Original Article

Trends in Hearing
Volume 23: 1–14
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2331216519837866
journals.sagepub.com/home/tia
SAGE

Visual Speech Benefit in Clear and Degraded Speech Depends on the Auditory Intelligibility of the Talker and the Number of Background Talkers

Catherine L. Blackburn[1], **Pádraig T. Kitterick**[2,3], **Gary Jones**[1], **Christian J. Sumner**[1,4], **and Paula C. Stacey**[1]

Abstract

Perceiving speech in background noise presents a significant challenge to listeners. Intelligibility can be improved by seeing the face of a talker. This is of particular value to hearing impaired people and users of cochlear implants. It is well known that auditory-only speech understanding depends on factors beyond audibility. How these factors impact on the audio-visual integration of speech is poorly understood. We investigated audio-visual integration when either the interfering background speech (Experiment 1) or intelligibility of the target talkers (Experiment 2) was manipulated. Clear speech was also contrasted with sine-wave vocoded speech to mimic the loss of temporal fine structure with a cochlear implant. Experiment 1 showed that for clear speech, the visual speech benefit was unaffected by the number of background talkers. For vocoded speech, a larger benefit was found when there was only one background talker. Experiment 2 showed that visual speech benefit depended upon the audio intelligibility of the talker and increased as intelligibility decreased. Degrading the speech by vocoding resulted in even greater benefit from visual speech information. A single "independent noise" signal detection theory model predicted the overall visual speech benefit in some conditions but could not predict the different levels of benefit across variations in the background or target talkers. This suggests that, similar to audio-only speech intelligibility, the integration of audio-visual speech cues may be functionally dependent on factors other than audibility and task difficulty, and that clinicians and researchers should carefully consider the characteristics of their stimuli when assessing audio-visual integration.

Keywords

speech in noise, visual speech, speech perception, integration, degraded speech

Date received: 20 July 2018; revised: 13 February 2019; accepted: 20 February 2019

Introduction

"Visual speech" information is defined as being able to see the movements of the talker's mouth, including the lips, tongue, and teeth (Peelle & Sommers, 2015). It provides phonetic and temporal cues that aid the perception of speech and is particularly beneficial to communication in noisy environments (Miller, 1947). The benefits of seeing a talker's face can be quantified in terms of the difference in the amount of noise that can be tolerated while maintaining high levels of speech understanding with and without access to visual speech information (Calvert, 2001; Grant & Walden, 1996; Lusk & Mitchell, 2016). For example, Middelweerd and Plomp (1987) found that when people were presented with auditory and visual speech information, the average signal-to-noise ratio (SNR) at which young adults could report 50% of words correctly (the speech reception threshold,

[1]Department of Psychology, Nottingham Trent University, UK
[2]Nottingham Biomedical Research Centre, UK
[3]Division of Clinical Neuroscience, School of Medicine, University of Nottingham, UK
[4]Medical Research Council Institute of Hearing Research, Nottingham, UK

Corresponding Author:
Catherine L. Blackburn, Department of Psychology, Nottingham Trent University, 50 Shakespeare Street, Nottingham NG1 4FQ, UK.
Email: catherine.blackburn@ntu.ac.uk

SRT50) was −14 dB compared with −9 dB when only auditory information was available.

Visual speech information is especially important to individuals with hearing impairments (Erber, 1975; Kaiser, Kirk, Lachs, & Pisoni, 2003), for whom difficulties with listening in noisy environments can make every day social interactions more demanding, and ultimately has detrimental effects on their quality of life (Hawthorne et al., 2004; Hilly et al., 2016; Saki et al., 2017). Those fitted with cochlear implants (CI) find listening in noisy environments particularly demanding (Turner, Gantz, Vidal, Behrens, & Henry, 2004). The loss of fine spectral and temporal information limits their ability to segregate speech from background noise and benefit from glimpses in speech (Bhargava, Gaudrain, & Başkent, 2016), and to differentiate between talkers due to the loss of pitch cues (Qin & Oxenham, 2003). As CI users report that most speech encounters occur when both visual and auditory information is available (Dorman et al., 2016), their ability to benefit from access to visual speech information is likely to play a key role in their capacity to communicate in everyday life.

