Vision perceptually restores auditory spectral dynamics in speech

John Plass[a,b,1], David Brang[a], Satoru Suzuki[b,c], and Marcia Grabowecky[b,c]

[a]Department of Psychology, University of Michigan, Ann Arbor, MI 48109; [b]Department of Psychology, Northwestern University, Evanston, IL 60208; and [c]Interdepartmental Neuroscience Program, Northwestern University, Chicago, IL 60611

Edited by Dale Purves, Duke University, Durham, NC, and approved June 3, 2020 (received for review February 17, 2020)

Visual speech facilitates auditory speech perception, but the visual cues responsible for these benefits and the information they provide remain unclear. Low-level models emphasize basic temporal cues provided by mouth movements, but these impoverished signals may not fully account for the richness of auditory information provided by visual speech. High-level models posit interactions among abstract categorical (i.e., phonemes/visemes) or amodal (e.g., articulatory) speech representations, but require lossy remapping of speech signals onto abstracted representations. Because visible articulators shape the spectral content of speech, we hypothesized that the perceptual system might exploit natural correlations between midlevel visual (oral deformations) and auditory speech features (frequency modulations) to extract detailed spectrotemporal information from visual speech without employing high-level abstractions. Consistent with this hypothesis, we found that the time–frequency dynamics of oral resonances (formants) could be predicted with unexpectedly high precision from the changing shape of the mouth during speech. When isolated from other speech cues, speech-based shape deformations improved perceptual sensitivity for corresponding frequency modulations, suggesting that listeners could exploit this cross-modal correspondence to facilitate perception. To test whether this type of correspondence could improve speech comprehension, we selectively degraded the spectral or temporal dimensions of auditory sentence spectrograms to assess how well visual speech facilitated comprehension under each degradation condition. Visual speech produced drastically larger enhancements during spectral degradation, suggesting a condition-specific facilitation effect driven by cross-modal recovery of auditory speech spectra. The perceptual system may therefore use audiovisual correlations rooted in oral acoustics to extract detailed spectrotemporal information from visual speech.

audiovisual speech | speech perception | multisensory | spectrotemporal

Reliable speech perception is essential to human communication and plays a critical role in social, vocational, and emotional health. While speech is often thought of as being conveyed primarily through auditory signals, visual speech signals can also play a significant role in forming reliable speech percepts, especially when auditory speech is degraded by environmental noise (e.g., ref. 1) or hearing impairments associated with clinical disorders or aging (2, 3). However, the mechanisms underlying visual contributions to speech perception are still poorly understood.

Converging evidence suggests that visual speech can enhance the detection (4), comprehension (1, 5), and neural encoding (6) of auditory speech, but the audiovisual correspondences underlying these enhancements remain unclear. Timing cues provided by lip movements can facilitate auditory speech perception by entraining cortical oscillations to amplitude modulations in speech (7–10), but it is unclear whether and how more complex attributes of speech signals might be derived from visual speech to facilitate perception.

Early models of audiovisual speech perception emphasized higher-level representations of abstracted speech units, such as phonemes and "visemes," categorical representations of acoustic features and articulatory gestures associated with particular speech sounds. However, converging evidence suggests that phoneme–viseme correspondences are insufficient to account for the perceptually detailed subphonemic information provided by visual speech (11–13), suggesting that visual speech may also convey more fine-grained acoustic details about auditory speech signals (14).

Alternative models have posited that amodal representations of speech, such as motor or articulatory representations, could provide a common representational basis for auditory and visual speech signals (15). However, neurophysiology and neuroimaging studies suggest that potentially relevant motor and premotor speech representations are not sufficiently detailed to account for observed audiovisual effects. Intracranial ECoG recordings and multivariate fMRI suggest that areas involved in speech production do not encode heard speech according to corresponding articulatory features but, rather, an impoverished representation of acoustic-phonetic features (16, 17). For example, while the syllables /ba/, /da/, and /ga/ could be readily distinguished based on motor cortex activity (ECoG high gamma) during speech production, they produced indistinguishable response patterns during speech listening (16). Because these syllables are clearly distinguished in audiovisual speech perception (e.g., ref. 18), these representations are unlikely to play a role in observed audiovisual interactions. Moreover, the temporal dynamics of neural coupling

Significance

Multisensory signals can facilitate perception by clarifying unreliable unisensory signals. These multisensory interactions are particularly apparent in audiovisual speech perception, in which visual speech substantially enhances auditory speech processes, remediating perceptual deficits produced by noisy environments, hearing disorders, or age-related hearing loss. However, the audiovisual cues and integration mechanisms responsible for these effects remain unclear. Here, we demonstrate that the changing frequencies of oral resonances—which are used to discriminate between speech sounds—can be predicted from mouth shapes during speech, and that listeners exploit this relationship to extract acoustic information from visual speech. These results suggest that the perceptual system uses natural correlations between midlevel visual (oral deformations) and auditory speech features (frequency modulations) to facilitate speech perception.

Author contributions: J.P., D.B., S.S., and M.G. designed research; J.P. performed research; J.P. analyzed data; and J.P., D.B., S.S., and M.G. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Data deposition: Data from the present study have been deposited at Open Science Framework (https://osf.io/mk7c9/).

[1]To whom correspondence may be addressed. Email: jplass@umich.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2002887117/-/DCSupplemental.

between auditory and motor cortices suggest that motor regions may not encode speech signals with sufficient temporal resolution to capture acoustic-phonetic details. Phase-coupling between auditory and motor cortices is limited to a tightly restricted range of modulation frequencies (~4.5 Hz) within the range of syllabic rhythms in speech (4 to 7 Hz; ref. 19), suggesting that speech motor regions are unlikely to encode more rapidly occurring (20 to 50 Hz; ref. 20) acoustic-phonetic features.

We hypothesized that the perceptual system may extract acoustic details from visual speech by exploiting natural physical relationships between visible articulators (e.g., mouth shapes) and associated vocal resonances (formants). During speech, the changing configuration of orofacial articulators alters the resonant frequencies of resonant cavities within the vocal tract (21–24). Therefore, to the extent that these articulators are visible, they may provide cues to the time-varying spectral content of speech. Such cross-modal correspondences could allow signals conveying fine acoustic-phonetic detail to be recovered from visual speech, potentially enhancing acoustic-phonetic representations of speech at a subphonemic level to facilitate perception.

Results

Statistical Relationship between Lip Deformations and Speech Spectra. To uncover potential relationships between visible articulations and spectrotemporal speech signals, we analyzed the spectral content of auditory syllables produced with visually salient deformations of the lips. Because variations in the shape and position of the lips contribute directly to the resonant properties of the oral cavity (21–23), we hypothesized that changes in the shape of the oral aperture would be reflected in the frequency of corresponding resonant peaks (formants) in the auditory spectrogram. To test this hypothesis, we audio-recorded and filmed four volunteers (two female) as they expanded and contracted their lips horizontally and vertically to produce the voiced syllables wiː (English: "we"), juː (English: "ew," informal expression of disgust), wa (as in English "want"), and aʊ (as in English "hour").

For all four volunteers, horizontal expansion and contraction of the lips produced corresponding increases and decreases in the frequency of a single resonant peak, producing sigmoidal formant contours (i.e., frequency sweeps) spanning ~500 to 3,000 Hz (Fig. 1 *A*, *Upper*). However, only weak or negligible formant frequency modulations were observed with vertical expansions and contractions (Fig. 1 *A*, *Lower*). These results are consistent with previous suggestions that the spreading (retraction) and rounding (protrusion) of the lips may alter the resonant frequency of the "front" oral cavity by shortening and lengthening the distance between the oral aperture and any intraoral constrictions (21, 23, 24).

