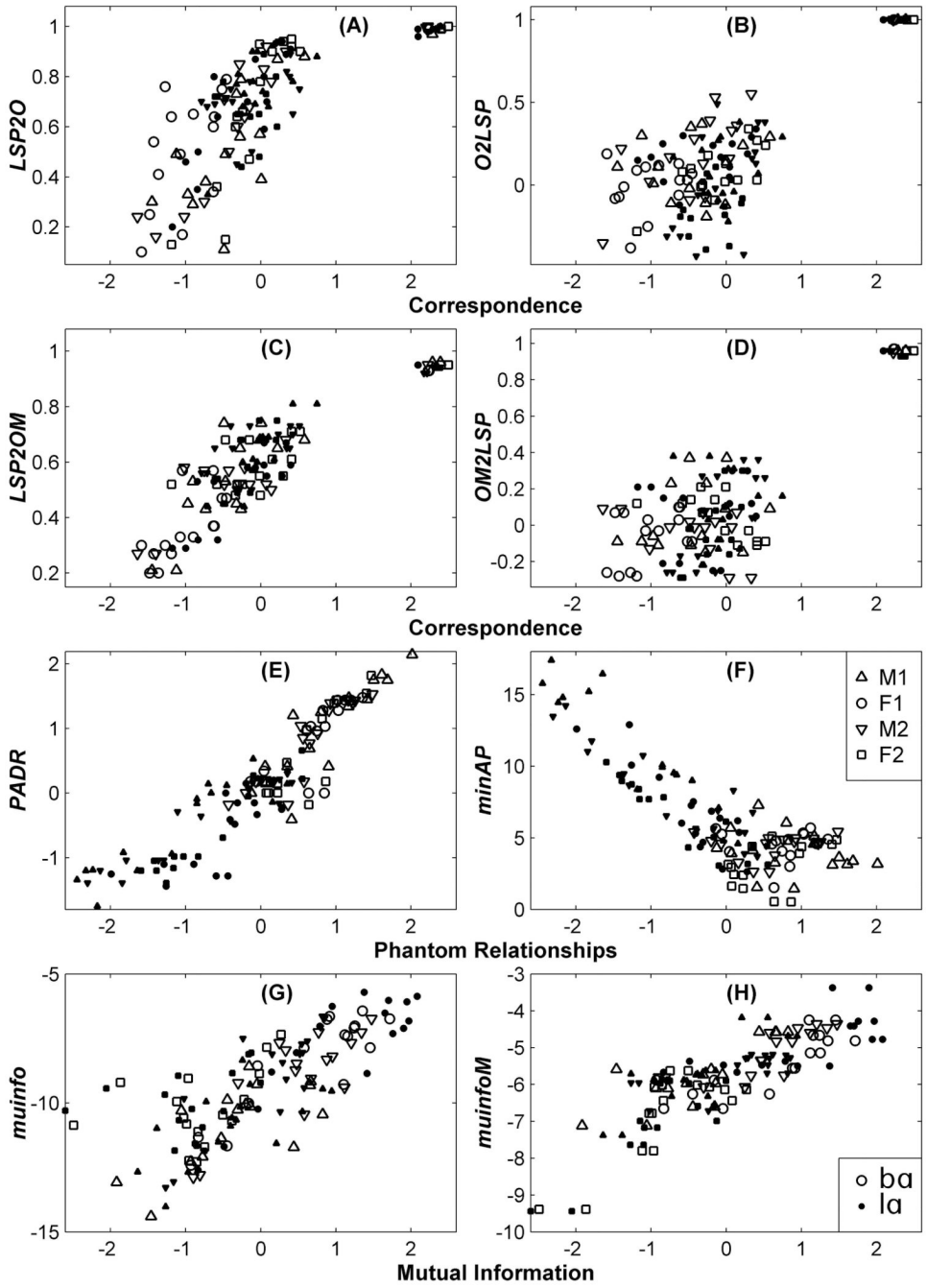

Figure 5.

