Acoustic noise and vision differentially warp the auditory categorization of speech

Gavin M. Bidelman,[a] Lauren Sigley, and Gwyneth A. Lewis[b]
School of Communication Sciences & Disorders, University of Memphis, 4055 North Park Loop, Memphis, Tennessee 38152, USA

(Received 18 March 2019; revised 5 June 2019; accepted 7 June 2019; published online 8 July 2019)

Speech perception requires grouping acoustic information into meaningful linguistic-phonetic units via categorical perception (CP). Beyond shrinking observers' perceptual space, CP might aid degraded speech perception if categories are more resistant to noise than surface acoustic features. Combining audiovisual (AV) cues also enhances speech recognition, particularly in noisy environments. This study investigated the degree to which visual cues from a talker (i.e., mouth movements) aid speech categorization amidst noise interference by measuring participants' identification of clear and noisy speech (0 dB signal-to-noise ratio) presented in auditory-only or combined AV modalities (i.e., A, A+noise, AV, AV+noise conditions). Auditory noise expectedly weakened (i.e., shallower identification slopes) and slowed speech categorization. Interestingly, additional viseme cues largely counteracted noise-related decrements in performance and stabilized classification speeds in both clear and noise conditions suggesting more precise acoustic-phonetic representations with multisensory information. Results are parsimoniously described under a signal detection theory framework and by a reduction (visual cues) and increase (noise) in the precision of perceptual object representation, which were not due to lapses of attention or guessing. Collectively, findings show that (i) mapping sounds to categories aids speech perception in "cocktail party" environments; (ii) visual cues help lattice formation of auditory-phonetic categories to enhance and refine speech identification.
https://doi.org/10.1121/1.5114822

[AKCL] Pages: 60–70

I. INTRODUCTION

In everyday life, we effortlessly combine information from multiple sensory systems to derive a robust unified percept of events. This ability is essential in the context of speech comprehension, whose success requires the many-to-one mapping of continuously varying acoustic signals onto discrete phonetic sound categories (Pisoni, 1973; Harnad, 1987a; Pisoni and Luce, 1987; Liberman and Mattingly, 1989). In the context of speech, this type of categorical perception (CP) is indicated when listeners hear gradually morphed speech sounds as one of only a few discrete phonetic classes, with an abrupt shift in perception near the midpoint of a stimulus continuum (Liberman *et al.*, 1967; Pisoni, 1973; Harnad, 1987b; Pisoni and Luce, 1987; Kewley-Port *et al.*, 1988). Given the rapid rate of speech transmission (~200 words per minute; Miller *et al.*, 1984), successful comprehension demands that observers process the incoming acoustic signal with maximal efficiency. CP facilitates speech perception by grouping unimportant differences within categories and boosting discriminability between categories, thereby providing the listener a more constrained, manageable perceptual space. Presumably, this "downsampling" process of CP also generates needed perceptual constancy in the face of individual variation along multiple acoustic dimensions (e.g., talker variability) (Prather *et al.*, 2009) or, as tested here, signal degradation (e.g., perceiving speech in noise).

CP also manifests in the visual domain, including the perception of faces (Beale and Keil, 1995), colors (Franklin *et al.*, 2008), and visual speech (O'Sullivan *et al.*, 2017). While there is a substantial literature on CP for single-cue (auditory or visual) contexts, less is known about its role in multisensory conditions such as audiovisual (AV) contexts. Multi-cue integration is necessary in face-to-face communication in which visual articulatory information from a talker's face provides a critical complement to what was said. In these AV contexts, dynamic speech features in auditory and visual channels reflect discrete representations of phonetic-linguistic units (phonemes) and corresponding representations of mouth shapes (visemes) (Peelle and Sommers, 2015). Such integration creates a "visual gain" compared to auditory-only speech, especially when the acoustic signal is degraded (Sumby and Pollack, 1954; MacLeod and Summerfield, 1987; Vatikiotis-Bateson *et al.*, 1998; Ross *et al.*, 2007, Golumbic *et al.*, 2013; Xie *et al.*, 2014).

Visual cues aid comprehension of speech in several ways. Synchronous visible mouth movements can guide comprehension by providing cues to both the timing and content of the acoustic signal (Peelle and Sommers, 2015). Spatial cues such as mouth shape help disambiguate less

[a]Also at: Institute for Intelligent Systems, University of Memphis, Memphis, TN 38152, USA. Electronic mail: gmbdlman@memphis.edu
[b]Also at: Institute for Intelligent Systems, University of Memphis, Memphis, TN 38152, USA.

salient speech sounds, while dynamic articulatory cues help predict upcoming elements in the speech stream (Peelle and Davis, 2012) and allocate attention to target sounds (Carlyon *et al.*, 2001). To varying degrees, facial kinematics and acoustic envelopes are correlated and this cross-modality comodulation can enhance target speech information and improve segregation amidst competing signals (Grant and Bernstein, 2019). In addition, lipreading and coherent visual cues enhance the mere detection of speech in noise (Grant and Seitz, 2000; Grant, 2001; Bernstein *et al.*, 2004; Schwartz *et al.*, 2004).

Sentence-level recognition depends on a variety of cues such as semantic context, lexical frequency, and other indexical cues that can aid speech perception, especially in noise (e.g., Boothroyd and Nittrouer, 1988; Helfer and Freyman, 2009). Moreover, prominent theories of multisensory facilitation typically explain AV speech benefits as an increase in redundancy (e.g., RACE models; Miller and Ulrich, 2003; Colonius and Diederich, 2006) or decrease in cognitive demand (Peelle and Sommers, 2015) offered by combining information from multiple modalities. We explore here an alternate, perhaps more fundamental mechanism to account for AV speech benefits that would not depend on such high-level context, lexical, or cognitive effects. Namely, that vision might alter the underlying *categorical (acoustic-phonetic) representations* of speech.

Our scientific premise was based on the notion (heretofore untested) that visual cues might "sharpen" the perceptual (category) units of speech itself rather than modulating listening effort or signal information/redundancy, *per se*. Supporting this hypothesis, visual cues have been shown to influence (alter) speech percepts even when the auditory signal is perfectly clear (McGurk and MacDonald, 1976). Such effects suggest that visual cues shape speech categories and that informational content of the visual signal systematically influences the perceptual identity of speech objects themselves (Massaro and Cohen, 1983; van Wassenhove *et al.*, 2005). Green and Kuhl (1989) theorized that visual and auditory cues of speech contribute complementary phonetic information about place and manner of articulation. They compared participants' CP of a nonvisible voicing feature of auditory stimuli presented in AV and auditory-only conditions and found that AV items yielded perception of a longer voicing boundary relative to the auditory-only speech, suggesting that AV cues for speech are processed together as a global percept.

Presumably, the inherent process of categorizing can be further beneficial to speech perception in degraded (noisy) listening conditions. Phonetic categories (a higher-level code) are thought to be more robust to noise than physical surface features of a stimulus (lower-level sensory code) (Gifford *et al.*, 2014; Helie, 2017; Bidelman *et al.*, 2019). That is, the construction of a perceptual object and natural filtering process of CP might enable category members to "pop out" among a noisy feature space (e.g., Nothdurft, 1991; Perez-Gay *et al.*, 2018; Bidelman *et al.*, 2019). Thus, from a theoretical perspective, the mere process of grouping speech sounds into categories may aid perception of speech in noise. While the benefits of visual cues on degraded speech recognition are well documented (Sumby and Pollack, 1954; MacLeod and Summerfield, 1987; Vatikiotis-Bateson *et al.*, 1998; Ross *et al.*, 2007; Xie *et al.*, 2014), we are unaware of any studies directly assessing how the operation of speech *categorization* itself—a fundamental mode of perception—is influenced by visual cues of a talker (e.g., phoneme-viseme interactions), particularly as a function of sensory uncertainty (e.g., noise) (cf. Bejjanki *et al.*, 2011). This is surprising given the robust CP observed in the visual domain, including perception of faces (Beale and Keil, 1995) and colors (Franklin *et al.*, 2008). Thus, a novel aspect of the current study is to characterize the extent to which visual information (and noise) interact during the core process of *categorization* (i.e., the acoustic-phonetic conversion).

Additionally, a large portion of speech research on AV integration examines "conflict situations" that involve testing performance for incongruent speech cues in the auditory and visual modalities, as in McGurk experiments (McGurk and MacDonald, 1976). Such paradigms can produce intersensory biasing, in which listeners develop a propensity to categorize speech based on cues from the other modality (Walden *et al.*, 1990; Bertelson *et al.*, 2003). Indeed, studies have shown that normal listeners categorizing incongruent AV speech rely more on auditory cues whereas cochlear implant users rely more on visual cues (see also Schorr *et al.*, 2005; Desai *et al.*, 2008). Thus, it is difficult to draw conclusions about AV benefits from studies employing such conflict situations that require congruency resolution.

