Published in final edited form as:

Eur J Neurosci. 2014 April ; 39(8): 1323–1331. doi:10.1111/ejn.12471.

A Visual or Tactile Signal Makes Auditory Speech Detection More Efficient by Reducing Uncertainty

Bosco S. Tjan¹, Ewen Chao², and Lynne E. Bernstein^{1,3}

¹University of Southern California, Department of Psychology, Neuroscience Graduate Program, Los Angeles, CA 90089

²Albany Medical College, Albany, NY 12208

³George Washington University, Department of Speech and Hearing Sciences, Washington, DC 20052

Abstract

Acoustic speech is easier to detect in noise when the talker can be seen. This finding could be explained by integration of multisensory inputs or refinement of auditory processing from visual guidance. In two experiments, we studied two-interval forced choice detection of an auditory “ba” in acoustic noise, paired with various visual and tactile stimuli that were identically presented in both observation intervals. Detection thresholds were reduced under the multisensory conditions versus the auditory-only condition, even though the visual and/or tactile stimuli alone could not inform the correct response. Results were analyzed relative to an ideal observer for which intrinsic (internal) noise and efficiency were independent contributors to detection sensitivity. Across experiments, intrinsic noise was unaffected by the multisensory stimuli, arguing against the merging (integrating) of multisensory inputs into a unitary speech signal; but sampling efficiency was increased to varying degrees, supporting refinement of knowledge about the auditory stimulus. The steepness of the psychometric functions decreased with increasing sampling efficiency, suggesting that the “task-irrelevant” visual and tactile stimuli reduced uncertainty about the acoustic signal. Visible speech was not superior for enhancing auditory speech detection. Our results reject multisensory neuronal integration and speech-specific neural processing as explanations for enhanced auditory speech detection under noisy conditions. Instead, our results support a more rudimentary form of multisensory interaction – the otherwise task-irrelevant sensory systems inform the auditory system about when to listen.

Keywords

speech detection; multisensory enhancement; ideal-observer analysis

---

Corresponding author: lynnebernstein@gmail.com.

Authors' Contributions

L.E.B. and B.T. developed the study concept. All authors contributed to the study design. Testing and data collection were performed by E. C. E. C. performed data analysis under the supervision of B.T., and B.T. carried out independent data analysis. B.T., L.E.B., and E.C. drafted the paper. B.T. and L.E.B. provided critical revisions. All authors approved the final version of the paper for submission.

Conflict of Interest

The authors declare they have no conflict of interests regarding this study.

Introduction

Stimulation to one sensory system can enhance perception of stimuli presented to a different sensory system. For example, auditory stimuli can enhance the perceived intensity of light (B. E. Stein, London, Wilkinson, & Price, 1996), and conversely, light can enhance the perceived intensity of acoustic white noise (Odgaard, Arie, & Marks, 2004). Vibrotactile pulses aid in the detection of tones and increase their perceived loudness (Gillmeister & Eimer, 2007; Ro, Hsu, Yasar, Elmore, & Beauchamp, 2009; Schurmann, Caetano, Jousmaki, & Hari, 2004). Under noisy acoustic conditions, seeing a talker can lower the auditory speech detection threshold (Bernstein, Auer, & Takayanagi, 2004; Eskelund, Tuomainen, & Andersen, 2011; Grant, 2001; Grant & Seitz, 2000; Kim & Davis, 2004; Schwartz, Berthommier, & Savariaux, 2004). The speech detection enhancement could be attributable to audiovisual integration that leads to an amodal “integrated neural signal [that] is different (e.g., bigger, smaller, having a different temporal evolution)” (B. E. Stein et al., 2010). Alternatively, it could be due to visual guidance for listening to the speech in noise (Nahum, Nelken, & Ahissar, 2008).

We investigated speech detection enhancement with respect to an ideal observer, which is a theoretically optimal detector (Green & Swets, 1966; Pelli & Farell, 1999). We used the ideal observer as a standard yardstick to quantify system-level changes with and without multisensory inputs. The ideal observer has the full knowledge about the acoustic stimulus to be detected. Its performance is limited *only* by noise in the stimulus and uncertainty inherent in the task (e.g., uncertainty due to multiple stimuli for the same response).

Ideal-observer model

An ideal-observer *model* can be used to quantify multisensory facilitations using two orthogonal factors: (1) a non-acoustic stimulus could reduce the internal noise of the perceiver (e.g., a visual speech stimulus might recruit an auditory speech-specific process that may be less noisy than a generic sound-detection process); and (2) it could facilitate the extraction of the acoustic signal from the noisy input by appropriately focusing perceptual resources to the relevant information in the signal (e.g., by providing a temporal marker or by correlating with the structure of the auditory stimulus). Multisensory processing could, at least in principle, worsen one factor while improving the other (e.g., integrating a noisy but informative non-acoustic signal with the task-relevant acoustic signal could add noise but also increase efficiency).

An additive noise ideal observer model (Green & Swets, 1966; Pelli & Farell, 1999) explicitly represents and dissociates these two factors. Discriminability between a noisy signal and noise alone, as measured in d' , is expressed as:

$$d'^2 = \eta E / (N + N_{eq}), \quad (1)$$

where E is signal energy, N is the spectral density of the external noise in the stimulus, N_{eq} is the additive noise in the perceptual system, expressed as an equivalent noise source at the input, and η is the sampling efficiency of the perceptual system. For an ideal observer, $\eta = 1.0$ and $N_{eq} = 0$. For humans, $N_{eq} > 0$, and $\eta < 1$.

Intuitively, the internal or “intrinsic” noise, expressed as the equivalent input noise, is the perceptual system’s precision for signal transduction and sensory measurement. For a given sensory-perceptual system, different neural pathways might operate with different amounts of intrinsic noise. The measurable intrinsic noise of a human observer depends on which of the subsystems are recruited, and how/whether the signals are combined. For example, if an auditory speech-specific subsystem had lower intrinsic noise than a general-purpose auditory system, and if a visual stimulus led to an increased utilization of the hypothetical lower-noise speech subsystem, intrinsic noise reduction should be observed. Alternatively, if the visual signal is combined with the auditory signal to form an amodal speech signal (multisensory integration), the noise in the visual system should contribute to the observed intrinsic noise.