To assess the extent to which the frequency of this resonance could be inferred from visual speech, we computed linear correlations between the width of the oral aperture and the frequency of this resonance at each frame refresh throughout the normal pronunciation of each syllable. Fig. 1*B* shows the linearly transformed mouth widths overlaid on the time–frequency trajectories of the observed resonance. As evidenced by the plots, the width of the oral aperture predicted the frequency of the resonant peak with surprising precision (all $r > 0.95$, all $P < 0.001$), suggesting that mouth width may provide a precise and visually salient cue to the time-varying spectral content of speech.

To verify that the observed spectrotemporal modulations were indeed attributable to changes in the resonant frequency of the oral cavity, we placed a small speaker cone ~3 mm above the tip of two speakers' (one female) tongues and recorded the spectral modulations produced by their lip movements as the speaker cone emitted synthesized glottal pulses (100 Hz). Horizontal lip movements produced frequency sweeps like those observed in natural speech, with narrow oral apertures producing lower-frequency spectral speaks and wide oral apertures producing higher-frequency spectral peaks (Fig. 1*C*).

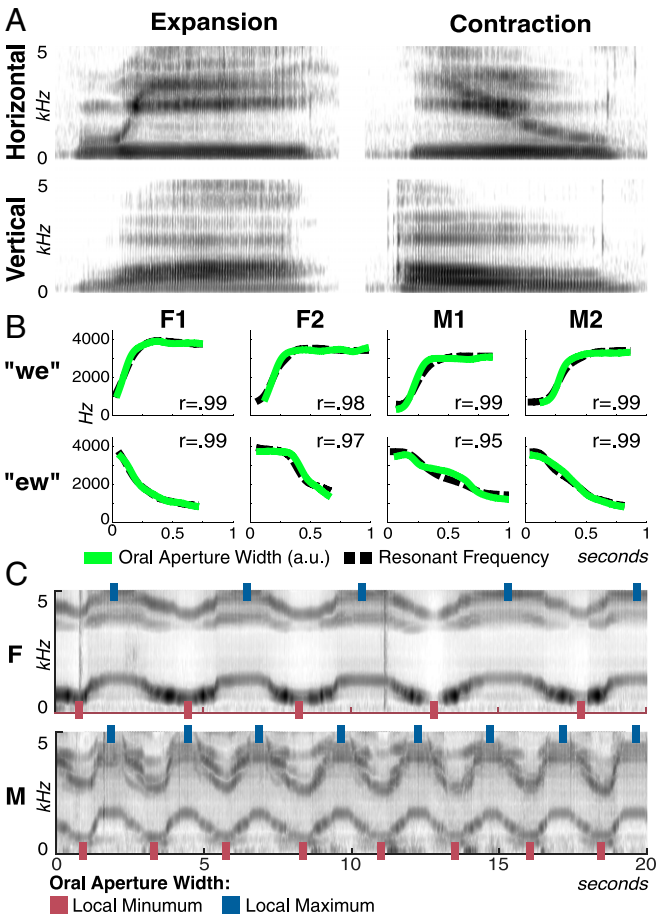

Fig. 1. Statistical relationship between lip deformations and speech spectra. (*A*) Four participants were recorded as they produced four voiced phonemes by expanding or contracting their lips horizontally or vertically. Resultant spectrograms for one male speaker (M2). Horizontal expansion and contraction (*Top*; wiː and juː) produced opposing rising and falling frequency sweeps (formant contours) spanning ~500 to 3,000 Hz. Vertical expansion and contraction (*Bottom*; wa and aʊ) produced negligible modulations of formant frequencies. (*B*) To assess the relationship between lip width and the observed formant frequencies, we calculated the linear correlation between the two variables as participants expanded and contracted their lips horizontally, producing the phonemes wiː (English "we") and juː (English: "ew"). Linearly transformed lip widths (black dashed curves) overlaid with corresponding resonant peaks (green curves; both curves spline-interpolated for display). (*C*) To verify that the observed frequency sweeps resulted from changes in oral resonances produced by lip movements, we recorded the sounds produced by horizontal lip deformations when we replaced the glottal source with artificial glottal pulses played from a small speaker placed in the mouth. Horizontal lip movements produced frequency modulations similar to those produced in natural speech, with narrow oral apertures producing lower frequencies and wide oral apertures producing higher frequencies.

Finally, to verify that the observed correlation generalized across other phonemes, we computed correlations between the average width of the oral aperture and the average frequency of the dominant spectral peak between 500 and 3,000 Hz during the production of 14 English vowels by two speakers (one female) (25). As shown in Fig. 2*A*, substantial linear correlations were observed for both speakers (M, $r = 0.90$, $P < 0.001$; F, $r = 0.76$, $P < 0.005$), suggesting that mouth width may provide a useful cue to speech spectra associated with multiple phonemes.

To further examine the relationship between lip width and resonant frequency across vowels, we performed two additional

analyses. First, as shown in Fig. 2*B*, we computed pairwise distances between *z*-transformed lip widths (rows) and resonant frequencies (columns) to assess how well vowels could be discriminated visually given the relationship between lip width and resonant frequency. In this analysis, the discriminability of each vowel's visual-to-auditory distance profile (i.e., the rows in Fig. 2*B*) indicates how well lip widths can discriminate between vowel-specific resonant frequencies. As can be appreciated from the matrices, distances are generally lowest along the diagonal, indicating the relationship between the two variables across vowels. However, many of the rows are qualitatively difficult to discriminate, suggesting that lip widths do not uniquely specify individual vowels' resonant frequencies across utterances. Thus, lip widths may provide general information to constrain hypotheses about, but not specify, vowel identity.

To assess the discriminability of each vowel's distance profile quantitatively, we used hierarchical agglomerative clustering to group rows based on their vector distances (Fig. 2*B*, dendrograms). Clustering was performed using the mean Euclidean vector distance between clusters and thresholded at a default 70% of the maximum linkage. Consistent with visual inspection of the distance matrices, hierarchical clustering identified three clusters of vowels for each speaker. For both speakers, a low-frequency/narrow-lipped cluster, consisting of "hoed" (IPA: oʊ), "hood" (ʊ), "hawed" (ɔ), "hoyd" (ɔɪ), and "heard" (ɝ), was discriminated from a high-frequency/wide-lipped cluster consisting of "how'd" (aʊ), "hide" (aɪ), "had" (æ), "hid" (ɪ), "head" (ɛ), and "heed" (i). Intermediate between these two clusters, "hud" (ʌ) was classified differently for each speaker. Finally, one vowel at the audiovisual extremes could be discriminated from these main clusters for each speaker (female, "who'd," [u:]; male, "hayed," [eɪ]). These results further suggest that lip widths provide information that can aid discrimination between vowels, but do not uniquely specify individual vowels across utterances. We note, however, that the relationship between lip width and resonant frequency is likely stronger within individual utterances containing multiple vowels than across separate utterances containing individual vowels, as analyzed here, since there is less expected variability in the configuration of the rest of the vocal tract within individual utterances.