In the present study, we tested the hypothesis that visual cues aid speech perception by sharpening categorical representations for speech, particularly under noise-degraded listening conditions. We measured participants' behavioral categorization as they identified speech stimuli along an acoustic-phonetic /da/ to /ga/ continuum presented in clear or noise-degraded listening conditions. This paradigm is particularly advantageous for assessing multisensory processing because stimulus features change orthogonal to the behavioral percept (Bidelman *et al.*, 2013), and therefore might be more veridical for studying perceptual "binding" vs mere "integration" in speech perception (for review, see Bizley *et al.*, 2016). To assess acoustic-visual interactions in CP, continua were also presented with or without visual cues (i.e., visemes) of the talker. We predicted that mouth movements accompanying the auditory input should further warp (bias) participants' auditory perceptual space to one or the other end of the speech continuum. Comparing the slopes of listeners' psychometric functions across conditions assessed the degree to which noise and visual cues influenced the strength/precision of CP. Previous studies have shown dissociations in the accuracy (%) and speed (i.e., response times, RTs) of listeners' categorization (Binder *et al.*, 2004). These dual properties of behavior might also be supported by different brain regions (e.g., %-correct: auditory cortex; RTs: inferior frontal cortex; Binder *et al.*, 2004; Chang *et al.*, 2010; Bidelman and Lee, 2015), suggesting categorization can be parsed into sensory-perceptual and decision processes (cf. "early- vs late-stage" or "pre- vs post-labelling" models of AV integration; Braida, 1991; Peelle and Sommers,

2015). Thus, we measured both the accuracy and speed of listeners' speech categorization to assess visual and noise-related modulations in behavior and to tease apart these factors in relation to CP. We hypothesized that noise would weaken the categorical representations of speech but that visual cues would partially counteract noise-related decrements in CP. While some work suggests visually presented syllables (visemes) are categorically perceived (Weinholtz and Dias, 2016), visual contributions are limited under normal circumstances. Thus, we predicted visual phonetic cues might become more effective for ambiguous or degraded speech classification (Massaro and Cohen, 1983).

II. METHOD

A. Participants

Fifteen young adults were recruited to the experiment. One participant's data were lost due to a technical error in data logging. Thus, the final sample consisted of fourteen participants [six males, eight females; age: mean = 26.9, standard deviation = 3.0 years]. All exhibited normal hearing sensitivity [i.e., <25 dB hearing level (HL) thresholds, audiometric frequencies]. Each was strongly right-handed (82.4 ± 17.3% laterality index; Oldfield, 1971), had obtained a collegiate level of education, and had normal or corrected-to-normal vision. Musical training is known to modulate categorical processing and speech-in-noise listening abilities (Parbery-Clark *et al.*, 2009; Bidelman *et al.*, 2014; Smayda *et al.*, 2015; Mankel and Bidelman, 2018). Consequently, we required that participants have minimal music training throughout their lifetime (1.9 ± 2.9 years). All received payment for their time and gave written informed consent in compliance with a protocol approved by the University of Memphis Institutional Review Board.

B. Stimuli: AV speech continua

We used a 7-step, stop-consonant /da/ to /ga/ sound continuum (varying in place of articulation) to assess CP for speech [Fig. 1(A)]. Each sound token (Tk) was separated by equidistant steps acoustically yet was perceived categorically from /da/ to /ga/. Stimulus morphing was achieved by altering the *F*2 formant region in a stepwise fashion using the STRAIGHT software package (Kawahara *et al.*, 2008). We chose a consonant-vowel (CV) continuum because compared to other speech sounds (e.g., vowels), CVs are perceived more categorically (Pisoni, 1973; Altmann *et al.*, 2014) and carry more salient articulatory gestures and visual cues for perception (Moradi *et al.*, 2017). Original video material consisted of a single talker (Talker #6) from the "congruent" set of AV CVs described in Nath and Beauchamp (2012) and Mallick *et al.* (2015).[1] The total length of each video clip ranged from 1.5 to 2.0 s to start and end each speaker in a neutral, mouth-closed position. The acoustic portion of each video (~350 ms corresponding to where the talker was opening/closing her mouth) came from the same talker in the video and was temporally centered within each clip.

Though it would have been a desirable, creating a morphed video channel from /da/ to /ga/ is technically challenging given the time-varying nature of images on the screen and integration of the sound channel. Therefore, to investigate if phonetic visual cues (visemes[2]) enhance the salience of speech CP, we superimposed the morphed acoustic continuum (each of the seven steps) on a prototypical "da" or "ga" video production. For the AV conditions, this resulted in an overlay of audio tokens 1–3 onto the "da" video and tokens 4–7 onto the "ga" video [Fig. 1(A)]. Thus, each half of the acoustic-phonetic /da/-/ga/continuum was latticed with either "da" (Tk 1–3) or "ga" (Tk 4–7) visual cues.[3] That is, we intended to further warp participants' auditory perceptual space to one or the other end of the

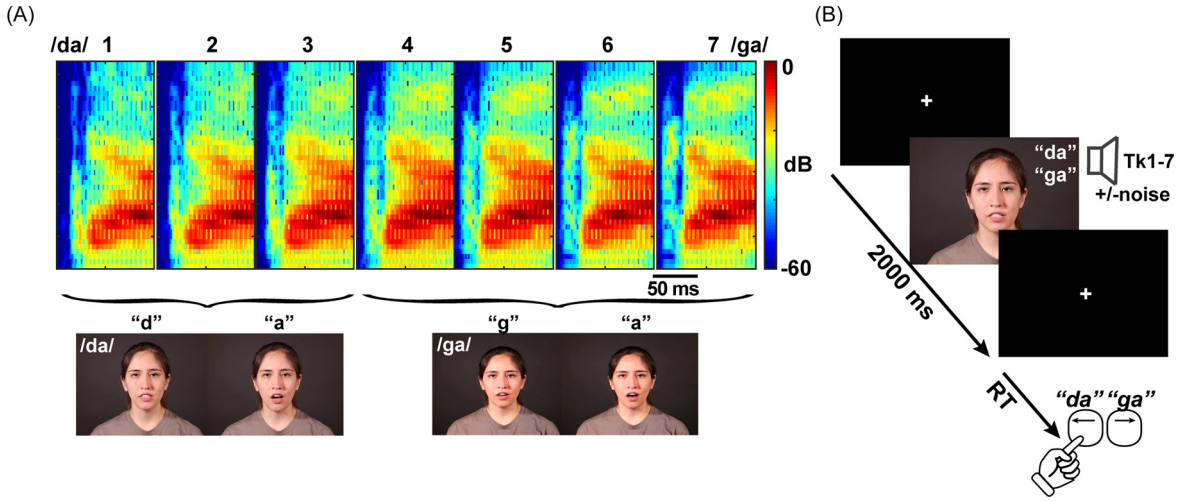

FIG. 1. (Color online) Stimuli and task design. (A) Spectrogram of the /da/-/ga/ continuum. Each of the seven acoustic tokens of the morphed speech continuum was overlaid onto a video of either a prototypical /da/ (Tk 1–3) or /ga/ (Tk 4–7) (Nath and Beauchamp, 2012; Mallick *et al.*, 2015). We intended to further warp (bias) participants' auditory perceptual space to one or the other end of the speech continuum with visual cues of the talker. (B) Single trial time course. After a brief orienting screen (+), participants rapidly identified whether they perceived each audio(visual) token as a "da" or "ga" via computer keypress. Speech stimuli were presented in four different blocks which varied in the number of sensory cues from the talker and the presence/absence of acoustic noise interference: A, A+noise, AV, AV+noise.

continuum via biasing visual cues of the talker. Mixing of the auditory and visual channels was achieved using FFmpeg software and custom routines coded in MATLAB 2013 b (The MathWorks, Inc.).

In addition to these AV conditions, we examined participants' speech categorization for identical tokens where the video channel was absent from the screen [i.e., auditory-only (A) condition]. To further investigate the impact of visual cues on speech categorization, we also constructed a similar AV continuum by partially masking the sound channel with multi-talker babble noise [signal-to-noise ratio (SNR) = 0 dB SNR] (Killion *et al.*, 2004). This SNR was chosen on the basis of previous studies showing behavioral performance is most dynamic under challenging (0 dB or negative) SNRs (e.g., Sumby and Pollack, 1954; Xie *et al.*, 2014; Reetzke *et al.*, 2016).

In total, the stimulus set consisted of four /da/-/ga/ continua that varied in AV cues and noise degradation to the sound channel: A, A+noise, AV, AV+noise. $N = 5$ of the participants also took part in a V-only condition where the speech continuum was presented with a muted sound channel. This pilot control allowed us to assess whether participants could categorize speech sounds based solely on visual cues of the talker. While listeners could identify V-only speech above chance, performance was highly variable. Given that V-only speech did not produce consistent/reliable identification and previous findings that visual speech offers impoverished phonetic detail in isolation (Walden *et al.*, 1977; Kuhl and Meltzoff, 1988; Bernstein and Liebenthal, 2014), we did not consider this condition further.