Sampling efficiency (sometimes called *statistical efficiency* or *calculation efficiency*) is the fraction of the noise-limited stimulus information that a perceptual system utilizes to perform a task. For example, a system that uses the visual stimulus onset time to attend synchronously to the auditory input will exhibit a higher sampling efficiency for detecting the auditory stimulus than one that ignores that information. In general, the more a system uses the spatiotemporal properties specific to the stimulus, the higher should be its sampling efficiency. We expect efficiency to increase if the perceiver can use knowledge about visual and/or tactile stimuli to pick out the auditory signal from its noise background.

Under the assumption of an additive noise ideal observer model, changes in intrinsic noise (N_{eq}) and sampling efficiency (η) are theoretically independent. As Equation 1 suggests, these parameters can be empirically determined for both unisensory and multisensory conditions by adding external noise (N) to the signal. To minimize effects of nonlinearity in the perceptual system associated with performance level, measurements are typically made at a constant d' . The ideal observer model of Equation 1 can then be rewritten to make explicit that at a constant d' , the signal energy (E) required to achieve the specific d' is linearly proportional to the total spectral density of the internal and external noise (N), and the proportional constant is inversely proportional to sampling efficiency:

$$E = \left(d'^2 / \eta \right) (N + N_{eq}). \quad (2)$$

Hence, an experiment that measures the threshold signal energy as a function of external noise at a constant d' provides a straightforward means to estimate intrinsic noise (N_{eq}) and sampling efficiency (η).¹

The additive noise ideal observer model can account for a broad range of tasks, from simple signal detection to object identification (Green & Swets, 1966; Legge, Kersten, & Burgess, 1987; Pelli & Farell, 1999; Tjan, Braje, Legge, & Kersten, 1995). Furthermore, numerous studies have used similar observer models to study the effects of attention and perceptual

¹The additive noise ideal observer model is known to be an incomplete model for human signal detection performance. While E is often linearly related to N for a given d' , as dictated by Eq. 2, the slope and intercept of this relationship often depend on d' . That is, the squared d' is nonlinearly related to the net signal-to-noise ratio $E / (N + N_{eq})$. More elaborated models attempt to account for this nonlinearity (Lu & Dosher, 2008). Here however, it is sufficient to isolate efficiency from intrinsic noise and to characterize qualitatively the nonlinearity between d' and the net signal-to-noise ratio.

learning on human performance, and in doing so, they demonstrated that efficiency and intrinsic noise are empirically dissociable (Gold, Bennett, & Sekuler, 1999; Lu & Doshier, 1998; Sun, Chung, & Tjan, 2010); see also (Lu & Doshier, 2008) for an extensive review and elaborated theoretical analysis).

The current study

We used the ideal-observer model of Equations 1 and 2 to investigate the data from two experiments. In Experiment 1, speech detection thresholds were measured at four external noise levels, including a no-noise condition, while holding d' constant. Stimulus conditions were audio-only (AO), audio-tactile (AT), audiovisual with a stationary rectangle (AVR), and audiovisual speech (AVS). The tactile stimulus extended generalizability to an additional sensory system. Having demonstrated that intrinsic auditory noise does not change across different multisensory stimuli, in Experiment 2, a more sensitive paradigm was used to examine further whether visible speech stimuli confer significant additional advantage for detection. The four conditions from Experiment 1 and the combination of the visual rectangle with the tactile stimulus (AVRT) and the visual speech with the tactile stimulus (AVST) were presented. Multisensory integration and speech-specific processing are ruled out as explanations for the auditory speech detection enhancement with audiovisual speech. The results point to the ability to use knowledge about visual and/or tactile stimuli to pick out the auditory signal from its noise background.

Experiment 1: Efficiency and Intrinsic Noise

In Experiment 1, speech detection thresholds were measured at four external noise levels, including a no-noise condition, while holding d' constant. We wanted to determine if the ideal-observer model could account for the data, and if so, how efficiency and intrinsic noise might vary across multisensory conditions. Stimulus conditions were audio-only (AO), audio-tactile (AT), audiovisual with a stationary rectangle (AVR), and audiovisual speech (AVS).

Materials and Methods

Participants—We tested four participants (ages 19-37 years, mean 25; 1 male) with American English as their first language, normal or corrected-to-normal vision, normal pure tone thresholds for ten standard frequencies from 250Hz to 8000Hz (ANSI, S3.6-2004), and normal composite scores on the Hearing in Noise Test (HINT) (Nilsson, Soli, & Sullivan, 1994). The participants had average or better lipreading ability (Auer & Bernstein, 2007). They gave informed consent and were paid \$12/hr for their participation. Testing took place over 4-6 sessions (mean 5.5), distributed over 8-71 days (mean 33). Human subject testing was approved by the Institutional Review Board of the St. Vincent's Hospital, Los Angeles, California, which oversees human subjects research at House Ear Institute, Los Angeles, California where the data were collected. The experiments were undertaken with the understanding and written consent of each subject, and the study conforms to the Code of Ethics of the World Medical Association (Declaration of Helsinki), printed in the *British Medical Journal* (18 July 1964).

Stimuli

Auditory—The speech stimulus was a video-recorded “ba” spoken by a female (Bernstein et al., 2004). The 543-ms acoustic syllable was adaptively adjusted in sound level during testing (see below). White noise was presented at 0, 40, 50, and 60 dB SPL. A large (90-sec) file of computer-generated acoustic white noise was sampled randomly for each trial, extending across both intervals and between them, at a constant level throughout a run. The acoustic stimulus and the white noise were mixed using a calibrated audio system, including a custom attenuator and were presented through calibrated ER-3A insert earphones (Etymotic Research Inc., external noise exclusion 30 dB SPL).