Scatter plots of latent factors (Factor 1: correspondence, Factor 2: phantom relationships, and Factor 3: mutual information) versus corresponding physical measures for which the rotated component loadings are .40 and above. Data points are graphed separately for stimuli with auditory /bA/ and /lA/. (Note: *LSP2O*, acoustic-to-optical correspondence; *O2LSP*, optical-to-acoustic correspondence; *LSP2OM*, acoustic-to-optical correspondence with the magnetometry data set; *OM2LSP* optical-to-acoustic correspondence with the magnetometry data set; *PADR*, phantom-to-acoustic log duration ratio; *minAP*, minimum acoustic-to-phantom distance; *muinfo*, mutual information; *muinfoM*, mutual information with the magnetometry data set).

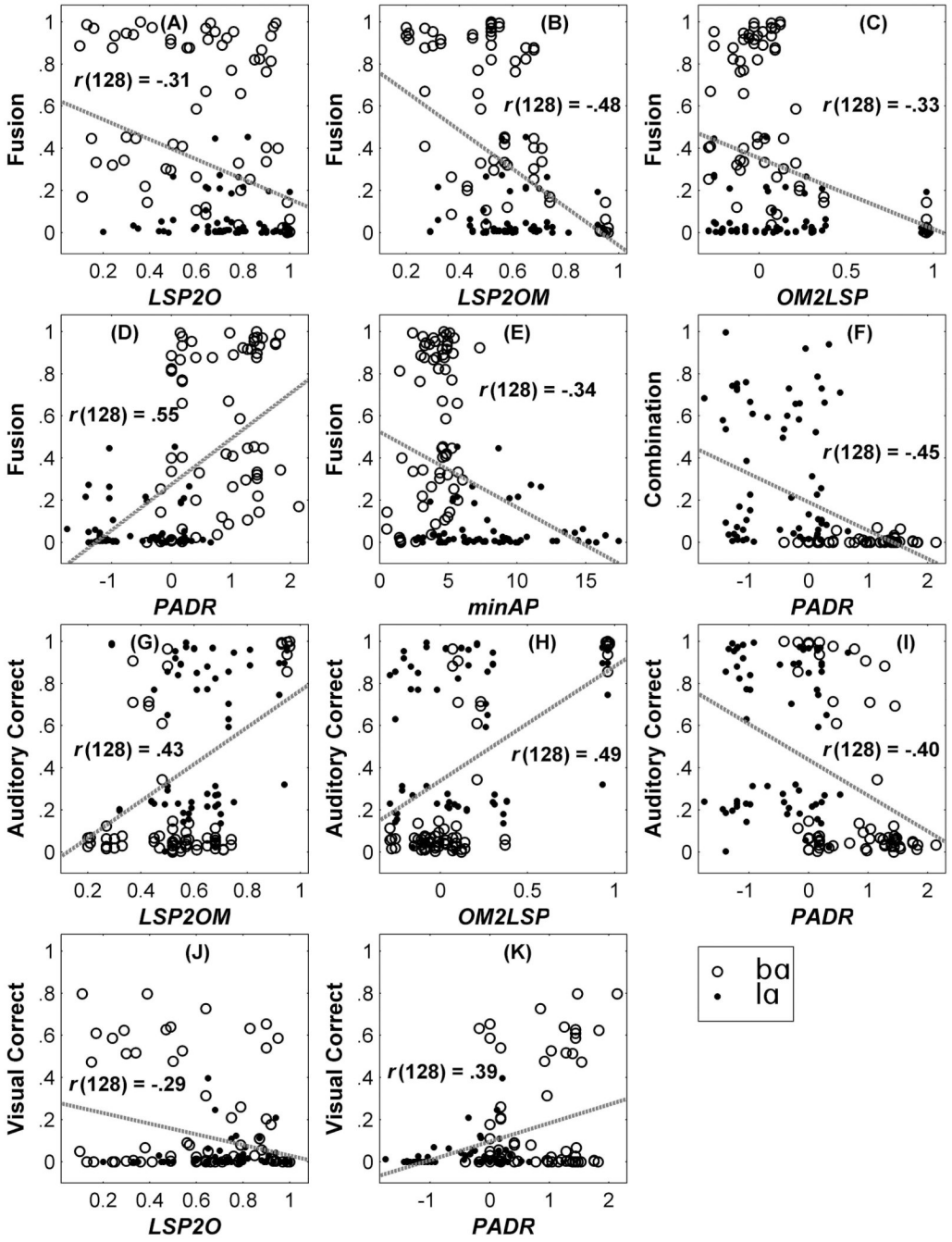

Figure 6.

Scatter plots of physical stimulus measures versus perceptual response proportions across talkers. Individual plots correspond to significant correlations in Table 5 that contributed more than 1% of total variance in the group regression analyses. Each point corresponds to an individual consonant-onset-aligned audiovisual stimulus. (Note: *LSP2O*, acoustic-to-optical correspondence; *O2LSP*, optical-to-acoustic correspondence; *LSP2OM*, acoustic-to-optical correspondence with the magnetometry data set; *OM2LSP* optical-to-acoustic correspondence with the magnetometry data set; *PADR*, phantom-to-acoustic log duration ratio; *minAP*, minimum acoustic-to-phantom distance; *muinfo*, mutual information; *muinfoM*, mutual information with the magnetometry data set).