C. Task procedure

Participants sat in a double-walled sound attenuating chamber (Industrial Acoustics, Inc.) ~90 cm from a computer monitor. Where necessary, participants wore corrective lenses/contacts for the experiment. All confirmed the screen and visual stimuli were clearly visible before initiating the experiment. MATLAB was used as a driving engine to present AV stimuli via the VLC media player as well as collect response data. Stimuli appeared at the center of the screen on a black background, subtending a 7.0° visual angle (Samsung SyncMaster S24B350HL; nominal 75 Hz refresh rate). High-fidelity circumaural headphones (Sennheiser HD 280 Pro) delivered the auditory channel binaurally at a comfortable level [80 dB sound pressure level (SPL)].

Participants heard 210 trials of each individual CV (30/token) presented in the different AV stimulus conditions (four separate blocks). Blocking was used to minimize adding unnecessary cognitive effort or distractibility to the task that would have occurred from trial-to-trial switching between modalities. Block order was counterbalanced across participants according to a Latin square sequence (Bradley, 1958). On each trial, participants labeled the perceived speech token with a binary response via the computer keyboard ("da" or "ga"). They were encouraged to respond as quickly and accurately as possible. Both percent identification and response times (RTs) were logged. Breaks were allowed between blocks to avoid fatigue.

D. Data analyses

For each stimulus condition, we measured the steepness of participants' psychometric function as the slope of each curve where it straddled the CP boundary (i.e., slope $= [PC_{Tk3} - PC_{Tk5}]/2$, where PC_n is the identification score at token n). Larger slopes reflect steeper psychometric functions and hence, stronger CP (Xu *et al.*, 2006; Bidelman and Lee, 2015; Bidelman and Walker, 2017).We measured the location of the CP boundary as the point (token number) along the continuum where the psychometric functions crossed, measured via the MATLAB function InterX.[4] Comparing these metrics between AV and noise conditions assessed possible differences in the location and "steepness" (i.e., rate of change) of the perceptual boundary as a function of AV context and noise interference.

Behavioral RTs for speech labeling speeds were computed as participants' median response latency across trials for a given condition. We excluded outliers (RTs outside 250–6000 ms) from further analysis since these reflect fast guesses and lapses of attention (e.g., Bidelman and Walker, 2017).

Psychometric slopes were analyzed using a mixed-model analysis of variance (ANOVA) (subject = random factor) with fixed effects of SNR (two levels: clean, noise) and modality (two levels: audio, audiovisual) (PROC GLIMMIX, SAS® 9.4). RT data were analyzed with fixed effects SNR, modality, and token (seven levels: Tk 1–7). False discovery rate (FDR) was used to adjust for multiple comparisons (Benjamini and Hochberg, 1995). Effects sizes are reported as partial-eta squares (η_p^2) and mean differences for omnibus ANOVAs and *post hoc* contrasts, respectively.

III. RESULTS

A. Psychometric identification functions

Figure 2 shows psychometric identification functions for the different AV conditions. An ANOVA revealed a significant SNR × modality interaction on psychometric slopes [$F_{1,\ 39} = 7.50$, $p = 0.0092$; $\eta_p^2 = 0.16$]. FDR-corrected paired contrasts revealed that noise weakened CP for auditory stimuli (A vs A+noise: $t_{13} = 4.78$, $p = 0.00145$; mean difference = 19.7%) [Fig. 2(B)]. AV speech was also perceived more categorically than A speech [$t_{13} = 2.51$, $p = 0.035$; mean difference = 5.2%], suggesting that multisensory cues enhance CP. The AV and AV+noise did not differ [$t_{13} = 1.44$, $p = 0.17$; mean difference = 4.7%], implying that CP is robust for multisensory speech even in the presence of noise. Lastly, psychometric slopes were sharper for the AV+noise compared to the A+noise condition [$t_{13} = 4.17$, $p = 0.0011$; mean difference = 20.1%], suggesting that the addition of visual cues helped counteract the negative effects of noise on auditory speech categorization.

Figure 3 shows an alternate presentation of these data, plotting each participant's psychometric slope for one condition vs another for the four major stimulus contrasts of interest. This visualization highlights the relative improvement or decrement in CP with added visual cues and noise, respectively. Points in the upper left half of each plot (above the

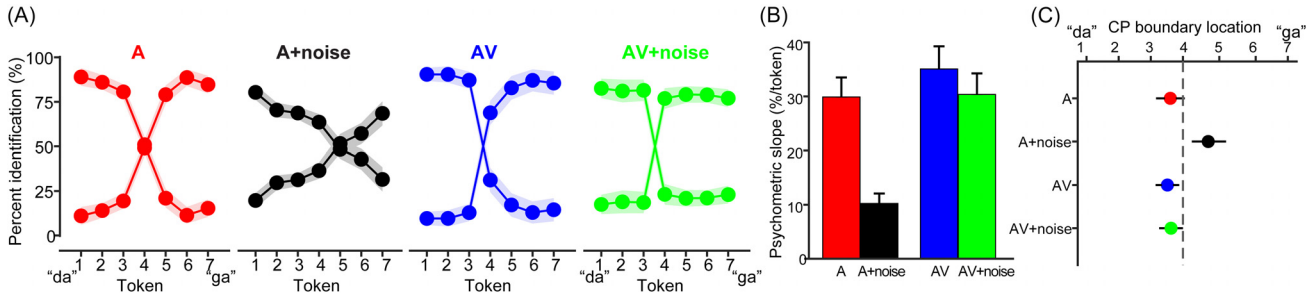

FIG. 2. (Color online) Perceptual identification for AV speech with and without noise interference. (A) Psychometric functions show an abrupt shift in perception when classifying speech indicative of discrete perception (i.e., CP). Degree of categorization varies depending on the sensory modality and quality of cues that are available. Note the two curves within each panel are mirror images since the categorization task is a binary decision. (B) Psychometric function slopes. Sharper identification curves are observed for clean A and AV speech. Acoustic noise weakens CP for speech (A vs A+noise), but this decrement is counteracted by the aid of visual cues of the talker (AV vs AV+noise). (C) Location (cross-over) of the CP boundary. Participants show slight bias to respond "ga" in the A+noise condition but otherwise location varies little across stimulus conditions. Shading and error bars = ±1 standard error of the mean (s.e.m.).

diagonal) show an improvement in categorization for the ordinate relative to the abscissa condition. We found that a majority of participants showed stronger CP when categorizing clean compared to noisy speech (panel A), again confirming that noise weakened phonetic representations of speech. Similarly, stronger CP (steeper identification functions) for AV compared to A speech (panel B) confirms that visual cues helped strengthen phonetic categories. Last, visual cues enhanced categorization for degraded speech compared to the degraded auditory channel alone (panel D).

While the location of the perceptual boundary was largely centered across conditions [Fig. 2(C)], it did shift depending on both noise and multisensory cues [SNR × modality interaction: $F_{1,\,39} = 4.87$, $p = 0.033$; $\eta_p^2 = 0.11$]. In particular, clean AV speech produced a leftward shift in identification curves (cross-over = 3.5; t-test against perceptual boundary at Tk 4; $t_{13} = -2.43$, $p = 0.029$), indicating a small but measurable bias toward producing more frequent "ga" responses across a wider range of the continuum. However, this might be expected given that "ga" visemes were more prevalent across the auditory perceptual continuum than "da" videos. Contrastively, A+noise yielded a rightward shift in the psychometric function (cross-over location = 4.7; $t_{13} = 2.50$, $p = 0.026$) suggesting a bias to more frequently respond "da" in noisy listening conditions. We note that this noise-related effect occurred even though "ga" videos were overall more prevalent across continua [see Fig. 1(A)]. This latter bias effect may reflect top-down influences of lexical knowledge because /da/ has a higher frequency of occurrence than /ga/ in spoken language (Denes, 1963). All other stimulus conditions produced symmetric psychometric functions (ps > 0.09).