Visual—The visual stimuli included the corresponding video of the talker as she pronounced the “ba” syllable (in AVS and AVST conditions) and a static rectangular image (AVR, AVRT) (Figure 1). The visual speech stimulus movement onset coincided with the acoustic syllable onset in the signal-present interval (Figure 2). The visible syllable was longer than the acoustic signal, as is often true with isolated audiovisual speech syllables. To equate for the contrast energy in the visual stimuli, the non-speech visual stimulus was a static rectangle filled with pixels randomly selected from the rectangular region of the visual speech stimulus including the face (Figure 1). The viewing distance was 1m. The face and the rectangle stimuli subtended 6.0 degrees of visual angle horizontally and 8.2 degrees vertically. A fixation cross during AO trials was presented continuously against a grey background and subtended 0.72 degrees of visual angle.

Tactile—A Bruel & Kjaer 4810 minishaker mounted on a wooden stand that incorporated an armrest delivered a vibration stimulus to the right index fingertip. The stimulus was a 200-Hz haversine pulse train (i.e., pulse duration of 2.5ms) of total duration 534ms, with the same onset and offset as the acoustic “ba,” presented via a 0.25-in diameter circular probe. A custom stimulus delivery system incorporated compensation for finger loading. The minishaker was encased in a foam-lined box to attenuate acoustical emissions, and participants wore earmuffs (Bilsom Comfort model #2315, NRR 25dB) throughout testing to guard against detecting acoustic radiation from the vibrating device, although no evidence suggested that vibration was detectable in the presence of the acoustic masking noise. The tactile intensity level was set to the average level at which the stimulus was judged to be equal in intensity to the visual rectangle (7.2 micron peak displacement), following an informal cross-modal intensity matching experiment.

Timing—Synchronized onsets between auditory and visual stimuli, and between auditory and tactile stimuli were permanently established using a pre-recorded stimulus DVD. Figure 2 illustrates the timing within a trial, during which the auditory “ba” stimulus was randomly presented in only one of two observation intervals. The visual speech stimulus began with freeze frames but motion onset coincided with acoustic onset. The visual square and tactile stimuli onset coincided with acoustic onset timing. A total jitter of 167 ms was randomly inserted at the onset and offset of the two observation intervals such that all trials were the same duration. In the AO condition, a fixation cross was presented for the entire 2135 ms of each observation interval, the total duration of the video speech, including freeze frames.

Uniform gray frames of 167 ms duration separated observation intervals in addition to the jitter.

Procedure

A two-interval forced-choice paradigm with adaptive three-down one-up staircase algorithm (Levitt, 1971) was used to obtain 79.4% ($d' = 1.16$ for a 2IAFC design) detection thresholds. Within each testing block, stimulus condition and noise level were fixed, and the “ba” stimulus amplitude was varied. The adaptive step sizes were as follows: At the beginning of the block, 3-dB steps were used until the first reversal following an error; then 2-dB steps until the third reversal; 1-dB until the fifth reversal; 0.5 dB until the eighth reversal; and 0.1 dB for the final four reversals. Thresholds were the arithmetic mean in dB units of all 12 reversal points. In the noise conditions, the initial SNR was -6 dB. In a no-added-noise (quiet) condition, the initial speech level was 10 dB SPL. Two subsequent blocks in each type of condition were initiated with SNRs of 6 dB above the threshold from the previous corresponding stimulus block.

Participants received 15 practice trials per condition and then executed a variable number of testing blocks per session. The conditions were pseudo-randomly ordered and each condition was presented at every noise level once before any were repeated, resulting in 48 blocks (3 repetitions $\times$ 4 conditions $\times$ 4 noise levels). Because the paradigm used adaptive testing, the number of test trials per participant varied somewhat, averaging 65 trials per block.

Participants were told to attempt to detect the auditory stimuli and keep their gaze on the video monitor. They were not explicitly told to attend to the tactile stimuli. It was obvious to the participants that the visual and tactile stimuli were presented in both the signal-present and signal-absent intervals. Participants were instructed to respond as quickly and as accurately as possible when they detected the “ba” auditory stimulus. Responses were made using a two-button box with each button assigned to one of the stimulus intervals. Participants were free to respond during the first interval if they detected the stimulus there. Response times were recorded but not analyzed. LEDs affixed to the sides of the monitor and on the button box indicated the correct response after each response. Testing took place in a double-walled sound booth.

Results and discussion

Each participant contributed 16 thresholds (4 noise levels $\times$ 4 stimulus conditions) averaged over 3 blocks (about 200 trials per threshold). In a repeated measures ANOVA, stimulus type [$F(3, 9) = 36.24, p < .0001$] and noise level [$F(3, 9) = 2015, p < 10^{-12}$] had strong effects on signal thresholds without any significant interaction. Post-hoc pairwise contrasts, corrected for multiple comparisons, revealed the order of signal threshold magnitudes to be AO (27.9 dB SPL) $>$ (AT $\approx$ AVR) $>$ AVS (25.6 dB SPL). That is, all multisensory conditions improved speech signal detection, with visual speech providing the largest gain.

The ideal-observer model (Equation 2) provided a good fit to the data of each participant, accounting for 99% of the variance (Figure 3a). Intrinsic noise (N_{eq}), efficiency (η), and the

standard errors of the estimates were obtained by fitting Equation 2 to the data (Figures 3b-c). Intrinsic noise did not vary across stimulus condition [$F(3, 9) = 1.003, p = .435$]. The average level of intrinsic noise was equivalent to an input noise of 15.8 dB SPL, which is very low compared to the external noise.

In contrast, efficiency was reliably affected by the stimulus condition [$F(3, 9) = 23.42, p < .001$]. Post-hoc pairwise comparisons showed that efficiency was AO (2.2%) < (AT $\approx$ AVR) < AVS (3.6%). Efficiency averaged across conditions was 2.8%. Thus, no evidence was obtained for multisensory integration (i.e., either reduced or increased internal noise), but there was reliable evidence for increased auditory efficiency from visual and tactile stimuli.

Experiment 2: Efficiency and Linearity of Speech Detection

Experiment 1 showed that the equivalent input noise was very low relative to external noise (equivalent to an external noise at 15.8 dB SPL). Equation 1 implies that $d' \approx (\eta E/N)$ whenever external noise is sufficiently high relative to intrinsic noise ($N \gg N_{eq}$). That is, d' measured at high external noise is unaffected by intrinsic noise and can therefore be used as a surrogate for efficiency. This fact was used to obtain a more precise assessment of multisensory facilitation, particularly, the relative effect of visual speech. It was also used to characterize any nonlinearity between d' and SNR, which provides additional insight about the basis for multisensory enhancement.