Table 1

Main effects *F*-values for arcsine-transformed perceptual response categories (auditory correct, visual correct, combination, and fusion), within-subjects stimulus factors (video, talker, audiovisual pairing, and audiovisual alignment), and between-subjects factor (blocked versus mixed audio). The *F*-values for two-way interactions involving alignment and blocking are also listed. The Video x Alignment interaction was not computed due to insufficient residual degrees of freedom.

		$A_{bA}V$			$A_{lA}V$	
Effect	**(df_1,df_2)**	**Auditory correct**	**Visual correct**	**Fusion**	**Auditory correct**	**Combination**
Video	(7, 2)	242.6*	30.3	61.4	4.0	2.5
Talker	(3, 6)	7.9	2.8	18.3*	20.2*	28.0*
Pairing	(1, 8)	7.0	0.1	6.7	30.4*	12.5*
Alignment	(2, 7)	0.3	1.5	0.2	0.8	1.6
Blocking	(1, 8)	1.3	0.3	0.3	0.2	0.2
Video x Blocking	(7, 2)	38.4	32.8	0.2	0.6	1.0
Talker x Blocking	(3, 6)	2.9	1.7	1.7	2.5	9.0
Pairing x Blocking	(1, 8)	0.6	0.0	1.7	0.1	1.8
Talker x Alignment	(6, 3)	0.3	3.1	1.1	1.1	0.6
Pairing x Alignment	(2, 7)	0.6	3.6	1.5	65.8*	0.3

Notes.

* indicates $p < .05$ (with Bonferroni correction). (df_1, df_2) is the degrees of freedom for the *F*-tests.

Table 2

F-values for arcsine-transformed perceptual response categories, auditory correct, visual correct, combination, and fusion, and within-subjects stimulus factors (video, talker, and audiovisual pairing), and their interactions. The *Huynh-Feldt* corrected degrees of freedom were used when the sphericity assumption was violated.