To further examine the discriminability of vowels based on lip width, resonant frequency, and their combination, we used linear discriminant analysis (LDA). Using LDA allowed us to assess how well individual vowels could be discriminated given the (co)variances and intervowel differences of each cue. Classifiers were trained on frame-wise values for each vowel (M = 7.36 frames per vowel; SD = 1.36) using a uniform prior and tested on the mean values for each vowel. Each classifier was trained and tested using only resonant frequency ("auditory," A), only lip width ("visual," V), or both ("audiovisual," AV). Fig. 2*C* shows the posterior probability assigned to each vowel (columns) after testing with each vowel's mean (rows). As shown in the A and V matrices, vowels could not be unambiguously discriminated based on the unimodal cues alone, with variable discriminability across the range of vowels and across modalities. Classifying based on both cues (AV) improved classifier accuracy (i.e., increased posterior probability along the main diagonal) for all vowels (Fig. 2*D*). Audiovisual accuracy was reduced to near chance (7.1%) when the frame-wise values for each cue were permuted within each vowel (M = 7.3%, SD < 0.001%; 5,000 permutations), illustrating the importance of consistent audiovisual covariation within and across vowels for accurate classification. Taken together, our cross-modal distance (Fig. 2*B*) and LDA (Fig. 2 *C* and *D*) analyses suggest that lip-width cues could plausibly aid vowel perception by constraining or supplementing perceptual inferences about vowel identity.

Articulatory Deformations Enhance Sensitivity for Corresponding Spectral Dynamics. Because both auditory frequency modulations (26–28) and visual shape properties (e.g., aspect ratio, curvature, convexity, and taper; ref. 29) are neurally encoded as midlevel perceptual features, we hypothesized that the perceptual system might exploit correspondences between feature-level representations of mouth shapes and oral resonances to recover degraded spectrotemporal information from visual speech. Such a mechanism would allow for perceptually informative spectrotemporal signals, such as formants and formant transitions, to be recovered from visual speech without employing higher-order speech-specific representations, such as phonemes, visemes, syllabic units, or articulatory representations.

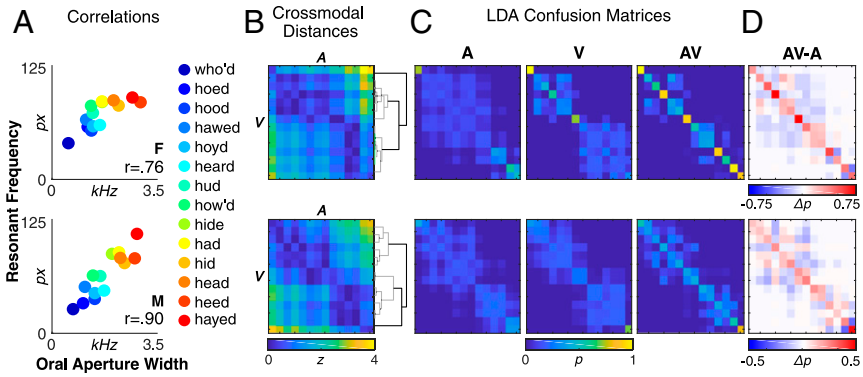

Fig. 2. Relationship between the peak resonant frequency (between 500 Hz and 3,000 Hz) and the width of the oral aperture for 14 vowels in /hVd/ context. Each row represents a single speaker (*Top*, female; *Bottom*, male). (*A*) Correlations between the average width of the oral aperture and the average resonant frequency during the pronunciation of each vowel. In the colored legend, vowels are ordered according to an orthogonal regression of the two variables, averaged across speakers. This same ordering is used for both axes of the matrices in *B* and *C*. (*B*) Pairwise distances between *z*-transformed lip widths (rows) and resonant frequencies (columns) for each vowel. The dendrograms to the right show hierarchical clustering of the visual-to-auditory distance profiles (i.e., the values in each row) for each vowel. Clusters below a default threshold of 70% of the maximum linkage are shown in gray. Three clusters were identified for each speaker, suggesting that lip widths can aid in distinguishing between vowel-associated resonances, but do not uniquely specify individual vowels. (*C*) Vowel confusion matrices for linear discriminant analysis (LDA) based on resonant frequency ("auditory," "A"), lip width ("visual," "V"), or both ("audiovisual," "AV"). Classifiers were trained on the frame-wise values for each vowel and tested on their means. Cell colors indicate the posterior probability assigned to each vowel (columns) after testing with each vowel's mean (rows). Vowels could not be unambiguously discriminated unimodally (A and V), with variable discriminability across the range of vowels and across modalities. Audiovisual classification (AV) provided improved accuracy (i.e., higher posterior probability along the main diagonal) for all vowels. (*D*) Differences in the posterior probabilities assigned to each vowel after auditory versus audiovisual classification. Red indicates increased probability with audiovisual classification; blue indicates decreased probability.

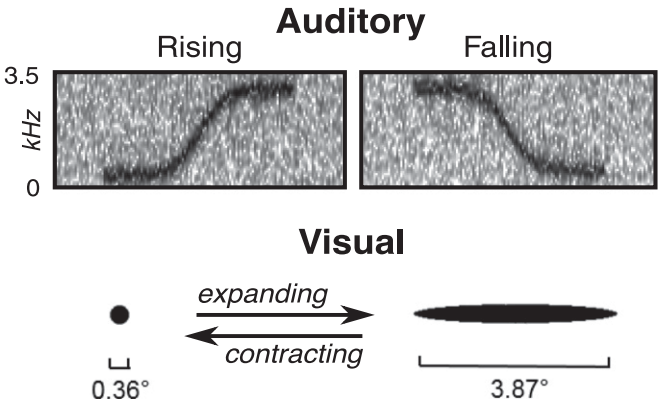

Fig. 3. Stimuli used in psychophysical detection task. (*Top*) Spectrograms of auditory frequency sweep stimuli embedded in white noise (−2 dB SNR). Frequency sweeps rose or fell from 500 to 3,000 Hz over 350 ms, approximating the trajectory of the formant contours observed in natural pronunciation of the syllables wi: (English: "we") and ju: (English: "ew"). A total of 100 ms of white noise preceded and followed each auditory stimulus. On each trial, participants detected which of two noise-filled intervals contained either of the frequency sweeps. Detection thresholds for different audiovisual conditions were estimated using an interleaved adaptive staircase procedure. (*Bottom*) Initial and final frames of visual deformation stimuli. The ellipses expanded or contracted in synchrony with the rising or falling of the frequency sweep stimuli. Deformations could be naturally correlated or unnaturally anticorrelated with the frequency sweeps and expand/contract horizontally (natural cross-modal relationship) or vertically (control). Visual stimuli appeared at the onset of the auditory frequency sweeps and disappeared when they terminated.

To test the hypotheses that viewing horizontal articulatory deformations of the mouth enhanced listeners' sensitivity for corresponding formant contours, we produced artificial stimuli that contained natural speech-based horizontal motion and resonant frequency dynamics, but no other auditory (e.g., amplitude modulation) or visual cues (e.g., facial features) typically present in speech (Fig. 3 and https://osf.io/mk7c9/ provide example stimuli).

To reproduce naturalistic frequency modulations in the oral resonance of interest, we generated frequency sweeps that rose or fell sigmoidally from 500 to 3,000 Hz over 350 ms, approximating the trajectory of the formant contours observed in natural pronunciation of the syllables wi: (English: "we") and ju: (English: "ew"). The frequency sweeps were generated by band-pass filtering white noise with a 100-Hz-wide filter to approximate the bandwidth of observed formants. The center frequency of the filter was updated every 4 ms to follow a logistic function with a steepness parameter based on averaged least-squares fits to the formant contours discussed in the previous section.