Experiment 2 was carried out in two phases. In the preliminary phase, SNR thresholds were obtained adaptively at $d' = 1.16$ with the signal fixed at 55 dB SPL and external noise varied. The relatively high signal intensity was used to ensure that performance would not be limited by the weak intrinsic noise. In the main experiment, a common range of SNRs, chosen based on the results from preliminary experiment and applicable to all participants, was used to measure d' s by using the method of constant stimuli (i.e., with both noise and signal fixed within blocks). Two additional conditions were tested in Experiment 2, AVRT and AVST, for which tactile stimuli were presented synchronously with the AVR and AVS stimuli, respectively.

Materials and Methods

Participants—Applying the same selection criteria as in Experiment 1, six participants (age 19-44 years, mean 26; 2 male) took part in the main part of the experiment. The testing was completed in 3 to 7 sessions for each subject, collected over 6 to 93 days. (See supplementary material for the preliminary phase of threshold testing.)

Procedures—The signal level was fixed at 55 dB SPL, and external noise levels of 68, 69 and 70 dB SPL were tested, resulting in three SNR levels. Each test run comprised 36 two-interval forced-choice trials per d' estimate. Participants were tested in all six conditions at all three SNR levels repeated three times, with each pseudorandomly ordered set of 18 tests completed before the next set.

Results and discussion

Accuracy data in the form of proportion correct for each testing block were converted to d' [$d' = 2 \Phi^{-1}(p)$, where p is proportion correct, and Φ^{-1} is the inverse cumulative normal distribution] (Figure 4). The d' values were submitted to a repeated measures ANOVA with three factors: stimulus type (6), SNR (3), and block (3). There were significant main effects of stimulus type [$F(5, 25) = 8.904, p < .001$] and SNR [$F(2, 10) = 130.737, p < .001$] only. Post-hoc pairwise contrasts revealed that in terms of d' , AO (1.2) < AT < (AVR, AVRT, AVS, AVST) (mean = 1.8). The AVS, AVST, AVRT, and AT stimuli were not reliably different from each other. Visible speech was not a reliably better stimulus than non-speech multisensory stimuli for enhancing auditory speech detection in noise.

Results across experiments

In Experiment 1, threshold signal energy (E) and noise power spectral density (N) followed a linear relationship, as dictated by the ideal-observer model (Eq. 2) at the tested d' of 1.16. We estimated the parameters (N_{eq} and η) of the ideal-observer model with data from Experiment 1 and used the model (Eq. 1) to predict the d' values from Experiment 2 (Figure 4, dashed line). This between-subjects cross-experiments prediction was particularly good in the vicinity of the tested d' of 1.16, even though the corresponding external noise level for this d' level in Experiment 2 was 10 dB higher than the highest noise level used in Experiment 1.

However, the ideal-observer model estimated using data from Experiment 1 systematically underestimated the d' s from Experiment 2 at higher SNR levels (at external noise levels of 69 and 68 dB SPL). The slopes between $\log(d')$ and $\log(E/N)$ were significantly greater than the predicted value of 1/2 from Equation 1 (Figure 4). That is, the human psychometric functions, measured in Experiment 2, have steeper log-log slopes than that of the ideal-observer model. A steeper log-log slope is consistent with an observer who does not know the signal exactly and has to consider one of many possibilities (Graham, 1989; Pelli, 1985; Tanner, 1961; Tjan, Lestou, & Kourtzi, 2006). This can be understood intuitively. At high SNR (high E/N), simultaneously considering one of many signal possibilities, even when there was just one specific signal has no impact on performance. This is because only one of possibilities is a good match to the high-SNR signal. In this case, the real observer is essentially like the ideal observer who knows the signal exactly. At low SNRs however, contemplating multiple possibilities other than the exact specification of the signal increases the false alarm rate, and thus, reduces d' relative to a signal-known-exactly ideal observer. Hence, compared to that of an ideal observer, a real observer who is uncertain about the precise specification of the signal will have a steeper slope, with a disproportionately poorer performance at low SNR.

Inasmuch as d' is monotonically related to efficiency at high external noise level (Equation 1), d' can be used as a surrogate for efficiency to test whether uncertainty reduction is the mechanism responsible for the multisensory facilitation. Figure 5 shows the log-log slope for individual participants against their averaged d' across all SNR levels. There is a strong negative correlation ($r = -0.67, p < .001$) between the log-log slope and d' . If we interpret the log-log slope as uncertainty and d' as efficiency, then the negative correlation means that

the higher a participant's uncertainty was about the target, the lower was the participant's efficiency. This can be seen also in terms of the specific conditions in Figure 5, with low efficiency and high log-log slope for AO (and AT for one participant), and high efficiency and lowest log-log slope for individual AVS, AVR, AVST, and AVRT points. This supports the view that visual by itself or with tactile stimuli during auditory speech detection in noise reduces the uncertainty about when or how to listen.

General Discussion

In Experiment 1, the task-irrelevant visual and/or tactile stimuli increased sampling efficiency for auditory speech detection without having any effect on intrinsic noise (Figure 3). The relationship of threshold acoustic signal energy estimates across conditions was $AO > (AT \approx AVR) > AVS$. In Experiment 2, using the method of constant stimuli with three external noise levels at least five orders of magnitude (50 dB) higher than the intrinsic noise, the order of d' values was $AO < AT < (AVR, AVRT, AVS, AVST)$, providing further evidence for a multisensory sampling efficiency advantage but with no evidence for a visual speech advantage. Furthermore, across participants and multisensory conditions, lower log-log slope of the psychometric function (d' vs. E/N) was associated with higher efficiency (higher d' obtained in high external noise), which is attributable to reduced uncertainty.

Our results show that perception of visual and/or vibrotactile stimuli increases the statistical efficiency of auditory speech detection in noise without altering noise intrinsic to the perceiver. This is a functional statement about the perceptual system that does not uniquely translate into a specific neural implementation. Nevertheless, this functional finding constrains what might be the probable biological implementation, which we discuss below. There are several available mechanisms to account for our findings, but a few mechanisms that have been previously proposed seem unlikely in light of the results reported here.