B. RTs

Behavioral RTs, reflecting the speed of participants' categorization, are shown in Fig. 4. RTs are plotted relative to the average RT across all conditions to highlight differential changes in categorization speed with noise and visual cues. RTs showed main effects of noise [$F_{1,351} = 137.96$, $p < 0.0001$; $\eta_p^2 = 0.28$] and modality [$F_{1,351} = 5.55$, $p = 0.019$; $\eta_p^2 = 0.016$]. On average, RTs were faster when classifying clean compared to noise-degraded speech (A vs A+noise: $p < 0.0001$; mean difference = 318 ms). Similarly, clean AV speech elicited faster RTs than degraded AV speech (AV vs AV + noise: $p < 0.0001$; mean difference = 403 ms). Participants were equally fast at categorizing clear A and AV speech ($p = 0.49$; mean difference = 30 ms). However, in noise, they were faster at classifying A+noise vs AV+noise speech ($p = 0.0086$; mean diff. = 114 ms). While these findings reveal a prominent effect of noise on the speed of categorical processing, the relative pattern of RTs is often more meaningful: CP is characterized by a slowing in response speed near the ambiguous midpoint of

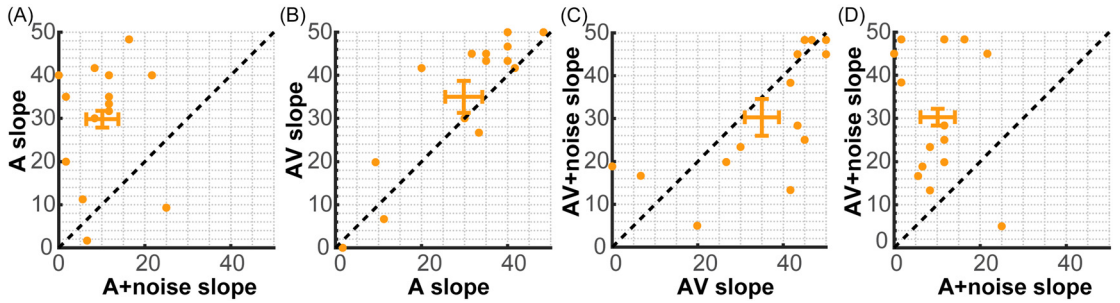

FIG. 3. (Color online) Multisensory cues enhance the CP of clear and especially noise-degraded speech. Individual points show each participant's psychometric slope for two different stimulus continua plotted against one another. Errorbars (±1 s.e.m.) show variance around the group average centroid. Points in the upper left half of each plot (above the diagonal) show an improvement in CP for that ordinate relative to the abscissa condition. A majority of participants show stronger CP when categorizing (A) clean compared to noisy speech sounds, (B) AV compared to A speech, and (D) AV+noise compared to A+noise speech. (C) AV and AV+noise speech yeild similar slopes.

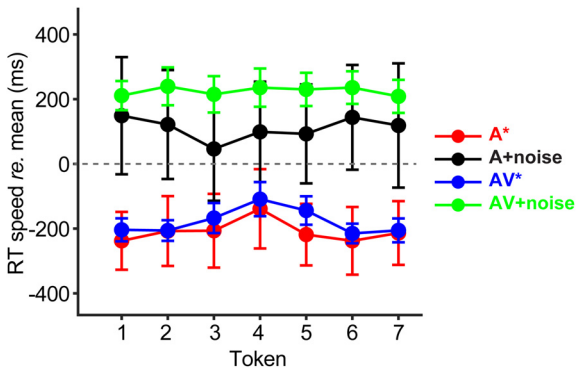

FIG. 4. (Color online) Speech classification speeds [response times (RTs)] vary with sensory modality and clarity of the speech signal. RTs are plotted relative to the mean RT across conditions (dotted line) to highlight the differential in speed of categorization judgments with noise and visual cues. Clean A and AV speech produce the fastest labeling speeds. Participants are 200–300 ms slower at categorizing A and AV speech amidst noise. Only clear A and AV speech produce a categorical pattern of RTs ($*p < 0.05$) where ambiguous speech tokens elicit slower decisions than tokens with clear phonetic categories [i.e., contrast: Tk 4 vs mean (Tk 1,2,3,5,6,7)] (Pisoni and Tash, 1974; Bidelman and Walker, 2017). errorbars = ±1 s.e.m.

the continuum (Pisoni and Tash, 1974; Bidelman *et al.*, 2013; Bidelman *et al.*, 2014; Bidelman and Walker, 2017; Reetzke *et al.*, 2018). To assess this effect, we conducted *a priori* contrasts of RTs on the perceptual boundary (Tk 4) vs others along the continuum [mean (Tk 1,2,3,5,6,7)]. This confirmed CP for A speech [$t_{13} = -2.51$, $p = 0.026$; mean diff. (Tk_{123567} vs Tk_4) = 81.5 ms]. RTs to AV speech similarly showed the categorical (inverted-V) pattern [$t_{13} = -3.16$, $p = 0.0075$; mean difference = 81.3 ms]. A categorical RT effect was not observed for A+noise [$t_{13} = 0.45$, $p = 0.66$] nor AV+noise [$t_{13} = -0.42$, $p = 0.68$; mean diff. = 12.3 ms], further suggesting that noise weakened categorical decisions in these degraded stimulus conditions.

Even in the absence of significant differences in central tendency, RTs can differ in terms of intrasubject variability (Bernstein *et al.*, 2014; Bidelman *et al.*, 2017). Reduced RT variability in certain conditions might also reflect improved cognitive processing (cf. Strauss *et al.*, 2002). Visual inspection of RT variance suggested more restricted, less variable response speeds for multisensory AV stimuli (cf. error bars for A vs AV). Pooling across tokens, formal tests of equal variance (two-sample F-test) revealed that RT dispersion was indeed smaller when categorizing AV compared to the A speech for both the clean [$F_{13,13} = 7.72$, $p = 0.0008$] and noise-degraded [$F_{13,13} = 10.46$, $p = 0.0002$] conditions.

C. Signal detection theory (SDT) modeling

To better understand multisensory and noise effects on CP [Fig. 2(A)], we modeled our empirical data using SDT (e.g., Rozsypal *et al.*, 1985; Braida, 1991; Getz *et al.*, 2017) [Fig. 5(A)]. We estimated different properties of each participant's psychometric functions using Bayesian inference via the *psignifit* toolbox (Schütt *et al.*, 2016). This allowed us to measure individual lapse (λ) and guess (γ) rates from their identification data. Lapse rate (λ) is computed as the difference between the upper asymptote of the psychometric function and 100% and reflects the probability of an "incorrect" response at infinitely high stimulus levels [i.e., responding "da" for Tk 7; see Fig. 2(A)]. Guess rate (γ) is defined as the difference between the lower asymptote and 0% and reflects the probability of a "correct" response at infinitely low stimulus levels [i.e., responding "ga" for Tk 1; see Fig. 2(A)]. For an ideal observer $\lambda = 0$ and $\gamma = 0$. λ and γ were measured from each participant's individual psychometric function per AV stimulus condition [see Fig. 5(A)].[5]

An ANOVA revealed that lapse rates depended on stimulus condition [$F_{3,39} = 5.42$, $p = 0.0032$; $\eta_p^2 = 0.29$] [Fig. 5(B)]. However, this effect was solely attributable to more lapses in the AV+noise compared to A+noise condition ($p = 0.0016$), paralleling the RT effect between these conditions (Fig. 4). No other pairwise comparisons differed in

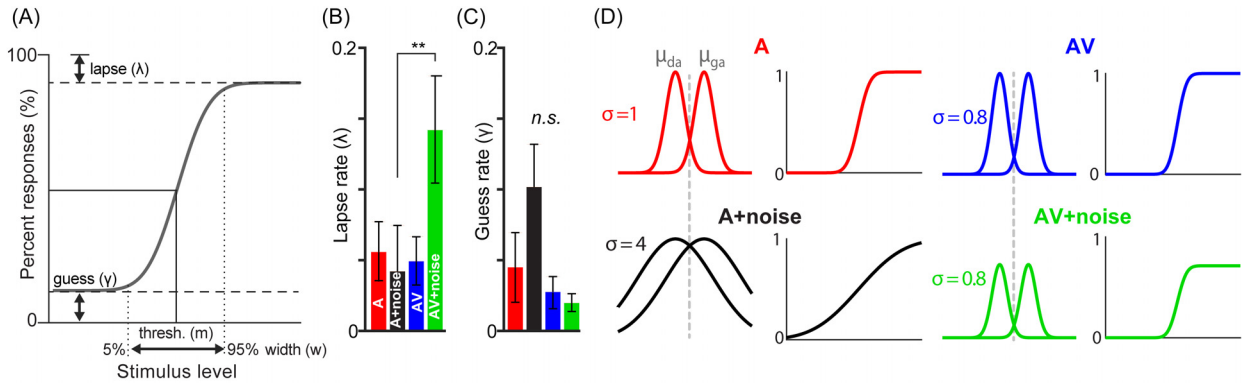

FIG. 5. (Color online) Signal detection model of AV and noise effects in speech CP. (A) Schematic identification curves illustrating definitions of lapse and guess rates from the psychometric function. (B) Lapse and (C) guess rates during CP across stimulus conditions. (D) SDT framework for understanding noise and AV effects in CP. (*left*) Observers' responses are modeled along a perceptual decision axis. In a binary classification task, the probability of responding one or another stimulus class (i.e., /da/ or /ga/) is modeled as two Gaussians. μ_{da} and μ_{ga} represent the means of the /da/ and /ga/ distributions; σ is their widths, reflecting response variability. An observer responds "ga" if the signal "energy" falls above the decision criterion (dotted line) and "da" below. (*right*) Integrating either probability curve results in a cumulative density function, modeling observers' psychometric functions. In an SDT framework, changes in the slope of observers' psychometric functions with multisensory cues and noise [Fig. 2(A)] are well modeled as changes in response variance σ^2. Reduced height of the psychometric function in the AV+noise condition can be attributed to lapses of attention (see panel B, "AV+noise"), which prevent full unity at the asymptotic end of the curve (Schütt *et al.*, 2016). errorbars = ± 1 s.e.m., $**p < 0.01$.

lapse rate ($ps > 0.09$–1.0). Similarly, guess rates were invariant across stimuli [$F_{3,39} = 1.85$, $p = 0.154$; $\eta_p^2 = 0.12$] [Fig. 5(C)]. Given that lapse and guess rate parameters were stable across stimuli, these data indicate that while visual cues and noise modulated the degree of CP for speech (Fig. 2), those effects were not driven by obvious differences in lapses of attention or guessing, *per se* (Shen and Richards, 2012; Schütt *et al.*, 2016).