To generate visual stimuli with corresponding horizontal deformations, we produced animated ellipses with widths (0.36 ° to 3.87 ° visual angle) determined by a linear transformation of the peak frequencies in the auditory stimuli, but static vertical extent (0.36 °). Like point-light walkers (30), these stimuli isolated the biological motion of interest while removing other potentially confounding features.

In order to assess whether the horizontal deformations facilitated the perception of naturally corresponding spectral dynamics, we used a criterion-free two-interval forced choice task to estimate participants' detection thresholds for the rising and falling auditory frequency sweeps during the visual presentation of naturally correlated horizontal deformations, unnaturally anticorrelated horizontal deformations, vertically deforming control stimuli, and static disks that provided only temporal information about the onset and offset of the auditory stimulus. On each trial, the task was to identify which of two noise-filled intervals (50 dB SPL) contained either of the auditory stimuli. To enforce attention to the visual stimuli, participants were asked to identify the type of visual deformation (expanding, contracting, or static) displayed on a subset (~9%) of trials. To estimate the auditory detection benefit provided by each animated visual stimulus over and above the effects of simple temporal cuing, we subtracted detection thresholds for trials with the static disk from detection thresholds for trials with each visual animation. This yielded cross-modal facilitation indices for each visual animation, which are plotted "negative up" in Fig. 4 so that taller bars indicate greater sensitivity (i.e., greater threshold decrement) (31).

As shown in the left half of Fig. 4, auditory sensitivity for the frequency sweeps was significantly increased when naturally corresponding visual deformations were presented simultaneously (blue bars; rising/expanding, $t[17] = 3.38$, $P = 0.004$; falling/contracting, $t[17] = 3.62$, $P = 0.002$). By contrast, there was no evidence of facilitation when unnaturally anticorrelated deformations (Fig. 4, orange bars) were presented (all $P > 0.45$). A significant sound-by-deformation interaction indicated a significant effect of cross-modal congruency ($F[1,17] = 9.23$, $P = 0.007$). Main effects of sound ($F[1, 17] = 0.05$, $P = 0.82$) and deformation ($F[1, 17] = 0.20$, $P = 0.67$) were not significant.

By contrast, as shown in the right half of Fig. 4, the same animations provided no benefit when they were rotated 90 ° so that they expanded and contracted vertically (gray bars; all $P > 0.08$), and there was no sound-by-deformation interaction ($F[1,17] = 0.46$, $P = 0.507$). Main effects of sound ($F[1, 17] = 0.10$, $P = 0.76$) and deformation ($F[1, 17] = 0.25$, $P = 0.62$) were also not significant. These results suggest that the observed benefits are likely due to perceptual exploitation of the natural statistical relationship between horizontal deformations and frequency modulations in speech, and not other potentially similar cross-modal correspondences such as looming-based associations between auditory frequency and visual expansion (32).

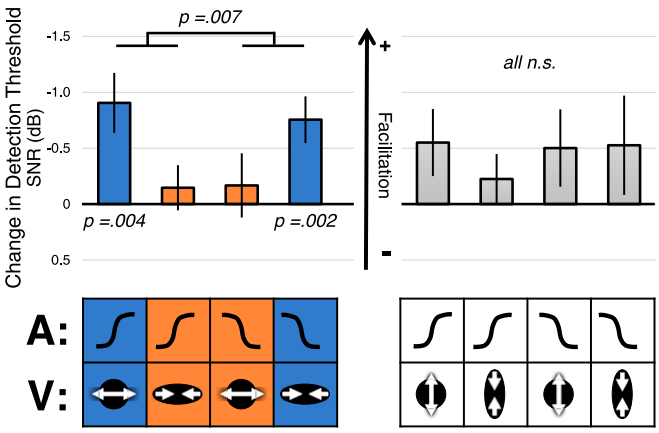

Fig. 4. Results of psychophysical detection experiment. Detection enhancements for auditory frequency sweeps during the presentation of naturally correlated (blue), unnaturally anticorrelated (orange), and vertical control (gray) visual deformations. Facilitation scores are plotted "negative up" so that taller bars indicate increased sensitivity (i.e., greater threshold decrement). Stimulus combinations for each condition are depicted in the lower table, with sigmoidal contours in the top row denoting rising and falling frequency sweeps and arrows in the bottom row indicating the direction of visual expansion or contraction. (*Left*) Sensitivity was significantly enhanced when frequency sweeps and visual deformations were correlated as they would be in natural speech (blue bars). By contrast, unnaturally anticorrelated stimuli produced no benefit (orange bars). Horizontal brackets indicate a significant sound-by-deformation interaction. (*Right*) Audiovisual congruency effects were abolished when the visual stimuli were rotated so that they expanded and contracted vertically (gray bars). Error bars represent ±1 SEM.

Interestingly, auditory sensitivity was, on average, improved in the vertical expansion conditions relative to the static baseline condition ($t[17] = 2.46$, $P = 0.025$), potentially indicating a general facilitative effect of visual motion or motion-based arousal. However, the lack of sound-by-deformation interaction for vertically deforming stimuli suggests that this effect did not reflect directionally specific cross-modal associations between frequency sweeps and vertical deformations.

In postexperimental interviews, none of the participants reported perceiving either the auditory or visual stimuli as speech or speech-related, suggesting that the effect does not require explicit processing of the stimuli as speech (33).

Finally, to estimate the size of the cross-modal modulation effect (i.e., congruent versus incongruent) in the horizontal deformation condition versus the vertical deformation condition, we compared cross-modal facilitation indices across all conditions using a linear mixed-effects model. The model included repeated fixed factors (unstructured covariance) for sound (rising vs. falling), deformation (expanding vs. contracting), orientation (horizontal vs. vertical), and their interactions. The three-way interaction between sound, deformation, and orientation had an estimated value of 1.00 dB SNR (95% CI, −0.03, 2.02), suggesting that the horizontal deformations produced ∼1 dB more cross-modal modulation than the vertical control condition. This result is consistent with previous reports using real speech stimuli, which typically show a 1- to 2-dB difference in detection thresholds for audiovisually congruent versus incongruent speech in noise (e.g., ref. 4).

Visual Speech Perceptually Restores Degraded Speech Spectra to Facilitate Comprehension. Our psychophysical detection experiment showed that visual speech cues can enhance perceptual sensitivity for auditory spectral information. Next, we examined whether this type of cross-modal interaction could improve speech intelligibility. Because formants and formant transitions are critical cues for speech perception and comprehension (34–36), we hypothesized that spectrotemporal information extracted from visual speech could facilitate the comprehension of degraded auditory speech. Moreover, we hypothesized that this cross-modal information might be particularly beneficial for the perception of natural, continuous speech, in which phoneme-level visual cues (e.g., visemes) might be obscured by coarticulation or ambiguous segmentation (37), but cross-modal correlations rooted in the acoustic physics of the oral cavity would be retained. To test this hypothesis, we selectively degraded the spectral or temporal dimension of auditory sentence spectrograms to assess how well visual speech improved speech intelligibility when each type of information was degraded to the point of near-complete unintelligibility (∼10% accuracy).