Multisensory integration and intrinsic noise

Multisensory neuronal integration is frequently offered to explain reduction in perceptual detection thresholds relative to unisensory thresholds (e.g., Gillmeister & Eimer, 2007; Odgaard et al., 2004; Ro et al., 2009; B. E. Stein et al., 1996). However, the multisensory enhancements reported for the current study are not attributable to *multisensory integration*, a term which we take to mean the combination of inputs from multiple sensory representations into a resultant amodal representation².

Our definition of multisensory "integration" is close to the notion of "feature fusion," for which the separate feature representations from different modalities are merged into a single data representation before a perceptual decision is made. An alternative to feature fusion is "decision fusion," where modality specific representations are using to make perceptual decisions before these sometimes ambiguous decisions are merged to form a percept. Based

²The term *integration* is used with various amounts of precision. For some investigators, any type of response effects of multisensory stimulus combinations that cannot be accounted for by various statistical approaches to data analysis (such as summing unisensory responses) are referred to as effects of *integration*. We would prefer to reserve the term *integration* for effects that can be validly attributed to convergence of multisensory representations with resultant amodal representations. All the other effects can be grouped as multisensory interactions, following the example of Kayser and Logothetis (2007).

solely on computational considerations involved in building an automatic audio-visual speech recognition system (including factors such as differences in speech segmentations, data rates, and distinguishable tokens), Meyer et al. argued in favor of decision fusion and against feature fusion, a view that is supported by our empirical findings. Of course, an automatic recognition system is not in general expected to be an instantiation of a neural system.

We cannot attribute the observed multisensory enhancement to multisensory integration (akin to feature fusion), because an integrated representation (B. E. Stein et al., 2010) from multisensory input would in theory comprise each sensory system's representation of the stimulus as well as the system's internal noise; resulting in a change in the net (increased or decreased) intrinsic noise of the perceiver, which we did not observe. Consistent with our observed lack of change in intrinsic noise, Chandrasekaran, Lemus and Ghazanfar (2013) found no change in the magnitude and variability of the firing rates of monkeys' auditory neurons when behavioral performance for vocalization detection was enhanced (speeded up) by the presence of a visible vocalizing face.

While observations of improvements in perceptual performance might seem to imply noise reduction from combining inputs across modalities, such reduction is not guaranteed. For example, from the perspective of estimating noise, when a task-relevant acoustic signal is averaged with a non-acoustic signal that occurs identically with and without the target, all that is contributed is a noisy channel. This increases intrinsic noise without affecting efficiency. (Efficiency is not affected, because a sufficiently high external noise in the acoustic stimulus would render the noise from this uninformative channel inconsequential.) In our experiments, no change in intrinsic noise was observed, providing no support for multisensory integration.

Multisensory stochastic resonance

While the phenomenon of multisensory stochastic resonance (SR) (Harper, 1979) seems to resemble some of our findings, they are unlikely to share the same neural mechanism. Stochastic resonance is a low-signal low-noise phenomenon, in which a non-informative low-amplitude noise added to a weak signal causes part of the signal to exceed an internal threshold of a nonlinear system (Lugo, Doti, & Faubert, 2012). The fact that we did not see any multisensory benefit on intrinsic noise, which is a limiting factor on performance in the regime of low external signal and noise, suggests that stochastic resonance is not the relevant mechanism. Indeed, as shown in Figure 4, our SNRs were 15.8 to 17.8 dB, whereas (within-modal) stochastic resonance demonstrated with normal hearing adults detecting an acoustic pure tone has been demonstrated, for example, at -15 or -20 dB SNR (Zeng, Fu, & Morse, 2000). The SNRs in our experiments therefore seem too high to benefit from stochastic resonance. Another signature of stochastic resonance is that there should be a U-shape function between detection threshold and external noise level (Lugo, Doti, Wittich, & Faubert, 2008), but Figures 3a and 4 show no evidence of an inverted-U shape for the uni- and multi-sensory conditions. Interestingly, Harper (1979) introduced the phenomenon as an "arousal mediator of the sensory interaction." This general qualification, which suggests an up-regulation of processing during a more precisely defined stimulus interval (when the

task-irrelevant noise was on), is in line with our findings, as we discuss in more detail below.

Sampling efficiency

Multisensory enhancements led to increases in sampling efficiency. This means that the perceiver had improved knowledge about the task-relevant acoustic signal whenever a visual and/or tactile stimulus was presented in synchrony, even though the latter stimuli could not by themselves signal which interval contained the acoustic target. Consistent with this interpretation, we observed that the lowering of the slope of the psychometric function was associated with increased efficiency, which is a signature for reduced uncertainty about the signal. This reduction could be due to a more precise marking of the onset moment of the acoustic speech stimuli by the synchronous onset of visual and/or tactile inputs. The visual and/or tactile signal/s could be used to deploy attention at an advantageous moment in time (Megevand, Molholm, Nayak, & Foxe, 2013). That is, their onset could cue the moment to attend to the possible occurrence of the auditory stimulus (Power, Lalor, & Reilly, 2011), thus excluding any influence of internal (intrinsic) or external (stimulus) noise outside of the expected temporal interval of the stimulus, increasing sampling efficiency as a result.

Furthermore, if the onset signal from the visual and/or tactile modality is produced under high threshold (requiring high signal-to-noise ratio), then the intrinsic noise in the non-acoustic channel will have little effect in the onset signal, leading to no detectable changes in the measured intrinsic noise, consistent with our finding. The actual magnitude of the effect on sampling efficiency could depend on the stimulus-onset-asynchrony of the non-auditory signal (Ghazanfar, Maier, Hoffman, & Logothetis, 2005; Kayser, Petkov, & Logothetis, 2008; Megevand et al., 2013; Raji et al., 2010; Schroeder & Foxe, 2002), which we did not explicitly manipulate at the stimulus level but may vary across sensory modalities because of differences in sensory processing. Such variation may explain the observed differences in sampling efficiency across multisensory conditions.