Stacey, Kitterick, Morris, and Sumner (2016) investigated the potential benefits of visual speech information available to CI users by assessing the performance of normal hearing listeners when listening to clear speech and speech sine-wave vocoded to simulate the information provided by a CI. This was assessed for a single background noise type: multitalker babble. They identified a modest (∼2 dB) but consistent increase in the size of visual speech benefit when participants listened to vocoded speech compared with when they listened to unprocessed speech. The finding that the value of visual information increases when the acoustic signal is degraded, in this case by removing informative temporal fine structure (TFS), is compatible with the principle of inverse effectiveness (PoIE; Sumby & Pollack, 1954). According to the PoIE, as unimodal performance declines multisensory integration is improved (Meredith & Stein, 1986). By extension, where audio input is most degraded the visual gain is at its greatest (Callan, Callan, Kroos, & Vatikiotis-Bateson, 2001). This principle is also supported by neurophysiological evidence as brain activity in response to visual speech information is enhanced in the presence of background noise (Callan et al., 2003).

The extent to which auditory-alone speech perception is degraded in everyday environments itself depends on numerous factors, including the nature of the background "noise." Rosen, Souza, Ekelund, and Majeed (2013) found that speech perception in normal-hearing individuals dropped off markedly as the number of background talkers was increased from one to two, likely due to an inability to disentangle similar streams of information (Durlach et al., 2003), but was relatively stable as the number of talkers increased thereafter. The intelligibility of the target talker also affects auditory-only speech perception. Different talkers have different levels of intelligibility (Brungart, 2001; Gagne, Masterson, Munhall, Bilida, & Querengesser, 1994; Lander & Davies, 2008). As such, the choice of target talker may change the task demands by placing greater reliance on linguistic knowledge, contextual information about a topic of conversation, or familiarity with the talker. Different listeners also respond differently to different talkers, a finding that is consistent with the idea that high-level cognitive functions also influence speech perception abilities (Conrey & Gold, 2006).

Since the nature of target talker, the background "noise," and any signal processing or degradation imposed by a hearing instrument can all influence speech understanding based on auditory information alone, an outstanding question is whether the benefits of visual speech information interact with these multiple factors. It is necessary to address this question in order to understand how to maximize the benefit of visual speech and the real-world value of auditory prostheses to hearing impaired listeners (e.g., by guiding rehabilitation strategies).

A testable formulation of this question is as follows: Do people integrate audio and visual sources of sensory information in the same way under all circumstances? There are different ways of combining information from two senses that would plausibly lead to different patterns of benefit from visual speech. For example, seeing the face of a talker could provide additional information about the identity of words being spoken, even in the absence of any background sounds. Alternatively, in the presence of background noise or a competing talker, the visual information might be used to help the listener segregate the target speech from the background speech by indicating when in time the target speaker is talking and audible. In this study, we assessed multisensory performance against the null-hypothesis of signal detection theory (SDT) models (Micheyl & Oxenham, 2012; Stacey et al., 2016), which assume that information is integrated in a fixed way. Our premise is that if a single SDT model can account for data across a range of experimental conditions, then the results are compatible with the idea that the process of sensory integration is fixed.

According to SDT models, audio-visual speech performance is a result of combining the information from the underlying unisensory "representations" in the brain to form an integrated multisensory representation of the speech to be recognized. A decision is then made about the words (from memory) which are most likely to give rise to this combined audio-visual representation. The model makes several simple assumptions, including: that information is integrated *before* any decisions are made about the words spoken; that audio and visual streams contribute statistically independent sources of information; and that information is combined in an optimal way, maximizing multisensory performance by weighting the individual sources according to their reliability. Thus, using SDT, it is possible compare the observed data against the predictions of an ''ideal observer'' which optimally integrates sensory information (Micheyl & Oxenham, 2012).

Despite the relative simplicity of the model, it can influence how experimental results are interpreted. For example, while Stacey et al. (2016) found that visual speech benefit was greater when speech was vocoded than when it was natural, a single SDT model accounted for the data under both listening conditions, implying that there was no change in the underlying processing. In the current study, SDT was again employed to provide a basis for assessing whether integration of audio-visual information was optimal and unchanging across the experimental manipulations. Conversely, if a single model cannot account for the data, it implies that audio-visual integration changes across experimental conditions.

The specific aim of the current study was to evaluate the effects of manipulating the number of background talkers (Experiment 1) and the auditory intelligibility of the target talker (Experiment 2) on the ability of listeners to benefit from access to visual speech information. The effects of these manipulations were assessed using both clear and vocoded speech. Vocoding allowed us to examine the interaction between informative TFS information and visual speech benefit which may give us insight into the visual speech benefits obtained by people with CIs. Following the PoIE principle, whereby multisensory integration is improved as unimodal performance declines, it was expected that the beneficial effect of having access to visual speech information would increase as the number of background talkers increased and would be greater for target talkers with poorer levels of auditory intelligibility.