In an SDT framework, observers' responses can be modeled along a perceptual decision axis where the probability of responding one or another stimulus class ("da" or "ga") follows Gaussian normal distributions of the form $y = e^{-(x-\mu)^2/2\sigma^2}$, where μ_{da} and μ_{ga} represent the means of the /da/ and /ga/ normal curves and σ, their width, reflecting variance due to the probabilistic nature of decision. An observer responds "ga" if the signal energy falls above the decision criterion (dotted line) and "da" below. The degree of categorization can then be expressed in terms of sensitivity, $d' = (\mu_{da} - \mu_{ga})/\sigma$ (Geschneider, 1997; p. 118). Theoretically, d' represents the strength of perceptual difference between /da/ and /ga/ classes. If we assume that the distance between internal representations (μ_{da}, μ_{ga}) for each phonetic class are fixed (e.g., because CV phonemes are overlearned sounds), more or less precise classification across conditions must result from changes in response variance σ^2. Integrating the Gaussian functions results in a cumulative density function, mirroring observers' psychometric functions (Fig. 5, right). Our empirical findings across AV conditions [Fig. 2(A)] are well-modeled as changes in response variance σ^2. We attribute this reduction in σ^2 as a sharpening of speech categories.

IV. DISCUSSION

A key aspect of speech comprehension is the ability to categorize variable acoustic input into discrete phonetic units. How different sensory modalities influence and potentially enhance this ability is important for understanding human speech comprehension. By measuring participants' CP for speech, we were able to assess whether the availability of additional sensory cues and the quality of the auditory signal influenced the degree of categorical speech processing. A primary finding was that noise and visual cues exerted opposite effects on CP, with noise eliciting weaker CP and visual cues eliciting stronger CP. Thus, noise appears to blur categorization whereas visual cues help compensate for the compromised auditory modality. Our results support the notion that observers integrate information from multiple sensory domains to categorize signals, especially those which lack sensory precision as in the case of noise degradation (Hélie, 2017; Bidelman *et al.*, 2019). Additionally, we extend previous findings demonstrating that visual cues enhance speech comprehension in noise (Sumby and Pollack, 1954; MacLeod and Summerfield, 1987; Vatikiotis-Bateson *et al.*, 1998; Ross *et al.*, 2007; Golumbic *et al.*, 2013; Xie *et al.*, 2014) by showing these multisensory benefits extend to the level of individual/isolated phonetic speech units and the fundamental process of CP (cf. Massaro and Cohen, 1983).

We found participants exhibited weaker CP for noisy relative to clear speech, as evidenced by flatter identification curves and slower RTs. While clear AV and A speech slowed responses to more ambiguous syllables near the midpoint of the continuum, this hallmark of CP was not elicited for the two degraded continua, implying weaker categorical processing in those latter conditions. We also found an enhancement in CP for AV compared to A speech [e.g., Fig. 3(B)]. This suggests visual cues helped lattice internalized acoustic-phonetic representations, allowing the formation of more well-formed (sharper) speech categories. Moreover, while the overall speed of access to phonetic labels (i.e., average RTs) was similar with (AV) and without (A) viseme information of the talker, visual cues reduced the overall *variability* in participants' response speeds for both clean and noise-degraded speech (Fig. 4). This reduction in decision variance is consistent with the idea that multisensory cues provide more precise access to speech categories post perceptually.

Overall, we found that AV benefits in CP were more prominent for perceptual identification (%) compared to RT data. This suggests a differential effect of multisensory cues on "early vs late" (Peelle and Sommers, 2015) or "pre- vs post-labeling" (Braida, 1991) stages of categorization. Similar dissociations in behavioral identification and response timing have been noted in previous neuroimaging studies examining speeded speech labeling tasks (Binder *et al.*, 2004). For example, in conditions with highly experienced listeners (Bidelman and Lee, 2015) or overlearned stimuli (Binder *et al.*, 2004; Chang *et al.*, 2010), early auditory cortex is sufficient to generate discrete neural representations that code discrete acoustic-phonetic categories. However, the decision process, as indexed by the speed of listeners' categorical judgments (i.e., RTs) are largely determined by activation in inferior frontal brain regions (Binder *et al.*, 2004). Other animal (Bizley and Cohen, 2013) and human studies (Bidelman and Howell, 2016; Bidelman *et al.*, 2018) have shown that functional interplay between frontal and superior temporal areas is necessary for robust speech recognition and neural coding within this network varies with signal clarity (SNR), intelligibility, and linguistic experience (Adank *et al.*, 2012; Scott and McGettigan, 2013; Bidelman and Dexter, 2015; Bidelman and Howell, 2016; Alain *et al.*, 2018; Bidelman *et al.*, 2018). While the purely behavioral nature of our data cannot adjudicate brain mechanisms, future neuroimaging experiments could address the neural underpinnings, temporal dynamics, and multisensory benefits in CP.

Noise manipulations revealed that acoustic interference weakened speech identification to the point where participants heard continua in a near continuous rather than categorical manner. This suggests a robust (perhaps expected) effect of signal clarity on the formation of categorical percepts; noise in the sensory input blurs acoustic-phonetic mapping inhibiting a strong match between the external signals and internalized memory templates. Nevertheless, while noise compromised categorical representations for speech, we found that visual cues counteract these behavioral disadvantages (at least for the moderate SNR tested here).

Interestingly, we find that the influence of visemes on auditory categorical processing is also larger when the speech signal is degraded. That is, the benefits of AV integration on CP appear stronger for degraded relative to clear speech [cf. Figs. 3(D) vs 3(B)]. Thus, in addition to providing useful groupings and perceptual constancy of sensory space (Prather *et al.*, 2009), our findings reveal another important benefit of CP: building phonetic categories (a higher-level discrete code) is more robust to noise than the physical surface features of a stimulus (lower-level sensory code) (cf. Gifford *et al.*, 2014; Hélie, 2017; Bidelman *et al.*, 2019). Consequently, our data imply that the mere process of grouping speech sounds into categories seems to aid speech comprehension in adverse listening conditions. Future studies should test whether AV effects on CP observed here for CVs also applies more broadly to CP for other speech (and non-speech) stimuli (e.g., vowels which carry more obvious visual cues).

Our empirical multisensory and noise effects on CP are parsimoniously described via concepts of SDT (e.g., Rozsypal *et al.*, 1985; Braida, 1991; Getz *et al.*, 2017). Under an SDT framework, changes to internal response variability of the observer account for the flattening of the psychometric function with additive noise (A+noise) and conversely, (re-) sharpening with visual cues (AV speech): external noise increases σ, leading to wider spread identification curves whereas visual cues reduce σ and steepen the psychometric function (see also Gifford *et al.*, 2014). The subtle reduction in height of the psychometric function in AV+noise (see Fig. 2) can be attributed to attentional lapses in this condition [see Fig. 5(B)], which prevent full unity at the asymptotic end of the identification curve (Shen and Richards, 2012; Schütt *et al.*, 2016). Although accuracy and decision speed (RT) are dissociable in categorization tasks (Binder *et al.*, 2004), our RT data further support these notions. Delayed response speeds in the AV+noise condition also suggest attentional lapses, at least in that condition [cf. Figs. 4 and 5(B)]. Indeed, RTs were correlated with attentional lapses but only for AV+noise speech ($r = 0.71$, $p = 0.0043$). Still, the fact that lapse effects were only limited to the difficult AV+noise condition—requiring parsing of speech from noise *and* the integration of phoneme and viseme information—suggests effects in the other conditions are probably not attributable to attention or post-labeling decision stages, *per se* (cf. Braida, 1991), but a sharpening (de-sharpening) of internalized categories. While SDT does not explicitly account for the speed of an observer's decision (only % identification/accuracy), the reduction in RT variability with visual cues (and increased variability with noise absent any V cues) suggests that these stimulus factors modulate observers' response variability (σ) during CP in opposite directions—a reflection of changes in the precision of the underlying perceptual object(s).