Beyond an explanation tied to attention, the improvement in efficiency of the non-informative multisensory stimulus could be attributable to amplification of ongoing neuronal activity in primary auditory areas by a high-threshold non-auditory signal. For example, an auditory stimulus that is paired with a simultaneous somatosensory stimulus will elicit stronger A1 neuronal responses than an auditory stimulus presented in isolation, as shown in awake behaving macaques (Lakatos, Chen, O'Connell, Mills, & Schroeder, 2007). Schroeder and colleagues (Schroeder, Lakatos, Kajikawa, Partan, & Puce, 2008) theorize that “visual cues amplify the cortical processing of accompanying vocalizations by shifting the phase of ongoing neuronal oscillations so that the auditory inputs tend to arrive during a ‘high excitability state’.” As those authors point out, for a weak auditory signal, the effects of phase shifting could determine whether or not inputs generate reliable post-synaptic effects. Compatible with the current results, their explanation for multisensory enhancement need not imply integration of stimulus representations (e.g., syllables are not integrated with the static visual rectangles) or intrinsic noise reduction—merely auditory response enhancement during the interval when the auditory signal is expected. Thus, improved efficiency could arise at least in part due to enhancement in auditory processing induced by

the synchronous onset of the visual and/or tactile stimulus. Reducing timing uncertainty in the auditory channel by a visual input can speed up auditory processing, as observed in (Chandrasekaran et al., 2013), without affecting the underlying representation.

That the visual and/or tactile stimulus can be used as a cue to deploy attention is compatible with a multisensory interpretation (Bernstein, Auer, Jiang, & Eberhardt, 2013) of the reverse hierarchy theory of speech processing (Nahum et al., 2008). Reverse hierarchy theory (Hochstein & Ahissar, 2002) posits that under specific conditions, the perceiver can use higher-level knowledge to guide access to lower-level representations that are available in the input. The visual and/or tactile stimuli did not combine with the auditory stimulus but guided discernment of the signal embedded in its acoustic noise background. In fact, it is conceivable that such guidance targets different stages of neural processing, depending on the task and the kind of information available in each sensory stream (Megevand et al., 2013). Here, the onset cue could guide auditory attention to “glimpses” of the vowel that become available as the non-stationary acoustic noise signal varies in amplitude. That is, the onset also reveals approximately how long to attend.

Overall, based on the results in the current study, multisensory enhancement to auditory speech detection cannot be attributed to reduction in intrinsic noise. It can be attributed to more efficient use of auditory input. Unfortunately, behavioral experiments alone cannot adjudicate between several probable neural mechanisms. However, the ideal-observer model could be applied in conjunction with neural measures to further isolate the neural source for enhanced speech detection efficiency.

No speech-specific mechanism

We obtained scant support for the possibility that visual speech stimuli convey unique benefit to the detection task (Bernstein et al., 2004; Eskelund et al., 2011; Grant, 2001; Grant & Seitz, 2000; Kim & Davis, 2004; Schwartz et al., 2004). If there were something special about affording visual and auditory *speech* stimuli for this detection task, the AVS speech condition should have produced reliably better thresholds. When acoustic speech signal amplitude varied adaptively and noise was fixed, there was evidence for an AVS advantage (Experiment 1). However, the AVS advantage disappeared when compared with other multisensory stimuli (AVR, AVRT, and AVST) in Experiment 2 (see also additional results in supporting material). We also showed similarly enhanced detection with stationary stimuli (the AVRT and AVR condition) as with visual speech (Bernstein et al., 2004).

A prominent previous explanation for the AVS detection advantage is that fine-grained correlations between the acoustic and visual speech stimuli are used by the perceiver (Eskelund et al., 2011; Grant, 2001; Grant & Seitz, 2000; Schwartz et al., 2004). Such a mechanism would introduce the noise of the visual system into the percept in the AVS condition, but we could not detect any change in the intrinsic noise. The lack of any difference between the visual-speech condition and static rectangle conditions also argues against that explanation.

As suggested above, the correlation between onsets of visual and auditory stimulus events is likely critical to the multisensory advantage (Schwartz et al., 2004), even without recourse

to multisensory integrative mechanisms of the type that might result in amodal representations. In Bernstein et al. (2004), when the visual speech stimulus included a preparatory mouth gesture preceding the acoustic stimulus onset, there was evidence for a significant effect of visual speech that was not found when the preparatory gesture was removed from the stimulus.

Lifelong experience of audiovisual speech stimuli potentially makes available predictions about the natural relationships between auditory and visual speech features (Bernstein, Lu, & Jiang, 2008; Jiang & Bernstein, 2011). Natural running speech with its asynchronously available visual and auditory onsets of syllables or phonemes offers multiple opportunities to predict the features of auditory speech stimuli based on visual speech features. In the current experiment, those opportunities were not available, and there was no evidence that visual speech conferred a special advantage for auditory speech detection. An obvious extension of the current study would be to reintroduce the natural preparatory mouth gesture in the current visual speech stimulus to determine whether its effect is to enhance efficiency, which we predict, rather than reduce intrinsic noise.

Conclusion

Using ideal-observer analysis, we identified the functional mechanism for the multisensory enhancement observed in the detection of a speech token in noise. The enhancement is due to an increase in the statistical efficiency of the perceiver caused by a reduction in the uncertainty about the speech signal, mostly likely related to its onset time but possibly also due to knowing its temporal extent. Our analysis rejected mechanisms that require combining signals from multiple sensory streams in the multisensory conditions, because we did not observe any change in the perceivers' internal noise. We advocate a flexible scheme of multisensory facilitations for which the point(s) of multisensory interactions can be task-dependent.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by NIH (R01s DC008583 and DC008308, Bernstein PI; R01 EY017707, Tjan PI). The authors thank Brian Chaney and John Jordan for their hardware and software contributions to this research.