Material and Methods

Design

A within-participants design was used for both experiments. For Experiment 1, 12 conditions were created from the factorial combination of the factors speech type (whether speech was clear or vocoded), modality (whether stimuli were audio visual or audio only), and background talkers (1, 2, or 16 talkers in the background). An additional condition was also included that measured performance in a visual-only task. For Experiment 2, 16 conditions were created from the factorial combination of the factors speech type (clear or vocoded), modality (whether stimuli were audio visual or audio only), and talker identity (four talkers with varying levels of intelligibility). Visual-only performance for each of the four target talkers used in Experiment 2 was measured in a supplementary experiment.

Participants

Twenty-four participants took part in each experiment (Experiment 1: age 19–47 years, mean age 29, 7 males; Experiment 2: age 18–33 years, mean age 22, 2 males; Experiment 2 supplementary: age 19–31 years, mean age 21.5, 7 males). Participants were recruited from the student and staff population at Nottingham Trent University. Students were rewarded with research credits. Consent was obtained from each participant as agreed by the Nottingham Trent University Research Ethics Committee. Participants also confirmed normal hearing and normal or corrected-to-normal vision and had English as their first language.

Materials

Equipment. The experiments were conducted in a multi-person IAC Acoustics 40 a-5 audiology booth. Sound levels were calibrated by presenting the stimuli over headphones attached to an artificial ear (G.R.A.S. 43AA) and measured using a microphone (ACO 7052E) connected to a sound level meter (SVAN 977). Audio was played over HD280pro headphones (Sennheiser, Wedemark, Germany) via a custom-built digital-to-analogue converter. Visual stimuli were presented on a computer monitor with a screen measuring 41 cm × 26 cm. Stimulus presentation was controlled using E-Prime software (Version 2.0 Psychology Software Tools Inc., Sharpsburg, United States) and using the MATLAB programming environment (Mathworks, Nantick, United States).

Target stimuli. Sentences were selected from the IEEE corpus that comprises 720 short sentences grouped into phonetically balanced lists of 10 sentences (Rothauser et al., 1969). Examples of the sentences are (with key words are underlined) ''Cars and buses were stuck in snow drifts'' and ''Use a pencil to write the first draft.'' For Experiment 1, audio-visual recordings were made of three hundred sentences spoken by a single male talker.

For Experiment 2, 11 different talkers were recorded articulating the same 30 IEEE sentences, creating a corpus of 330 sentences. The relative auditory intelligibility of these talkers was assessed by conducting a pilot study with six participants who were asked to identify key words in all 330 sentences presented in a random order at an SNR of −8 dB. The order of the sentences was randomized for each participant to minimize the influence of order effects if present. The percentage of key words correctly identified was recorded for each of the 11 talkers. Results showed a large variation between the auditory intelligibility of the talkers, with overall mean correct scores ranging from 45% to 88% (see Supplementary Table 1). Four talkers were selected for use in the main experiment; the two talkers with the highest ranked auditory intelligibility (one male, average score 88%; and one female, average score 82%) and the two talkers with the lowest ranked intelligibility (one male, average score 54%; and one female, average score 45%). Each sentence lasted approximately 3 seconds. The audio and video recordings were processed using Adobe Premiere Pro CC (v9.2).

Background stimuli. Background noise stimuli for use in both experiments were created using an existing database of speech materials (Markham & Hazan, 2002) and were informed by the procedures outlined in Rosen et al. (2013). Each recording in the database is a 30-second narrative of the talker describing in free-form language the scene they had witnessed in a video. The free-form nature of the description ensured that talkers were not repeating the same set text and therefore avoided obvious repetition. Silences of more than 100 ms were removed from each recording, and all filler expressions (e.g., erm, eh) removed. Recordings from 16 male talkers within the database were chosen on the basis that they sounded most similar to the male talker used to create the target stimuli for Experiment 1 and were used to create the 1-talker, 2-talker, and 16-talker background noises. On each trial, 3-second segments were extracted from the continuous narratives spoken by each background talkers and started from different points within each 30-second recording to avoid repetition of information and words within the resulting background noise between trials. A 16-talker background noise was used in Experiment 2. This was created in the same way as for Experiment 1, but as the target talkers were both male and female it consisted of eight male and eight female talkers.