Sentence spectrograms were degraded by taking their 2D Fourier transforms and low-pass filtering the spectral (vertical) or temporal (horizontal) dimension, selectively "smearing" the spectrogram along each dimension while leaving the other dimension unaltered (Fig. 5*A* and https://osf.io/mk7c9 provide example stimuli) (38, 39). In the temporal degradation condition, temporal modulations were low-pass filtered below 2 Hz, obscuring amplitude modulations in the 2- to 7-Hz range in which rhythmically correlated mouth movements are thought to facilitate perception by strengthening neural entrainment to quasi-rhythmic dynamics in the speech amplitude envelope (7, 40). In the spectral degradation condition, information along the spectral axis was low-pass filtered below 0.01 cycles/kHz, obscuring spectral details associated with formants and formant transitions (4 cycles/kHz and lower; ref. 36) while preserving the broadband amplitude envelope. This level of spectral degradation was chosen to produce intelligibility comparable to the temporal degradation condition and to ensure that spectral peaks associated with individual formants could not be recovered in the absence of additional spectral information.

To assess the effects of visual speech on comprehension in each auditory degradation condition, we measured participants' ability to identify the words in typical American English sentences (41) after audiovisual, unimodal auditory, and unimodal visual presentations of each sentence. Speech intelligibility was indexed as the percentage of previously established keywords (41) that were correctly identified in each condition, and audiovisual intelligibility scores were compared to null probability summation (PS) models to test for significant multisensory interactions (31). As shown in Fig. 5*B*, accuracy in the audiovisual conditions was significantly higher than PS model predictions in both the temporal ($t[24] = 4.27$, $P < 0.001$) and spectral ($t[24] = 8.54$, $P < 0.001$) degradation conditions, suggesting the presence of multisensory facilitations in both conditions.

To compare the strength of these multisensory enhancements, we compared the differences between observed audiovisual accuracy and PS model accuracy across the two degradation conditions. We found a significant difference between the degradation conditions ($t[24] = 6.73$, $P < 0.001$), suggesting stronger enhancement

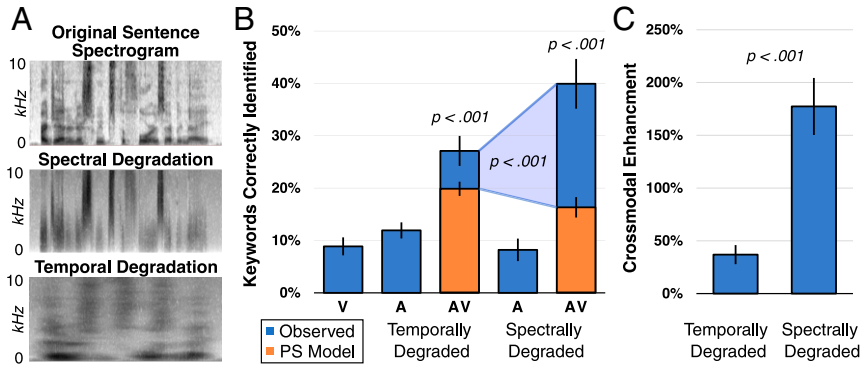

Fig. 5. Results of audiovisual speech comprehension experiment. (*A*) Sentence spectrograms were degraded along the spectral (vertical) or temporal (horizontal) dimensions to selectively obscure spectral (e.g., formants and formant transitions) or temporal (e.g., amplitude modulations) cues that are critical for speech intelligibility. (*B*) Percentage of previously established keywords accurately identified during unimodal visual, unimodal auditory, and audiovisual stimulus conditions. Comparing audiovisual accuracy to null-probability summation models (orange bars) revealed multisensory enhancements in speech comprehension for both temporally and spectrally degraded sentences, with a significantly larger enhancement in the spectral degradation condition. (*C*) Relative cross-modal enhancement indices (i.e., percent improvement relative to probability summation model) for temporally and spectrally degraded sentences. Cross-modal enhancements were substantially larger for spectrally degraded sentences, suggesting a mechanism that selectively facilitated comprehension in the spectral degradation condition by cross-modally restoring degraded spectral content. Error bars represent ±1 SEM.

in the spectral degradation condition (23.67 percentage points; $d_z = 1.71$) than in the temporal degradation condition (7.30 percentage points; $d_z = 0.85$). To facilitate comparison with relative enhancement metrics, we also computed relative cross-modal enhancement indices (i.e., percent improvement relative to PS model) for each condition. As shown in Fig. 5*C*, multisensory enhancements were substantially larger in the spectral degradation condition (186% increase relative to PS model; $d_z = 1.37$) than in the temporal degradation condition (43% increase; $d_z = 0.92$; $t[24] = 6.26$, $P < 0.001$). These results suggest the presence of an independent mechanism that selectively facilitated comprehension in the spectral degradation condition.

Because phoneme-level information provided by visual speech would be expected to enhance perception similarly regardless of the type of degradation, and because audiovisual cues to amplitude dynamics alone would not be sufficient to specify intelligible speech segments, these results suggest that visual speech improved comprehension in the spectral degradation condition by cross-modally restoring perceptually relevant spectral information. Phoneme representations are derived from integrated spectral and temporal acoustic-phonetic features (42, 43), so phoneme-level information extracted from visual speech would be expected to produce comparable enhancements when either type of information was degraded. Instead, we observed substantially larger improvements in the spectral degradation condition, suggesting that visual speech may have provided subphonemic spectral information that complemented auditory temporal cues to uniquely specify phoneme- or higher-level speech representations and facilitate comprehension.

Finally, we note that unimodal auditory accuracy was slightly (3.7 percentage points) but significantly ($t[24] = 2.38$, $P = 0.026$) higher in the temporal degradation than the spectral degradation condition, which could have contributed to observed differences in cross-modal enhancement through inverse effectiveness (i.e., greater multisensory enhancement when unimodal stimulation is less effective; refs. 44 and 45). However, previous research suggests that, inconsistent with the principle of inverse effectiveness, visual speech facilitates speech comprehension maximally at intermediate, rather than lower, signal-to-noise ratios (corresponding to 12 to 24% unimodal auditory accuracy), and that some apparent instances of inverse effectiveness in speech comprehension may result from methodological or quantitative artifacts (46–49). Because unimodal accuracy observed in this study was below previously observed points of maximal facilitation, our results are unlikely to have been influenced by inverse effectiveness.

Discussion

Taken together, our results suggest that visual speech signals facilitate perception by cross-modally restoring degraded spectrotemporal signals in auditory speech. These results both extend and constrain current models of audiovisual speech perception in multiple ways. First, they suggest that cross-modal correlations provide more detailed information about the contents of auditory speech than previously supposed. Whereas previous research has demonstrated that changes in the area of the oral aperture are associated with amplitude modulations in speech (40), our results suggest that the shape of the oral aperture can provide additional information about the spectral content of speech signals. Because linguistic information is encoded in both amplitude and frequency dynamics in auditory speech (38, 50), these results may help to explain the often drastic improvements in speech detection and comprehension produced by visual speech (1, 4, 5). Moreover, because these correlations are rooted in the acoustic physics of the oral cavity and its resonances, rather than abstracted categorical representations of speech (e.g., phonemes/visemes), they are maintained even when speech categories are obscured by articulatory variability or ambiguous segmentation (e.g., due to coarticulation) in natural, continuous speech. Thus, these cross-modal correspondences may be particularly effective in facilitating the perception of natural extended speech.