References

- ANSI. Specification for audiometers. S3.6-2004
- Auer ET Jr, Bernstein LE. Enhanced visual speech perception in individuals with early-onset hearing impairment. *Journal of Speech Language and Hearing Research*. 2007; 50(5):1157–1165.
- Bernstein LE, Auer ET Jr, Jiang J, Eberhardt SP. Auditory perceptual learning with degraded speech can be enhanced by audiovisual training. *Frontiers in Auditory Cognitive Neuroscience*. 2013:1–16.
- Bernstein LE, Auer ET Jr, Takayanagi S. Auditory speech detection in noise enhanced by lipreading. *Speech Communication*. 2004; 44(1-4):5–18.
- Bernstein LE, Lu ZL, Jiang J. Quantified acoustic-optical speech signal incongruity identifies cortical sites of audiovisual speech processing. *Brain Res*. 2008; 1242:172–184. [PubMed: 18495091]

- Chandrasekaran C, Lemus L, Ghazanfar AA. Dynamic faces speed up the onset of auditory cortical spiking responses during vocal detection. *Proceedings of the National Academy of Science*. 2013
- Eskelund K, Tuomainen J, Andersen TS. Multistage audiovisual integration of speech: dissociating identification and detection. *Experimental Brain Research*. 2011; 208:447–457. [PubMed: 21188364]
- Ghazanfar AA, Maier JX, Hoffman KL, Logothetis NK. Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *The Journal of Neuroscience*. 2005; 25(20):5004–5012. doi: 25/20/5004 [pii] 10.1523/JNEUROSCI.0799-05.2005. [PubMed: 15901781]
- Gillmeister H, Eimer M. Tactile enhancement of auditory detection and perceived loudness. *Brain Research*. 2007; 1160:58–68. [PubMed: 17573048]
- Gold JM, Bennett PJ, Sekuler AB. Ideal observer analysis of crowding and the reduction of crowding through learning. *Journal of Vision*. 1999; 10(5)
- Graham, NVS. *Visual Pattern Analyzers*. Vol. 16. Oxford University; Oxford, U.K.: 1989.
- Grant KW. The effect of speechreading on masked detection thresholds for filtered speech. *Journal of the Acoustical Society of America*. 2001; 109(5):2272–2275. [PubMed: 11386581]
- Grant KW, Seitz PF. The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*. 2000; 108(3 Pt 1):1197–1208. [PubMed: 11008820]
- Green, DM.; Swets, JA. *Signal Detection Theory and Psychophysics*. Wiley & Sons; New York: 1966.
- Harper DW. Signal detection analysis of effect of white noise intensity on sensitivity to visual flicker. *Perceptual and motor skills*. 1979; 48(3 Pt 1)
- Hochstein S, Ahissar M. View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron*. 2002; 36(791-804)
- Jiang J, Bernstein LE. Psychophysics of the McGurk and other audiovisual speech integration effects. *Journal of Experimental Psychology: Human Performance and Perception*. 2011; 37(4):1193–1209.
- Kayser C, Logothetis NK. Do early sensory cortices integrate cross-modal information? *Brain Struct Funct*. 2007; 212(2):121–132. doi: 10.1007/s00429-007-0154-0. [PubMed: 17717687]
- Kayser C, Petkov CI, Logothetis NK. Visual modulation of neurons in auditory cortex. *Cereb Cortex*. 2008; 18(7):1560–1574. [PubMed: 18180245]
- Kim J, Davis C. Investigating the audio-visual speech detection advantage. *Speech Communication*. 2004; 44:19–30.
- Lakatos P, Chen CM, O'Connell MN, Mills A, Schroeder CE. Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron*. 2007; 53(2):279–292. [PubMed: 17224408]
- Legge GE, Kersten D, Burgess AE. Contrast discrimination in noise. *J Opt Soc Am A*. 1987; 4(2):391–404. [PubMed: 3559785]
- Levitt H. Transformed Up-Down Methods in Psychoacoustics. *Journal of the Acoustical Society of America*. 1971; 49(2B):467–477. [PubMed: 5541744]
- Lu Z-L, Doshier BA. External noise distinguishes attention mechanisms. *Vision Res*. 1998; 38(9)
- Lu Z-L, Doshier BA. Characterizing observers using external noise and observer models: Assessing internal representations with external noise. *Psychological Review*. 2008; 115(1):44–82. [PubMed: 18211184]
- Lugo JE, Doti R, Faubert J. Effective tactile noise facilitates visual perception. *Seeing and Perceiving*. 2012; 25:29–44. [PubMed: 22353567]
- Lugo JE, Doti R, Wittich W, Faubert J. Multisensory integration: Central processing modifies peripheral systems. *Psychological Science*. 2008; 19(10)
- Megevand P, Molholm S, Nayak A, Foxe JJ. Recalibration of the multisensory temporal window of integration results from changing task demands. *PLoS One*. 2013; 8(8):e71608. doi: 10.1371/journal.pone.0071608. [PubMed: 23951203]
- Nahum M, Nelken I, Ahissar M. Low-level information and high-level perception: The case of speech in noise. *PLoS Biology*. 2008; 6(5):978–991.