Speech processing. Audio-visual sentence materials were processed using the MATLAB programming environment. The desired SNR was achieved by attenuating the target talker (for negative SNRs) or the background noise (for positive SNRs) before summing the two signals and normalizing the root mean square of the composite signal. The composite signal was then band-pass filtered into eight adjacent frequency bands spaced equally on an equivalent rectangular bandwidth frequency scale between 100 Hz and 8 kHz (Glasberg & Moore, 1990) using finite impulse response filters. In clear speech conditions, the auditory stimuli were constructed by summing the output of the eight band-pass filters. In vocoded conditions, the temporal envelope of each filter output was extracted using the Hilbert transform and used to modulate a sine wave at the center frequency of the filter and with alternating phase. The eight sine waves were then summed to form an auditory stimulus with uninformative TFS. This processing method ensured that the temporal envelopes were similar across clear and vocoded conditions (Eaves et al., 2011).

Procedure

The audio stimuli were presented at 70 to 73 dB SPL. Each video was displayed as an image 17 cm × 30 cm in the center of the screen. Participants were seated at approximately 0.5 m from the display monitor with the display at head height, meaning that the image subtended a horizontal visual angle of 19° and a vertical visual angle of 33°. They were instructed to watch the video and listen to the audio stimuli (AV conditions) or listen to the audio only (AO conditions) and repeat out loud any words they were able to understand at the end of each sentence. Stimuli were presented diotically but to help avoid floor effects a 0.001-second delay applied randomly to the copy of the target stimulus in the left or right ear. The result was that the target stimulus was perceived to originate toward the left or right ear and therefore from a different location to the background noise (London, Bishop, & Miller, 2012).

Participants first undertook a practice session in order to gain familiarity with the task in which five IEEE sentences were presented for each of four conditions: audio-only clear speech, audio-only vocoded speech, audio-visual clear speech, and audio-visual vocoded speech. In the audio-only and audio-visual conditions of the main experiments, an initial sentence was presented at an SNR of −16 dB and repeated at increasing SNRs in 4-dB steps until at least three out of the five key words were identified correctly. Once the key words were identified correctly, a different sentence was then presented on each subsequent trial, and the SNR was varied adaptively in 4-dB steps until two reversals were made and then in 2-dB steps for the remaining sentences in each condition. Twenty different sentences were presented in each condition. The SNRs for the final 10 sentences were averaged to produce a SRT50 for each participant, which represented the SNR measured in dB at which participants could report the key words within 50% of sentences correctly. In the visual-only conditions of both experiments, participants were asked to repeat any words they thought they could understand out loud at the end of each sentence and performance was scored in terms of percentage key words reported correctly.

In Experiment 1, the practice stimuli were presented at SNRs ranging from −8 dB to 8 dB in each type of background noise (1-, 2-, or 16-talker noise). During the practice sessions, SRTs were not estimated.

For the main experiment, 13 conditions were presented in random order: speech type (clear or vocoded), modality (audio visual or audio only), number of background talkers (1, 2, or 16), and a visual-only condition. Each condition was assigned a different list of 20 sentences from 260 IEEE sentences, and each list was presented in a random order. Therefore, each participant had different sentences for each condition and in a different presentation order.

In Experiment 2, practice stimuli were presented at an initial SNR of −8 dB, and the SNR was then reduced in 4-dB steps if three out of the five key words were correctly identified or otherwise increased in 4-dB steps. In the supplemental experiment that assessed visual-only performance, each participant was presented with 20 sentences from each of the four talkers. A different 20 sentences were presented for each talker. The 20 sentences for each talker were presented in a random order, and the order of the talkers was counterbalanced for each participant.

SDT Modeling

The modeling methods closely followed previous studies (Micheyl & Oxenham, 2012; Stacey et al., 2016). SDT (Green & Swets, 1966) posits that the participant's ability to discriminate sensory inputs depends on the differences in some internal representation between different signals, and the trial-to-trial variability of the representation (or "noise"): If the difference in the representation of two different stimuli is large relative to the variability associated with those representations, then it is easy to correctly identify the physical stimulus. SDT assumes that the noise associated with representations is normally distributed, and that the variance is the same for all possible stimuli. This means a participant's performance is determined by the number of standard deviations separating the variables' means. This is basis of the *d'* measurement in SDT (for a more detailed treatment of this, the reader is referred to Chapters 1 and 2 of Macmillan & Creelman, 1991). Here, we will illustrate how the model works and its extension to multisensory stimuli.

Consider first the simplest case, where a participant is required to identify a particular sensory input as belonging to one of two spoken words (referred to here by different numbers, i.e., stimulus $i=1$ or $i=2$). The incoming sensory stimulus is "represented" (somewhere in the brain) by a number, $a_{i,}$ the mean value of which depends on the word that is spoken. In addition, even for the same word, this number is variable from trial to trial. This reflects both unknown variability in the external stimulus (e.g., variability inherent in real speech) and the stochastic nature of processing by the nervous system. Figure 1(a) shows the distributions of internal representations which might arise from two different spoken words. The distributions shown here are not completely smooth as they are the result of Monte Carlo simulations, to emphasize that these are sampling distributions as if derived from perceiving the words to have been heard thousands of times each.