Additionally, our results suggest that cross-modal enhancements of speech perception need not rely on speech-specific supramodal or amodal representations of speech, but can occur at the level of nominally "unimodal," non–speech-specific feature coding. Previous models have posited the necessity of modality-neutral superordinate representations to act as a "common currency" subserving multisensory interactions in speech perception (15). However, the results of our psychophysical detection task suggest that speech-based multisensory interactions can occur between midlevel feature representations that are traditionally thought as part of "unimodal" processing hierarchies (26–29). The observed enhancements may therefore rely on cross-modal modulations of unisensory feature coding, rather than supramodal convergence, as has been observed in other, lower-level multisensory interactions (51). These feature-level interactions may arise from Hebbian-style learning of correspondences between environmentally cooccurring signals, so that the "common currency" subserving cross-modal enhancements is simply provided by the pattern of connections between neural populations that encode correspondent features. Such a mechanism could potentially be implemented in the dense reciprocal connections (52, 53) between the superior temporal gyrus, which encodes auditory spectrotemporal features (26–28), and the posterior superior temporal sulcus, which integrates visual form and motion signals associated with biological motion (54), including orofacial movements produced by visual speech (55, 56). Both regions, and interactions between the two, have frequently been implicated in audiovisual speech perception (6, 57, 58).

Moreover, because both visual shape signals and auditory spectrotemporal signals provide measures of the same underlying physical quantities (i.e., the spatial dimensions of resonant cavities in the vocal tract), the joint encoding of these quantities may provide a basis for perceptual interpolation between discrepant audiovisual speech signals, as is observed in audiovisual speech illusions such as the McGurk effect (59, 60). Indeed, previous results suggest that multisensory "fusion" percepts observed in the McGurk effect may arise from perceptual interpolation between formant trajectories associated with auditory and visual speech stimuli (61), and a similar mechanism could subserve speech-based modulations of visual shape perception (62). Such perceptual compromises could arise from precision-weighted averaging of audiovisual estimates of tract dimensions, allowing the perceptual system to use computational mechanisms employed in multisensory estimates of basic quantities [e.g., size (63), motion direction (64)] to enhance estimates of speech parameters as well (65).

The visual cues analyzed here likely work in concert with a variety of additional cues to supplement and corroborate auditory speech signals across multiple levels of speech encoding. Indeed, many articulatory features can affect the spectrotemporal content of speech, some of which are directly visible to observers, such as lingual (tongue) information (66–68), while others can be inferred based on correlations with visible features (69). Given these other potential sources of information, the present results likely reflect one instance of a larger phenomenon, in which the perceptual system maintains implicit statistical models relating multiple auditory and visual cues to facilitate perception. Together with additional visual cues to temporal dynamics and higher-level (e.g., phonemic, syllabic, phrasal) representations of speech, these cues likely help to support a robust and redundant perceptual system that can dynamically adapt to changing information availability throughout face-to-face conversations. This broader view, that multiple sources of visual information benefit speech perception, is consistent with previous suggestions that vision both supplements (referred to as the "complementary mode") and corroborates

(referred to as the "correlated mode") auditory speech signals to facilitate perception (70).

Materials and Methods

Experimental Model and Subject Details. Four student volunteers from Northwestern University (two female; ages 21 to 27 y) gave informed consent to participate in recordings of lip deformations and speech spectra. Author J.P. (male) and one female volunteer (ages 27 and 34 y, respectively) gave informed consent to participate in recordings with artificial glottal pulses. A total of 18 undergraduate students from Northwestern University (nine female; ages 18 to 21 y) gave informed consent to participate in the psychophysical detection task in exchange for partial course credit. A total of 25 undergraduate students from the University of Michigan (18 female; ages 18 to 21 y) gave informed consent to participate in the speech comprehension task in exchange for partial course credit. All participants had normal or corrected-to-normal vision and hearing. Experiments performed at Northwestern University were approved by the Northwestern University Social and Behavioral Sciences IRB. Experiments performed at the University of Michigan were approved by the University of Michigan Health Sciences and Behavioral Sciences IRB.

Method Details.

Audiovisual correlations. Volunteers were recorded from ∼1 m away with a Nikon D3200 digital camera (59.94 fps). Audio was simultaneously recorded with the built-in camera microphone and a Rode NTG2 condenser microphone placed 10 cm away from the speaker's mouth. The high-quality audio stream was digitized at a 48-kHz sampling rate with an E-MU 0404 analog–digital converter.

For analysis, high-quality audio was aligned to the camera's audio stream using peak alignment in Audacity 2.1.0. Formant peaks were extracted at the time of each video frame refresh using the Burg method in Praat 5.3.76 (71).

To measure the width of the oral aperture, two research assistants positioned digital markers on the left- and rightmost visible points of the oral aperture using either the mouse or arrow keys. Markers were placed at the point where the upper and lower lips converged to form a visible edge against the intraoral background. Width was measured as the Euclidean distance between the two points. The research assistants did not have access to the corresponding audio and were naïve to the study hypotheses.

Recordings of oral resonances. A 28-mm speaker (8 Ω, 0.5 W) was affixed to a wooden skewer and held facing upward ∼3 mm above the tip of the tongue. The speaker had a frequency range of ∼600 to 10,000 Hz and a resonant frequency of 680 Hz. Twenty-second recordings of artificial glottal pulse trains (100 Hz) generated in Praat ["PointProcess: To Sound (phonation)"; default parameters] were played through the speaker as it was held in the mouth of author J.P. and one female volunteer. J.P. and the volunteer repeatedly expanded and contracted their lips horizontally while the resultant auditory signal was recorded with a Rode NTG2 condenser microphone placed 5 cm away from the mouth. Audiovisual recording and alignment were performed in the same manner as outlined above. Time points containing the minimum and maximum oral aperture widths for each repetition were identified by a naïve research assistant who did not have access to the corresponding audio. The temporal midpoint was used when the local minimum or maximum persisted for longer than one frame.

Analysis of vowels. Audiovisual recording of participants F05 and M04 were retrieved from the Audiovisual Database of Spoken American English (25). Voiced portions of each of 14 vowels in /hVd/ context (colored legend in Fig. 2*A*) were first extracted in Praat using the Vocal Toolkit Plugin (www.praatvocaltoolkit.com/). The dominant spectral peak between 500 and 3,000 Hz was detected using the Burg method at the time of each video frame refresh. Oral aperture width was extracted from each video frame using a semiautomated snake segmentation algorithm (72). Hierarchical agglomerative clustering was performed using the MATLAB "linkage" function with the "average" method. LDA was performed using the MATLAB functions "fitcdiscr" and "predict."

Psychophysical threshold estimation. Auditory detection thresholds were estimated for each of 10 audiovisual stimulus pairs (audio, rising or falling frequency sweeps; video, horizontally expanding or contracting, vertically expanding or contracting, or static ellipses) using randomly interleaved QUEST staircases in a two-interval forced choice task (35 trials each) (73, 74). Each staircase was configured to estimate the threshold parameter of a logistic function with slope parameter $\beta = .68$, lapse rate $\lambda = .02$, and guess rate $\gamma = .5$. The slope parameter was selected by averaging the slope parameters estimated using the method of constant stimuli for two trained and two untrained observers in the static ellipse condition.

Visual stimuli and auditory white noise were presented during both intervals, but frequency sweeps were only presented during one interval. The main task was to identify which interval contained the frequency sweep. Participants were not asked to identify the direction of the frequency sweep.

To enforce attention to the visual stimuli, we also included 35 visual catch trials in which participants were asked to report the type of visual stimulus displayed (expanding, contracting, or static). To ensure similar distributions of trial types throughout the duration of the experiment, 1 of each of the 11 trial types (10 QUEST conditions and 1 catch condition) were always presented in random order before moving on to the next set of 11 trials. This process was repeated 35 times for a total of 385 trials.