We interpret multisensory effects on CP to reflect a sharpening of internalized speech categories, consistent with notions that auditory and visual components of speech are fused into a single global percept (Green and Kuhl, 1989). Still, an alternate interpretation of our data relates to a reduction in stimulus uncertainty (Gifford *et al.*, 2014); visual cues add another source of information that might reduce uncertainty in making categorical judgments. Similarly, vision might provide an added "gain" of sensory information relative to A-only speech (Sumby and Pollack, 1954; MacLeod and Summerfield, 1987; Vatikiotis-Bateson *et al.*, 1998; Ross *et al.*, 2007; Xie *et al.*, 2014). Under this interpretation, the observed sharpening of the CP boundary (Fig. 2) might be explained by a reduction in uncertainty due to AV facilitation.

In this vein, comparing AV to AV+noise responses for Tk 4 revealed that classification was more reliably "ga" in the presence of noise. This implies a Bayesian-like integration (Deneve and Pouget, 2004). When noise is absent, A cues still dominate the behavioral decision. However, when noise is added, the reliability of A cues is severely diminished and so V cues take over as they offer more reliable inference on which to form AV speech percepts (Bidelman *et al.*, in press). Indeed, studies have shown that in situations where visual cues are deemed unreliable (e.g., noise, sensory impairments), sound can trump vision to maintain robust perception (Alais and Burr, 2004; Narinesingh *et al.*, 2015; Myers *et al.*, 2017). Similar cue (re)weighting has been observed with analogous degradation to the auditory channel (Bidelman *et al.*, 2019a).

Still, evidence against a strict cue weighting explanation is the fact that RTs for clean A and AV speech were identical (Fig. 4), in contrast to the faster response speeds that would be expected by multisensory facilitation (e.g., RACE and redundant signal models; Miller and Ulrich, 2003; Colonius and Diederich, 2006). Similarly, the slower RTs for AV+noise vs A+noise speech (Fig. 4) also runs counter to a strict RACE framework, where multisensory speech would be expected to facilitate response speeds, even in noise. These data reveal a functional distinction in multisensory processing for isolated speech categorization (present study) that is perhaps fundamentally distinct from sentence-level recognition, where semantic context, lexical frequency, and other indexical cues can aid speech-in-noise perception (e.g., Boothroyd and Nittrouer, 1988; Helfer and Freyman, 2009). Instead, we attribute the overall slower RTs for AV+noise to reflect cognitive interference in attempting to reconcile the minimal categorical cues supplied by our CV visemes with the more salient ones from the auditory-phonetic input (Files *et al.*, 2015) coupled with the inherent listening effort associated with degraded-speech perception tasks (Picou *et al.*, 2016; Bidelman and Yellamsetty, 2017; Bidelman *et al.*, in press). Under this interpretation, the relative invariance in CP slope but counterintuitive slower RT in AV+noise could reflect a dissociable effect in pre- vs post-labeling aspects of CP (Braida, 1991); V cues help lattice (i.e., sharpen) the categorical object at a pre-perceptual stage resulting in robust identification [Fig. 2(A)] but noise impairs the speed of access to this representation post-perceptually, as reflect in the delayed RTs (Fig. 4).

The notion of AV integration itself has been questioned since animal work has shown visual stimuli modulate cortical responses in auditory cortical fields independently of visual stimulus category (Kayser *et al.*, 2008). Calvert *et al.* (1997) even suggested that activation of primary auditory

cortex during lip reading implies visual cues influence perception even before speech sounds are categorized into phonemes (for review, see Bernstein and Liebenthal, 2014). Visual cues precede the corresponding auditory signal, so they may also serve a predictive role in facilitating speech processing, particularly in noisy situations (Golumbic *et al.*, 2013). Unfortunately, behavior alone cannot delineate accounts of our data based on reduction in stimulus uncertainly due to additional (AV) information or true sharpening of the category—although these explanations need not be mutually exclusive. Ongoing neuroimaging experiments are currently underway in our laboratory to adjudicate these competing mechanisms (e.g., Bidelman *et al.*, 2019).

V. CONCLUSIONS

In sum, our findings support a view of multimodal integration in which observers use available cues from multiple sensory modalities to perceptually categorize speech, especially when the acoustic signal is impoverished. While the present study was not designed to provide a mechanistic account of how visual and auditory cues may combine or interact in CP, our results provide important evidence that the perceptual and categorical organization of speech is not a unimodal process. Visual cues reduce the precision of categorical representations leading to sharper phonetic identification whereas noise exerts the opposite pattern, increasing variability in CP and leading to less precise speech categories. Though controversial, dyslexia has been linked to poorer CP (Messaoud-Galusi *et al.*, 2011; Noordenbos and Serniclaes, 2015; Hakvoort *et al.*, 2016; Zoubrinetzky *et al.*, 2016)—a deficit which may be exacerbated in noise (Calcus *et al.*, 2016). Future work could focus on identifying possible contributions that information from visual and additional sensory modalities make to auditory speech comprehension in normal and disordered populations and in other complex perceptual scenarios (e.g., reverberation, visual noise interference).

ACKNOWLEDGMENTS

This work was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under award number NIH/NIDCD R01DC016267 (G.M.B.).

[1]For more information, see https://openwetware.org/wiki/Beauchamp:Stimuli

[2]The notion of the viseme as a unitary perceptual category has been recently questioned (Files *et al.*, 2015). There is also longstanding debate that visual speech is too impoverished to convey much phonetic information (Kuhl and Meltzoff, 1988). Consequently, the categorical morphing of visual speech, if even technically feasible, is likely ill-posed.

[3]We did not pair each video with all seven speech tokens because our goal was to assess the strength of AV benefits during CP. Exposure to incongruent cues (e.g., visual "da" with steps of auditory "ga") could result in cross-modal bias (Bertelson *et al.*, 2003) or McGurk-like AV illusions (McGurk and MacDonald, 1976), which would confound interpretation in this study.

[4]Available at https://www.mathworks.com/matlabcentral/fileexchange/22441-curve-intersections

[5]Our definitions of lapse (λ) and guess (γ) rates are based on conventional psychometric function analysis from *detection* experiments (i.e., %-correct vs increasing stimulus intensity) (Shen and Richards, 2012; Schütt *et al.*, 2016). However, for the application to categorical identification functions, which are mirror images for a binary response task, λ and γ could be described as common guessing terms at each end of the stimulus continuum. In that case, a single parameter (i.e., $\lambda = \gamma$) could be used to produce non-ceiling/floor responses at the extremes of the continuum and constrain the asymptotes (see Fig. 5A) to be symmetric in model fitting. Across AV conditions, model fits with independent λ and γ (asymmetric asymptotes) vs $\lambda = \gamma$ (symmetric asymptotes) showed better agreement with our empirical data (two-parameter$_{\text{deviance}} = 6.59 \pm 1.40$ vs one-parameter$_{\text{deviance}} = 8.68 \pm 2.35$; $t_{13} = -2.94$, $p = 0.0115$). Thus, we modeled psychometric functions with independent asymptotes ($\lambda \neq \gamma$), which is further supported by the asymmetry observed visually in our identification data (see Fig. 2A).

Adank, P., Davis, M. H., and Hagoort, P. (**2012**). "Neural dissociation in processing noise and accent in spoken language comprehension," Neuropsychologia **50**, 77–84.

Alain, C., Du, Y., Bernstein, L. J., Barten, T., and Banai, K. (**2018**). "Listening under difficult conditions: An activation likelihood estimation meta-analysis," Hum. Brain Mapp. **39**, 2695–2709.

Alais, D., and Burr, D. (**2004**). "The ventriloquist effect results from near-optimal bimodal integration," Curr. Biol. **14**, 257–262.

Altmann, C. F., Uesaki, M., Ono, K., Matsuhashi, M., Mima, T., and Fukuyama, H. (**2014**). "Categorical speech perception during active discrimination of consonants and vowels," Neuropsychologia **64C**, 13–23.

Beale, J. M., and Keil, F. C. (**1995**). "Categorical effects in the perception of faces," Cognition **57**, 217–239.

Bejjanki, V. R., Clayards, M., Knill, D. C., and Aslin, R. N. (**2011**). "Cue integration in categorical tasks: Insights from audio-visual speech perception," PloS One **6**, e19812.

Benjamini, Y., and Hochberg, Y. (**1995**). "Controlling the false discovery rate: A practical and powerful approach to multiple testing," J. R. Stat. Soc. Ser. B (Methodol.) **57**, 289–300, available at http://www.jstor.org/stable/2346101.

Bernstein, L. E., Auer, E. T., Jr., and Takayanagi, S. (**2004**). "Auditory speech detection in noise enhanced by lipreading," Speech Commun. **44**, 5–18.

Bernstein, L. J., Catton, P. A., and Tannock, I. F. (**2014**). "Intra-individual variability in women with breast cancer," J. Int. Neuropsychol. Soc. **20**, 380–390.

Bernstein, L. E., and Liebenthal, E. (**2014**). "Neural pathways for visual speech perception," Front. Neurosci. **8**, 1–18.