- Nilsson M, Soli SD, Sullivan JA. Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America*. 1994; 95(2):1085–1099. [PubMed: 8132902]
- Odgaard EC, Arieh Y, Marks LE. Brighter noise: sensory enhancement of perceived loudness by concurrent visual stimulation. *Cogn Affect Behav Neurosci*. 2004; 4(2):127–132. [PubMed: 15460919]
- Pelli DG. Uncertainty explains many aspects of visual contrast detection and discrimination. *Journal of the Optical Society of America A*. 1985; 2(9):1508–1532.
- Pelli DG, Farell B. Why use noise? *Journal of the Optical Society of America A*. 1999; 16(3):647–653.
- Power AJ, Lalor EC, Reilly RB. Endogenous auditory spatial attention modulates obligatory sensory activity in auditory cortex. *Cereb Cortex*. 2011; 21(6):1223–1230. doi: 10.1093/cercor/bhq233. [PubMed: 21068187]
- Raij T, Ahveninen J, Lin F-H, Witzel T, Jaaskelainen IP, Letham B, Israeli E, Sahyoun C, Vasios C, Stufflebeam S, Hamalainen M, Belliveau JW. Onset timing of cross-sensory activations and multisensory interactions in auditory and visual sensory cortices. *European Journal of Neuroscience*. 2010; 31:1772–1782. [PubMed: 20584181]
- Ro T, Hsu J, Yasar NE, Elmore LC, Beauchamp MS. Sound enhances touch perception. *Experimental Brain Research*. 2009; 195(1):135–143. [PubMed: 19305983]
- Schroeder CE, Foxe JJ. The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. *Cognitive Brain Research*. 2002; 14:187–198. [PubMed: 12063142]
- Schroeder CE, Lakatos P, Kajikawa Y, Partan S, Puce A. Neuronal oscillations and visual amplification of speech. *Trends Cogn Sci*. 2008; 12(3):106–113. [PubMed: 18280772]
- Schurmann M, Caetano G, Jousmaki V, Hari R. Hands help hearing: Facilitatory audiotactile interaction at low sound-intensity levels. *Journal of the Acoustical Society of America*. 2004; 115(2):830–832. [PubMed: 15000194]
- Schwartz JL, Berthommier F, Savariaux C. Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*. 2004; 93(2):B69–78. [PubMed: 15147940]
- Stein BE, Burr D, Constantinidis C, Laurienti PJ, Meredith AM, Perrault TJ Jr, Ramachandran R, Roder B, Rowland BA, Sathian K, Schroeder CE, Shams L, Stanford TR, Wallace MT, Yu L, Lewkowicz DJ. Semantic confusion regarding the development of multisensory integration: a practical solution. *Eur J Neurosci*. 2010; 31(10):1713–1720. doi: EJN7206 [pii] 10.1111/j.1460-9568.2010.07206.x. [PubMed: 20584174]
- Stein BE, London N, Wilkinson LK, Price DD. Enhancement of perceived visual intensity by auditory stimuli: A psychophysical analysis. *Journal of cognitive neuroscience*. 1996; 8(6):497–506. [PubMed: 23961981]
- Sun GJ, Chung ST, Tjan BS. Ideal observer analysis of crowding and the reduction of crowding through learning. *Journal of Vision*. 2010; 10(5):16. doi: 10.1167/10.5.16. [PubMed: 20616136]
- Tanner WP Jr. Physiological implication of psychophysical data. *Annals of the New York Academy of Science*. 1961; 89:752–765.
- Tjan BS, Braje WL, Legge GE, Kersten D. Human efficiency for recognizing 3-D objects in luminance noise. *Vision Res*. 1995; 35(21):3053–3069. [PubMed: 8533342]
- Tjan BS, Lestou V, Kourtzi Z. Uncertainty and invariance in the human visual cortex. *Journal of Neurophysiology*. 2006; 96(3)
- Zeng FG, Fu QJ, Morse R. Human hearing enhanced by noise. *Brain Res*. 2000; 869(1-2):251–255. [PubMed: 10865084]

Figure 1.

Visual stimuli used in different conditions. (A) Fixation cross. This stimulus appeared during the Audi0-Only (AO) speech stimulus condition. (B) The static rectangle comprised the pixels of the first frame of the visible speech stimulus, in a random static pattern. This stimulus appeared during the AudioVisual with a stationary Rectangle (AVR) condition and during the conditions with both the stationary rectangle and the Tactile stimulus (AVRT). (C) Visual speech stimuli (only one frame is shown). The natural moving visual speech stimulus was shown during the AudioVisual Speech condition (AVS). It was also shown during the audiovisual speech and Tactile (AVST) condition.

Figure 2.

Stimulus timing diagram. Each trial comprised of two temporal intervals, with the target acoustic “ba” presented in only one of the intervals (E). All other stimuli for a particular condition were repeated in both intervals. The horizontal extent of the stimuli in the figure corresponds to their temporal interval. The tactile, the visual rectangle, and the acoustic syllable had the same duration (534 ms). In the AVS condition, the talker’s face appeared at the beginning of the interval, but the mouth did not move until the acoustic signal onset (D). Dots on either end of the timeline (E) indicate the frames of temporal jitter – the total jitter around a particular interval was always 167 ms (5 frames). F indicates the noise duration. Up to six stimulus conditions were tested in this study; audio-only (AO) (E), audio-tactile (AT) (A and E), audiovisual with a stationary rectangle (AVR) (C and E), AVR with tactile (AVRT) (A, C, and E), audio with visual speech (AVS) (D and E), AVS with tactile (AVST) (A, D, and E). In the AO (E) and AT (A and E) conditions, a fixation cross was displayed for the entire interval (B).

a

Figure 3.

Ideal-observer analysis of speech-detection in noise (Experiment 1). (a) Energy (E) of the speech signal is plotted against the power spectral density (N) of the external noise in log units for each participant. The ideal-observer model of Eq. 2 provides an excellent fit to the individual data ($R^2 > 0.99$). Equivalent input noise (intrinsic noise) (b) and sampling efficiency (c) were estimated from the fits. Error bars represent $\pm$ one standard error of the estimates. Multisensory conditions had no effect on intrinsic noise but significantly improved efficiency. Efficiencies were $AO < (AT, AVR) < AVS$, with mean intrinsic noise level estimated at 15.8 dB SPL.

Figure 4.

Mean d' as a function of SNR across the six stimulus conditions tested in the main phase of Experiment 2 using the method of constant stimuli. The speech signal was fixed at 55 dB SPL, and the noise levels were at 68, 69, and 70 dB SPL. Bars indicate $\pm$ one standard error of the mean. The lines are slightly staggered horizontally for ease of viewing. The dashed line shows the predicted d' values from the ideal-observer model (Eq. 1) fitted with data from Experiment 1. There is a good agreement between the two experiments, except that the log-log slope of the empirical psychometric function (d' vs. E/N) is steeper than that predicted by the ideal-observer model, suggesting that uncertainty about the speech target was a limiting factor for the participants.

Figure 5.

Log-log slopes of the empirical psychometric functions obtained in Experiment 2 as a function of mean d' . In experimental conditions, when the external noise level is very high compared to the intrinsic noise (estimated at 15.8 dB SPL, Figure 3b), d' is completely determined by efficiency (Eq. 1) for a given E/N ratio and can therefore be used as a surrogate for efficiency. The log-log slope for each participant in each multisensory condition is plotted against their averaged d' s across the three external noise levels. A strong negative correlation between the log-log slope and d' (regression line, $R=-.67$, $p < .001$) implies that higher efficiency is associated with a reduction in uncertainty about the target speech signal.