Degraded sentence comprehension. Audiovisual recordings of participant F10 speaking 100 Central Institute for the Deaf (CID) Everyday Sentences (Lists A to J) were retrieved from the Audiovisual Database of Spoken American English (25, 41). Auditory streams for each sentence were spectrally and temporally degraded using ModFilter Sound Tools version 5 (38).

Twenty sentences were presented in each of five conditions: unimodal visual, unimodal auditory with temporal degradation, audiovisual with temporal degradation, unimodal auditory with spectral degradation, and audiovisual with spectral degradation. The sentences presented in each condition were counterbalanced across participants (at the level of 10-sentence CID Everyday Sentences lists) so that no sentence was ever presented twice to the same participant. The 10 lists were divided into 5 sets of 2 lists, and the 2 lists presented in each condition were counterbalanced across participants.

During the experimental trials, stimuli from each condition were presented in randomized order. After each stimulus presentation, participants were asked to type in any of the words that they thought that they had heard. Participants were encouraged to enter their best guesses even if they were not entirely certain of what they heard, and the experimenter emphasized that there was no penalty for inputting incorrect words. Accuracy was quantified as the percentage of keywords correctly identified in each condition. Keywords were those selected by the original authors of the sentences or homophones of those words (41).

Additional methodological details are given in *SI Appendix, Supplementary Materials and Methods*.

Data Availability Statement. Data and example stimuli are available at Open Science Framework, https://osf.io/mk7c9/.

ACKNOWLEDGMENTS. This research was supported by NIH Grant T32 NS047987.

1. W. H. Sumby, I. Pollack, Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* **26**, 212–215 (1954).
2. J. Irwin, L. DiBlasi, Audiovisual speech perception: A new approach and implications for clinical populations. *Lang. Linguist. Compass* **11**, e12237 (2017).
3. S. Puschmann *et al*., Hearing-impaired listeners show increased audiovisual benefit when listening to speech in noise. *Neuroimage* **196**, 261–268 (2019).
4. K. W. Grant, P.-F. Seitz, The use of visible speech cues for improving auditory detection of spoken sentences. *J. Acoust. Soc. Am.* **108**, 1197–1208 (2000).
5. J. C. Cotton, Normal "visual hearing.". *Science* **82**, 592–593 (1935).
6. M. S. Beauchamp, "Audiovisual speech integration: Neural substrates and behavior" in *Neurobiology of Language*, G. Hickok, S. L. Small, Eds. (Academic Press, San Diego, CA, 2016), pp. 515–526.
7. C. E. Schroeder, P. Lakatos, Y. Kajikawa, S. Partan, A. Puce, Neuronal oscillations and visual amplification of speech. *Trends Cogn. Sci.* **12**, 106–113 (2008).
8. H. Luo, Z. Liu, D. Poeppel, Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. *PLoS Biol.* **8**, e1000445 (2010).
9. E. Zion Golumbic, G. B. Cogan, C. E. Schroeder, D. Poeppel, Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". *J. Neurosci.* **33**, 1417–1426 (2013).
10. M. J. Crosse, J. S. Butler, E. C. Lalor, Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *J. Neurosci.* **35**, 14195–14204 (2015).
11. L. D. Braida, Crossmodal integration in the identification of consonant segments. *Q. J. Exp. Psychol. A* **43**, 647–677 (1991).
12. K. P. Green, "The use of auditory and visual information during phonetic processing: Implications for theories of speech perception" in *Hearing by Eye II*, R. Campbell, B. Dodd, D. Burnham, Eds. (Psychology Press, East Sussex, UK, 1998), pp. 3–26.
13. L. E. Bernstein, "Visual speech perception" in *Audiovisual Speech Processing*, E. Vatikiotis-Bateson, G. Bailly, P. Perrier, Eds. (Cambridge University, Cambridge, UK, 2012), pp. 21–39.
14. Q. Summerfield, Lipreading and audio-visual speech perception. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **335**, 71–78 (1992).