Bertelson, P., Vroomen, J., and De Gelder, B. (**2003**). "Visual recalibration of auditory speech identification: A McGurk aftereffect," Psych. Sci. **14**, 592–597.

Bidelman, G. M., Brown, B., Mankel, K., and Price, C. N. (**2019a**). "Psychobiological responses reveal audiovisual noise differentially challenges speech recognition," Ear Hear. (in press).

Bidelman, G. M., Bush, L. C., and Boudreaux, A. M. (**2019b**). "The categorical neural organization of speech aids its perception in noise," bioRxiv (published online).

Bidelman, G. M., Davis, M. K., and Pridgen, M. H. (**2018**). "Brainstem-cortical functional connectivity for speech is differentially challenged by noise and reverberation," Hear. Res. **367**, 149–160.

Bidelman, G. M., and Dexter, L. (**2015**). "Bilinguals at the 'cocktail party': Dissociable neural activity in auditory-linguistic brain regions reveals neurobiological basis for nonnative listeners' speech-in-noise recognition deficits," Brain Lang. **143**, 32–41.

Bidelman, G. M., and Howell, M. (**2016**). "Functional changes in inter- and intra-hemispheric auditory cortical processing underlying degraded speech perception," Neuroimage **124**, 581–590.

Bidelman, G. M., and Lee, C.-C. (**2015**). "Effects of language experience and stimulus context on the neural organization and categorical perception of speech," Neuroimage **120**, 191–200.

Bidelman, G. M., Lowther, J. E., Tak, S. H., and Alain, C. (**2017**). "Mild cognitive impairment is characterized by deficient hierarchical speech coding between auditory brainstem and cortex," J. Neurosci. **37**, 3610–3620.

Bidelman, G. M., Moreno, S., and Alain, C. (**2013**). "Tracing the emergence of categorical speech perception in the human auditory system," Neuroimage **79**, 201–212.

Bidelman, G. M., and Walker, B. (**2017**). "Attentional modulation and domain specificity underlying the neural organization of auditory categorical perception," Eur. J. Neurosci. **45**, 690–699.
Bidelman, G. M., Weiss, M. W., Moreno, S., and Alain, C. (**2014**). "Coordinated plasticity in brainstem and auditory cortex contributes to enhanced categorical speech perception in musicians," Eur. J. Neurosci. **40**, 2662–2673.
Bidelman, G. M., and Yellamsetty, A. (**2017**). "Noise and pitch interact during the cortical segregation of concurrent speech," Hear. Res. **351**, 34–44.
Binder, J. R., Liebenthal, E., Possing, E. T., Medler, D. A., and Ward, B. D. (**2004**). "Neural correlates of sensory and decision processes in auditory object identification," Nat. Neurosci. **7**, 295–301.
Bizley, J. K., and Cohen, Y. E. (**2013**). "The what, where and how of auditory-object perception," Nat. Rev. Neurosci. **14**, 693–707.
Bizley, J. K., Maddox, R. K., and Lee, A. K. C. (**2016**). "Defining auditory-visual objects: Behavioral tests and physiological mechanisms," Trends Neurosci. **39**, 74–85.
Boothroyd, A., and Nittrouer, S. (**1988**). "Mathematical treatment of context effects in phoneme and word recognition," J. Acoust. Soc. Am. **84**, 101–114.
Bradley, J. V. (**1958**). "Complete counterbalancing of immediate sequential effects in a Latin square design.," J. Amer. Statist. Assoc. **53**, 525–528.
Braida, L. D. (**1991**). "Crossmodal integration in the identification of consonant segments," Q. J. Exp. Psychol. A: Human Exp. Psychol. **43**, 647–677.
Calcus, A., Lorenzi, C., Collet, G., Colin, C., and Kolinsky, R. (**2016**). "Is there a relationship between speech identification in noise and categorical perception in children with dyslexia?," J. Speech. Lang. Hear. Res. **59**, 835–852.
Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., Woodruff, P. W., Iversen, S. D., and David, A. S. (**1997**). "Activation of auditory cortex during silent lipreading," Science **276**, 593–596.
Carlyon, R. P., Cusack, R., Foxton, J. M., and Robertson, I. H. (**2001**). "Effects of attention and unilateral neglect on auditory stream segregation," J. Exp. Psychol.: Human Percept. Perform. **27**, 115.
Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., and Knight, R. T. (**2010**). "Categorical speech representation in human superior temporal gyrus," Nat. Neurosci. **13**, 1428–1432.
Colonius, H., and Diederich, A. (**2006**). "The race model inequality: Interpreting a geometric measure of the amount of violation," Psychol. Rev. **113**, 148–154.
Denes, P. B. (**1963**). "On the statistics of spoken English," J. Acoust. Soc. Am. **35**, 892–904.
Deneve, S., and Pouget, A. (**2004**). "Bayesian multisensory integration and cross-modal spatial links," J. Physiol.-Paris **98**, 249–258.
Desai, S., Stickney, G., and Zeng, F.-G. (**2008**). "Auditory-visual speech perception in normal-hearing and cochlear-implant listeners," J. Acoust. Soc. Am. **123**, 428–440.
Files, B. T., Tjan, B. S., Jiang, J., and Bernstein, L. E. (**2015**). "Visual speech discrimination and identification of natural and synthetic consonant stimuli," Front. Psychol. **6**, 878.
Franklin, A., Drivonikou, G. V., Clifford, A., Kay, P., Regier, T., and Davies, I. R. (**2008**). "Lateralization of categorical perception of color changes with color term acquisition," Proc. Natl. Acad. Sci. U.S.A. **105**, 18221–18225.
Geschneider, G. A. (**1997**). *Psychophysics: The Fundamentals* (Lawrence Erlbaum Associates, Inc., Mahwah, NJ).
Getz, L., Nordeen, E., Vrabic, S., and Toscano, J. (**2017**). "Modeling the development of audiovisual cue integration in speech perception," Brain Sci. **7**, 32.
Gifford, A. M., Cohen, Y. E., and Stocker, A. A. (**2014**). "Characterizing the impact of category uncertainty on human auditory categorization behavior," PLoS Comput. Biol. **10**, e1003715.
Golumbic, E. Z., Cogan, G. B., Schroeder, C. E., and Poeppel, D. (**2013**). "Visual input enhances selective speech envelope tracking in auditory cortex at a 'cocktail party'," J. Neurosci. **33**, 1417–1426.
Grant, K. (**2001**). "The effect of speechreading on masked detection thresholds for filtered speech," J. Acoust. Soc. Am. **109**, 2272–2275.
Grant, K., and Bernstein, J. (**2019**). "Toward a Model of Auditory-Visual Speech Intelligibility: The Auditory Perspective," in *Springer Handbook of Auditory Research: Multisensory Processes: The Auditory Perspective*, edited by A. K. C. Lee, M. Wallace, A. B. Coffin, A. N. Popper, and R. R. Fay (Springer-Nature, Switzerland), pp. 33–57.
Grant, K. W., and Seitz, P.-F. (**2000**). "The use of visible speech cues for improving auditory detection of spoken sentences," J. Acoust. Soc. Am. **108**, 1197–1208.
Green, K. P., and Kuhl, P. K. (**1989**). "The role of visual information in the processing of place and manner features in speech perception," Percept. Psychophys. **45**, 34–42.
Hakvoort, B., de Bree, E., van der Leij, A., Maassen, B., van Setten, E., Maurits, N., and van Zuijen, T. L. (**2016**). "The role of categorical speech perception and phonological processing in familial risk children with and without dyslexia," J. Speech. Lang. Hear. Res. **59**, 1448–1460.
Harnad, S. (**1987a**). "Psychophysical and cognitive aspects of categorical perception: A critical overview," in *Categorical Perception: The Groundwork of Cognition* (Cambridge University Press, Cambridge, UK), pp. 1–52.
Harnad, S. R. (**1987b**). *Categorical Perception: The Groundwork of Cognition* (Cambridge University Press, Cambridge, UK).
Helfer, K. S., and Freyman, R. L. (**2009**). "Lexical and indexical cues in masking by competing speech," J. Acoust. Soc. Am. **125**, 447–456.
Helie, S. (**2017**). "The effect of integration masking on visual processing in perceptual categorization," Brain Cogn. **116**, 63–70.
Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., TIrino, T., and Banno, H. (**2008**). "Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 30–April 4, Las Vegas, NV, pp. 3933–3936.
Kayser, C., Petkov, C. I., and Logothetis, N. K. (**2008**). "Visual modulation of neurons in auditory cortex," Cereb. Cortex **18**, 1560–1574.
Kewley-Port, D., Watson, C. S., and Foyle, D. C. (**1988**). "Auditory temporal acuity in relation to category boundaries: Speech and nonspeech stimuli," J. Acoust. Soc. Am. **83**, 1133–1145.
Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., and Banerjee, S. (**2004**). "Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **116**, 2395–2405.
Kuhl, P. K., and Meltzoff, A. N. (**1988**). "Speech as an intermodal object of perception," in *Perceptual Development in Infancy*, edited by A. Yonas (Lawrence Erlbaum Associates, Inc., Hillsdale, NJ), pp. 235–266.
Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (**1967**). "Perception of the speech code," Psychol. Rev. **74**, 431–461.
Liberman, A. M., and Mattingly, I. G. (**1989**). "A specialization for speech perception," Science **243**, 489–494.
MacLeod, A., and Summerfield, Q. (**1987**). "Quantifying the contribution of vision to speech perception in noise," Br. J. Audiol. **21**, 131–141.
Mallick, D. B., Magnotti, J. F., and Beauchamp, M. S. (**2015**). "Variability and stability in the McGurk effect: Contributions of participants, stimuli, time, and response type," Psychon. Bull. Rev. **22**, 1299–1307.
Mankel, K., and Bidelman, G. M. (**2018**). "Inherent auditory skills rather than formal music training shape the neural encoding of speech," Proc. Natl. Acad. Sci. U.S.A. **115**, 13129–13134.
Massaro, D. W., and Cohen, M. M. (**1983**). "Evaluation and integration of visual and auditory information in speech perception," J. Exp. Psychol. Hum. Percept. Perform. **9**, 753–771.
McGurk, H., and MacDonald, J. (**1976**). "Hearing lips and seeing voices," Nature **264**, 746–748.
Messaoud-Galusi, S., Hazan, V., and Rosen, S. (**2011**). "Investigating speech perception in children with dyslexia: Is there evidence of a consistent deficit in individuals?," J. Speech. Lang. Hear. Res. **54**, 1682–1701.
Miller, J. L., Grosjean, F., and Lomanto, C. (**1984**). "Articulation rate and its variability in spontaneous speech: A reanalysis and some implications," Phonetica **41**, 215–225.
Miller, J., and Ulrich, R. (**2003**). "Simple reaction time and statistical facilitation: A parallel grains model," Cogn Psychol **46**, 101–151.
Moradi, S., Lidestam, B., Danielsson, H., Ng, E. H. N., and Ronnberg, J. (**2017**). "Visual cues contribute differentially to audiovisual perception of consonants and vowels in improving recognition and reducing cognitive demands in listeners with hearing impairment using hearing aids," J. Speech. Lang. Hear. Res. **60**, 2687–2703.
Myers, M. H., Iannaccone, A., and Bidelman, G. M. (**2017**). "A pilot investigation of audiovisual processing and multisensory integration in patients with inherited retinal dystrophies," BMC Ophthalmol. **17**, 1–13.
Narinesingh, C., Goltz, H. C., Raashid, R. A., and Wong, A. M. (**2015**). "Developmental trajectory of McGurk effect susceptibility in children and adults with amblyopia," Invest. Ophthalmol. Vis. Sci. **56**, 2107–2113.