	$A_{bA}V$			$A_{lA}V$	
Effect	**Auditory correct**	**Visual correct**	**Fusion**	**Auditory correct**	**Combination**
Video	$F(1.7,15.4)=157.9^{*}$	$F(3,27.1)=32.0^{*}$	$F(3.8,34)=53.4^{*}$	$F(1.5,13.9)=36.1^{*}$	$F(1.4,12.9)=22.7^{*}$
Talker	$F(1.6,14.1)=9.5^{*}$	$F(3,27)=5.3^{*}$	$F(1.8,16.2)=11.2^{*}$	$F(3,27)=22.7^{*}$	$F(2.1,18.8)=27.2^{*}$
Pairing	$F(1,9)=8.4^{*}$	$F(1,9)=0.1$	$F(1,9)=8.1^{*}$	$F(1,9)=32.8^{*}$	$F(1,9)=12.0^{*}$
Video x Talker	$F(21,189)=8.0^{*}$	$F(21,189)=9.3^{*}$	$F(21,189)=8.3^{*}$	$F(21,189)=8.1^{*}$	$F(21,189)=10.0^{*}$
Video x Pairing	$F(7,63)=0.6$	$F(7,63)=2.7^{*}$	$F(7,63)=2.0$	$F(7,63)=12.3^{*}$	$F(4.4,39.4)=5.4^{*}$
Talker x Pairing	$F(1.4,12.3)=6.6^{*}$	$F(3,27)=7.5^{*}$	$F(1.7,15.4)=12.3^{*}$	$F(1.9,16.9)=9.6^{*}$	$F(1.4,12.2)=15.0^{*}$
Video x Talker x Pairing	$F(21,189)=4.6^{*}$	$F(21,189)=4.1^{*}$	$F(21,189)=3.9^{*}$	$F(21,189)=4.4^{*}$	$F(21,189)=5.3^{*}$

Notes.

*indicates $p < .05$ (with Bonferroni correction). (df_1, df_2) is the degrees of freedom for the *F*-tests.

J Exp Psychol Hum Percept Perform. Author manuscript; available in PMC 2012 August 1.

Table 3

The rotated component loadings on each physical measure for the three latent factors. Loadings of less than .40 were omitted.

Physical measure	Factor1	Factor2	Factor3
LSP2OM	.89		
OM2LSP	.81		
O2LSP	.81		
LSP2O	.74		
PADR		.92	
minAP		−.81	
muinfoM			.83
Muinfo			.76

Notes. *LSP2OM*, acoustic-to-optical correspondence with the magnetometry data set; *OM2LSP*, optical-to-acoustic correspondence with the magnetometry data set; *LSP2O*, acoustic-to-optical correspondence; *O2LSP*, optical-to-acoustic correspondence; *PADR*, phantom-to-acoustic duration ratio; *minAP*, minimum acoustic-to-phantom distance; *muinfoM*, mutual information with the magnetometry data set; *muinfo*, mutual information.

Table 4

Correlations and regressions multiple R^2 across talkers for physical measures and perceptual response categories. All but the last line of the table are correlations, and R^2 is given in the last line. Regression analyses were performed with the stepwise method using only physical measures whose correlation with the response category was significant, as indicated by bolding. The order in which physical measures were entered was based on the magnitude of the correlation (from highest to lowest). Matched audiovisual speech stimuli were excluded (N = 112).

	Response category			
	Fusion	**Combination**	**Auditory Correct**	**Visual Correct**
LSP2O	−.21	.19	.20	−.23
LSP2OM	**−.40**$^{.10}$	.16	.14	.24
O2LSP	.09	−.20	−.02	.14
OM2LSP	−.17	.04	.18	−.06
PADR	**.56**$^{.32}$	**−.47**$^{.22}$	**−.44**$^{.02}$	**.39**$^{.15}$
minAP	**−.49**$^{.03}$	.24	**.50**$^{.25}$	**−.30**$^{.00}$
muinfoM	.08	−.13	.06	−.07
muinfo	.03	−.08	.01	.03
R^2	.44	.22	.27	.15

Notes. Bolding indicates $p < .05$ (with Bonferroni correction), 2-tailed. Superscript indicates the multiple R^2 contributed by the corresponding physical measure.

Table 5

Correlations and regression multiple R^2 across talkers for physical measures and perceptual response categories. All but the last line of the table are correlations, and R^2 is given in the last line. Regression analyses were performed with the stepwise method using only physical measures whose correlation with the response category was significant, as indicated by bolding. The order in which physical measures were entered was based on the magnitude of the correlation (from highest to lowest). Matched audiovisual speech stimuli were included (N = 128).