15. C. A. Fowler, "Speech as a supramodal or amodal phenomenon" in *The Handbook of Multisensory Processes*, G. A. Calvert, C. Spence, B. E. Stein, Eds. (MIT Press, Cambridge, MA, 2004), pp. 189–201.
16. C. Cheung, L. S. Hamiton, K. Johnson, E. F. Chang, The auditory representation of speech sounds in human motor cortex. *eLife* **5**, e12577 (2016).
17. J. S. Arsenault, B. R. Buchsbaum, No evidence of somatotopic place of articulation feature mapping in motor cortex during passive speech perception. *Psychon. Bull. Rev.* **23**, 1231–1240 (2016).
18. T. Paris, J. Kim, C. Davis, Visual speech form influences the speed of auditory speech processing. *Brain Lang.* **126**, 350–356 (2013).
19. M. F. Assaneo, D. Poeppel, The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm. *Sci. Adv.* **4**, eaao3842 (2018).
20. D. Poeppel, The analysis of speech in different temporal integration windows: Cerebral lateralization as "asymmetric sampling in time." *Speech Commun.* **41**, 245–255 (2003).
21. P. Delattre, The physiological interpretation of sound spectrograms. *PMLA* **66**, 864–875 (1951).
22. G. Fant, *Acoustic Theory of Speech Production: With Calculations Based on X-Ray Studies of Russian Articulations*, (De Gruyter Mouton, Revised edition, 1971).
23. G. M. Kuhn, On the front cavity resonance and its possible role in speech perception. *J. Acoust. Soc. Am.* **58**, 428–433 (1975).
24. H. Hermansky, D. J. Broad, "The effective second formant F2′ and the vocal tract front-cavity" in *International Conference on Acoustics, Speech, and Signal Processing*, (IEEE, Piscataway, NJ, 1989), Vol. vol.1, pp. 480–483.
25. C. Richie, S. Warburton, M. Carter, Data from "Audiovisual Database of Spoken American English." Linguistic data consortium. https://catalog.ldc.upenn.edu/LDC2009V01#:~:text=The%20Audiovisual%20Database%20of%20Spoken,speech%20production%20and%20speech%20recognition. Deposited 29 August 2017.
26. B. Tian, J. P. Rauschecker, Processing of frequency-modulated sounds in the lateral auditory belt cortex of the rhesus monkey. *J. Neurophysiol.* **92**, 2993–3013 (2004).
27. R. Santoro *et al*., Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLOS Comput. Biol.* **10**, e1003412 (2014).
28. P. W. Hullett, L. S. Hamilton, N. Mesgarani, C. E. Schreiner, E. F. Chang, Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. *J. Neurosci.* **36**, 2014–2026 (2016).
29. S. Suzuki, "High-level pattern coding revealed by brief shape aftereffects" in *Fitting the Mind to the World: Adaptation and After-Effects in High-Level Vision*, C. W. G. Clifford, G. Rhodes, Eds. (Oxford University Press, Oxford, UK, 2005), pp. 135–172.
30. G. Johansson, Visual perception of biological motion and a model for its analysis. *Percept. Psychophys.* **14**, 201–211 (1973).
31. J. Plass, D. Brang, S. Suzuki, M. Grabowecky, Supplemental materials for Vision Perceptually Restores Auditory Spectral Dynamics in Speech. Open Science Framework. https://osf.io/mk7c9/. Deposited 8 May 2020.
32. A. A. Ghazanfar, J. X. Maier, Rhesus monkeys (Macaca mulatta) hear rising frequency sounds as looming. *Behav. Neurosci.* **123**, 822–827 (2009).
33. K. Eskelund, J. Tuomainen, T. S. Andersen, Multistage audiovisual integration of speech: Dissociating identification and detection. *Exp. Brain Res.* **208**, 447–457 (2011).
34. B. E. Lindblom, M. Studdert-Kennedy, On the role of formant transitions in vowel recognition. *J. Acoust. Soc. Am.* **42**, 830–843 (1967).
35. S. Furui, On the role of spectral transition for speech perception. *J. Acoust. Soc. Am.* **80**, 1016–1025 (1986).
36. C. Liu, D. A. Eddins, Effects of spectral modulation filtering on vowel identification. *J. Acoust. Soc. Am.* **124**, 1704–1715 (2008).
37. A. Turkmani, *Visual Analysis of Viseme Dynamics*, (University of Surrey, 2008).
38. T. M. Elliott, F. E. Theunissen, The modulation transfer function for speech intelligibility. *PLOS Comput. Biol.* **5**, e1000302 (2009).
39. C. R. Holdgraf *et al*., Rapid tuning shifts in human auditory cortex enhance speech intelligibility. *Nat. Commun.* **7**, 13654 (2016).
40. C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, A. A. Ghazanfar, The natural statistics of audiovisual speech. *PLOS Comput. Biol.* **5**, e1000436 (2009).
41. H. Davis, S. R. Silverman, *Hearing and Deafness*, (Holt, Rinehart and Winston, 1970).
42. B. H. Repp, Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychol. Bull.* **92**, 81–110 (1982).
43. D. B. Pisoni, P. A. Luce, "Trading relations, acoustic cue integration, and context effects in speech perception" in *The Psychophysics of Speech Perception*, M. E. H. Schouten, Ed. (NATO ASI Series, Springer, Dordrecht, 1987), pp. 155–172.
44. M. A. Meredith, B. E. Stein, Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *J. Neurophysiol.* **56**, 640–662 (1986).
45. B. E. Stein, T. R. Stanford, Multisensory integration: Current issues from the perspective of the single neuron. *Nat. Rev. Neurosci.* **9**, 255–266 (2008).
46. L. A. Ross, D. Saint-Amour, V. M. Leavitt, D. C. Javitt, J. J. Foxe, Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cereb. Cortex* **17**, 1147–1153 (2007).
47. W. J. Ma, X. Zhou, L. A. Ross, J. J. Foxe, L. C. Parra, Lip-reading aids word recognition most in moderate noise: A Bayesian explanation using high-dimensional feature space. *PLoS One* **4**, e4638 (2009).
48. N. P. Holmes, The law of inverse effectiveness in neurons and behaviour: Multisensory integration versus normal variability. *Neuropsychologia* **45**, 3340–3345 (2007).
49. N. P. Holmes, The principle of inverse effectiveness in multisensory integration: Some statistical considerations. *Brain Topogr.* **21**, 168–176 (2009).
50. F.-G. Zeng *et al*., Speech recognition with amplitude and frequency modulations. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 2293–2298 (2005).
51. N. van Atteveldt, M. M. Murray, G. Thut, C. E. Schroeder, Multisensory integration: Flexible use of general operations. *Neuron* **81**, 1240–1253 (2014).
52. C. L. Barnes, D. N. Pandya, Efferent cortical connections of multimodal cortex of the superior temporal sulcus in the rhesus monkey. *J. Comp. Neurol.* **318**, 222–244 (1992).
53. B. Seltzer, D. N. Pandya, Parietal, temporal, and occipital projections to cortex of the superior temporal sulcus in the rhesus monkey: A retrograde tracer study. *J. Comp. Neurol.* **343**, 445–463 (1994).
54. M. A. Giese, T. Poggio, Neural mechanisms for the recognition of biological movements. *Nat. Rev. Neurosci.* **4**, 179–192 (2003).
55. N. Furl, R. N. Henson, K. J. Friston, A. J. Calder, Network interactions explain sensitivity to dynamic faces in the superior temporal sulcus. *Cereb. Cortex* **25**, 2876–2882 (2015).
56. L. L. Zhu, M. S. Beauchamp, Mouth and voice: A relationship between visual and auditory preference in the human superior temporal sulcus. *J. Neurosci.* **37**, 2697–2708 (2017).
57. A. A. Ghazanfar, C. Chandrasekaran, N. K. Logothetis, Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys. *J. Neurosci.* **28**, 4457–4469 (2008).
58. L. H. Arnal, B. Morillon, C. A. Kell, A.-L. Giraud, Dual neural routing of visual facilitation in speech processing. *J. Neurosci.* **29**, 13445–13453 (2009).
59. H. McGurk, J. MacDonald, Hearing lips and seeing voices. *Nature* **264**, 746–748 (1976).
60. Q. Summerfield, M. McGrath, Detection and resolution of audio-visual incompatibility in the perception of vowels. *Q. J. Exp. Psychol. A* **36**, 51–74 (1984).
61. K. P. Green, "The use of auditory and visual information in phonetic perception" in *Speechreading by Humans and Machines*, D. G. Stork, M. E. Hennecke, Eds. (NATO ASI Series, Springer, Berlin, 1996), pp. 55–77.
62. T. D. Sweeny, E. Guzman-Martinez, L. Ortega, M. Grabowecky, S. Suzuki, Sounds exaggerate visual shape. *Cognition* **124**, 194–200 (2012).
63. M. O. Ernst, M. S. Banks, Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433 (2002).
64. C. R. Fetsch, A. H. Turner, G. C. DeAngelis, D. E. Angelaki, Dynamic reweighting of visual and vestibular cues during self-motion perception. *J. Neurosci.* **29**, 15601–15612 (2009).
65. J. F. Magnotti, M. S. Beauchamp, A causal inference model explains perception of the McGurk effect and other incongruent audiovisual speech. *PLOS Comput. Biol.* **13**, e1005229 (2017).
66. J. Jiang, A. Alwan, P. A. Keating, E. T. Auer, L. E. Bernstein, On the relationship between face movements, tongue movements, and speech acoustics. *EURASIP J. Adv. Signal Process.* **2002**, 506945 (2002).
67. H.-W. Kim, H. Nam, C.-Y. Kim, [i] is lighter and more greenish than [o]: Intrinsic association between vowel sounds and colors. *Multisens. Res.* **31**, 419–437 (2018).
68. P. Rubin *et al*, "CASY and extensions to the task-dynamic model" in *1st ETRW on Speech Production Modeling: From Control Strategies to Acoustics. 4th Speech Production Seminar: Models and Data* (Institut de la Communication Parlée, Grenoble, France, 1996).
69. H. Yehia, P. Rubin, E. Vatikiotis-Bateson, Quantitative association of vocal-tract and facial behavior. *Speech Commun.* **26**, 23–43 (1998).
70. R. Campbell, The processing of audio-visual speech: Empirical and neural bases. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 1001–1010 (2008).
71. P. Boersma, Praat, a system for doing phonetics by computer. *Glot Int.* **5**, 341–345 (2002).
72. M. Kass, A. Witkin, D. Terzopoulos, Snakes: Active contour models. *Int. J. Comput. Vis.* **1**, 321–331 (1988).
73. A. B. Watson, D. G. Pelli, QUEST: A Bayesian adaptive psychometric method. *Percept. Psychophys.* **33**, 113–120 (1983).
74. N. Prins, F. A. A. Kingdom, Palamedes: Matlab routines for analyzing psychophysical data (2009). www.palamedestoolbox.org. Accessed 12 March 2013.