Nath, A. R., and Beauchamp, M. S. (**2012**). "A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion," Neuroimage **59**, 781–787.
Noordenbos, M. W., and Serniclaes, W. (**2015**). "The categorical perception deficit in dyslexia: A meta-analysis," Sci. Studies Read. **19**, 340–359.
Nothdurft, H. C. (**1991**). "Texture segmentation and pop-out from orientation contrast," Vision Res. **31**, 1073–1078.
O'Sullivan, A. E., Crosse, M. J., Di Liberto, G. M., and Lalor, E. C. (**2017**). "Visual cortical entrainment to motion and categorical speech features during silent lipreading," Front. Human Neurosci. **10**, 679.
Oldfield, R. C. (**1971**). "The assessment and analysis of handedness: The Edinburgh inventory," Neuropsychologia **9**, 97–113.
Parbery-Clark, A., Skoe, E., Lam, C., and Kraus, N. (**2009**). "Musician enhancement for speech-in-noise," Ear Hear. **30**, 653–661.
Peelle, J. E., and Davis, M. H. (**2012**). "Neural oscillations carry speech rhythm through to comprehension," Front. Psychol. **3**, 320.
Peelle, J. E., and Sommers, M. S. (**2015**). "Prediction and constraint in audiovisual speech perception," Cortex **68**, 169–181.
Perez-Gay, F., Sicotte, T., Theriault, C., and Harnad, S. (**2018**). "Category learning can alter perception and its neural correlate," arXiv:1805.04619 (published online).
Picou, E. M., Gordon, J., and Ricketts, T. A. (**2016**). "The effects of noise and reverberation on listening effort for adults with normal hearing," Ear Hear. **37**, 1–13.
Pisoni, D. B. (**1973**). "Auditory and phonetic memory codes in the discrimination of consonants and vowels," Percept. Psychophys. **13**, 253–260.
Pisoni, D. B., and Luce, P. A. (**1987**). "Acoustic-phonetic representations in word recognition," Cognition **25**, 21–52.
Pisoni, D. B., and Tash, J. (**1974**). "Reaction times to comparisons within and across phonetic categories," Percept. Psychophys. **15**, 285–290.
Prather, J. F., Nowicki, S., Anderson, R. C., Peters, S., and Mooney, R. (**2009**). "Neural correlates of categorical perception in learned vocal communication," Nat. Neurosci. **12**, 221–228.
Reetzke, R., Lam, B. P. W., Xie, Z., Sheng, L., and Chandrasekaran, B. (**2016**). "Effect of simultaneous bilingualism on speech intelligibility across different masker types, modalities, and signal-to-noise ratios in school-age children," PloS One **11**, e0168048.
Reetzke, R., Xie, Z., Llanos, F., and Chandrasekaran, B. (**2018**). "Tracing the trajectory of sensory plasticity across different stages of speech learning in adulthood," Curr. Biol. **28**, 1419–1427.
Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., and Foxe, J. J. (**2007**). "Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments," Cereb. Cortex **17**, 1147–1153.
Rozsypal, A. J., Stevenson, D. C., and Hogan, J. T. (**1985**). "Dispersion in models of categorical perception," J. Math. Psychol. **29**, 271–288.
Schorr, E. A., Fox, N. A., van Wassenhove, V., and Knudsen, E. I. (**2005**). "Auditory-visual fusion in speech perception in children with cochlear implants," Proc. Natl. Acad. Sci. U.S.A. **102**, 18748–18750.
Schütt, H. H., Harmeling, S., Macke, J. H., and Wichmann, F. A. (**2016**). "Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data," Vision Res. **122**, 105–123.
Schwartz, J.-L., Berthommier, F., and Savariaux, C. (**2004**). "Seeing to hear better: Evidence for early audio-visual interactions in speech identification," Cognition **93**, B69–B78.
Scott, S. K., and McGettigan, C. (**2013**). "The neural processing of masked speech," Hear. Res. **303**, 58–66.
Shen, Y., and Richards, V. M. (**2012**). "A maximum-likelihood procedure for estimating psychometric functions: Thresholds, slopes, and lapses of attention," J. Acoust. Soc. Am. **132**, 957–967.
Smayda, K. E., Chandrasekaran, B., and Maddox, W. T. (**2015**). "Enhanced cognitive and perceptual processing: A computational basis for the musician advantage in speech learning," Front. Psychol. **6**, 682.
Strauss, E., MacDonald, S. W., Hunter, M., Moll, A., and Hultsch, D. F. (**2002**). "Intraindividual variability in cognitive performance in three groups of older adults: Cross-domain links to physical status and self-perceived affect and beliefs," J. Int. Neuropsychol. Soc. **8**, 893–906.
Sumby, W. H., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**, 212–215.
van Wassenhove, V., Grant, K. W., and Poeppel, D. (**2005**). "Visual speech speeds up the neural processing of auditory speech," PNAS **102**, 1181–1186.
Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S., and Munhall, K. G. (**1998**). "Eye movement of perceivers during audiovisual speech perception," Percept. Psychophys. **60**, 926–940.
Walden, B. E., Montgomery, A. A., Prosek, R. A., and Hawkins, D. B. (**1990**). "Visual biasing of normal and impaired auditory speech perception," J. Speech Lang. Hear. Res. **33**, 163–173.
Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., and Jones, C. J. (**1977**). "Effects of training on the visual recognition of consonants," J. Speech. Lang. Hear. Res. **20**, 130–145.
Weinholtz, C., and Dias, J. W. (**2016**). "Categorical perception of visual speech information," J. Acoust. Soc. Am. **139**, 2018–2018.
Xie, Z., Yi, H.-G., and Chandrasekaran, B. (**2014**). "Nonnative audiovisual speech perception in noise: Dissociable effects of the speaker and listener," PloS One **9**, e114439.
Xu, Y., Gandour, J. T., and Francis, A. (**2006**). "Effects of language experience and stimulus complexity on the categorical perception of pitch direction," J. Acoust. Soc. Am. **120**, 1063–1074.
Zoubrinetzky, R., Collet, G., Serniclaes, W., Nguyen-Morel, M.-A., and Valdois, S. (**2016**). "Relationships between categorical perception of phonemes, phoneme awareness, and visual attention span in developmental dyslexia," PLoS One **11**, e0151015.