	Response category			
	Fusion	**Combination**	**Auditory Correct**	**Visual Correct**
LSP2O	**−.31**$^{.02}$	.09	**.38**$^{.00}$	**−.29**$^{.02}$
LSP2OM	**−.48**$^{.15}$	0	**.43**$^{.02}$	.03
O2LSP	−.20	−.23	**.39**$^{.00}$	−.09
OM2LSP	**−.33**$^{.02}$	−.11	**.49**$^{.24}$	−.20
PADR	**.55**$^{.31}$	**−.45**$^{.20}$	**−.40**$^{.15}$	**.39**$^{.15}$
minAP	**−.34**$^{.03}$	.27	.24	−.21
muinfoM	0	−.16	.18	−.11
muinfo	−.01	−.10	.08	0
R^2	.52	.20	.41	.17

Notes. Bolding indicates $p < .05$ (with Bonferroni correction), 2-tailed. Superscript indicates the multiple R^2 contributed by the corresponding physical measure.

Table 6

Correlations for the individual talkers and a summary of the pattern of significant correlations. Matched audiovisual speech stimuli were included.

	Talker F2				Talker M1				Talker F1				Talker M2				Summary			
	Response category				Response category				Response category				Response category				Response category			
	Fus	**Com**	**Aud**	**Vis**	**Fus**	**Com**	**Aud**	**Vis**	**Fus**	**Com**	**Aud**	**Vis**	**Fus**	**Com**	**Aud**	**Vis**	**Fus**	**Com**	**Aud**	**Vis**
LSP2O	−.16	−.03	**.25**[4]	−.06	**−.37**[3]	.04	**.64**[4]	−.57	**−.39**[3]	.19	**.34**[4]	−.29	**−.37**[3]	.20	**.37**[4]	−.22	−(3)	0	+(4)	0
LSP2OM	**−.39**[4]	−.04	**.41**[4]	.13	**−.44**[4]	−.17	**.48**[4]	.06	**−.59**[4]	.03	**.54**[4]	−.11	**−.64**[4]	.10	**.51**[4]	0	−(4)	0	+(4)	0
O2LSP	−.02	−.24	**.31**[4]	−.05	−.17	−.24	**.39**[4]	−.15	−.45	−.02	**.52**[4]	−.32	−.18	−.38	**.32**[4]	.03	0	0	+(4)	0
OM2LSP	−.19	−.18	**.52**[4]	−.19	−.39	−.02	**.40**[4]	−.05	−.43	−.17	**.61**[4]	−.32	−.31	−.07	**.49**[4]	−.29	0	0	+(4)	0
PADR	**.63**[4]	**−.55**[4]	**−.29**[4]	**.25**[4]	**.52**[4]	**−.42**[4]	**−.53**[4]	**.53**[4]	**.62**[4]	**−.40**[4]	**−.47**[4]	**.45**[4]	**.46**[4]	**−.50**[4]	**−.36**[4]	**.30**[4]	+(4)	−(4)	−(4)	+(4)
minAP	**−.37**[4]	**.56**[3]	.02	−.30	**−.45**[4]	**.42**[3]	.31	−.27	**−.24**[4]	.07	.21	−.11	**−.28**[4]	**.33**[3]	.23	−.23	−(4)	+(3)	0	0
muinfoM	.08	−.16	.18	−.19	−.10	.28	.03	−.20	−.17	.01	.24	−.25	.27	−.18	−.11	−.05	0	0	0	0
muinfo	.05	.01	.04	−.18	−.22	.14	−.04	.27	−.06	−.20	.26	−.23	.08	−.12	−.07	.09	0	0	0	0

Notes. Bolding indicates $p < .05$ (with Bonferroni correction) across talkers, 2-tailed, $N = 32$. Superscripts indicate the number of talkers that produced supra-threshold correlations for each physical measure and response category pair. The thresholds for significant correlations were set at: one talker, .59; two talkers, .42; three talkers, .32; four talkers, .23. Fus = fusion; Com = combination; Aud = auditory correct; Vis = visual correct.