If the listener has learned these distributions, as would be assumed for typical adults, then they could use this information to decide the most likely word being spoken on any given trial. For two overlapping distributions, one simply divides the possible range of values from the combined distributions in two based on where those distributions cross, and response selection is whichever side of the line the current value of a_i falls (see vertical dashed line in Figure 1(a)). The less these distributions overlap, the more likely a response is to be correct.

To create Figure 1(a), we performed a Monte Carlo simulation using two random variables with means of 0 and 1.5, both having a standard deviation of 1. This yields a d' of 1.5, which corresponds to getting ~ 77% of trials correct. In Figure 1(b), we show a similar simulation with a d' of 1. The distributions are closer together, and therefore the participant's performance will be lower (~69%). For each different stimulus (here, different words) that must be distinguished and each circumstance (here, difference in background sound), there will be a distribution with a different mean value.

Now consider that these two examples are independent, auditory and visual sources of information about the same two possible words (Auditory: Figure 1(a) and Visual: Figure 1(b)), thus two noisy variables, a_i and v_i. We assume that multisensory integration is a process of combining these variables to form another number: our audio-visual representation. There are an infinite number of ways of combining these, but arguably the simplest is to add them. This gives a new multimodal, noisy internal variable:

$$av_i = w_a a_i + w_v v_i \qquad (1)$$

where w_a and w_v are weights that reflect the relative influence of each modality. Figure 1(c) shows a Monte Carlo simulation of adding the audio only (a_i) and visual only (v_i), for the two different possible words ($i=1$, blue; $i=2$, red). For example in Figure 1(a) and (b), the difference in av_i is larger than either input individually. However, the variance of the new distribution, av_i, will be larger, reflecting the combined variability of both inputs.

Furthermore, these resulting distributions and the ability to discriminate stimuli are also going to depend on the weights. In Figure 1(c), the weights have been chosen to maximize the number of standard deviations separating the two words after integration ($\mathrm{w_a/w_b} = .414$; recall that both the means and the standard deviation of

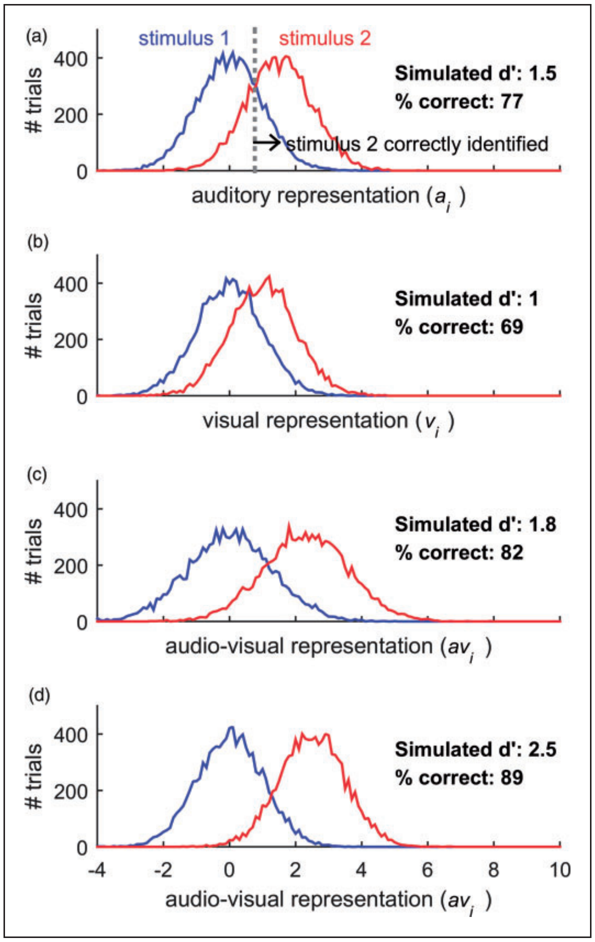

Figure 1. Monte Carlo simulations of noisy variables representing audio, visual, and audio-visual stimuli, representing a simple example of how SDT can be applied to audiovisual integration. (a) A simulation of the expected range of values of some internal representation of two possible audio stimuli. Each is represented as a normally distributed random variable with SD of 1 and means of 0 and 1.5, respectively. The vertical dashed line indicates the best possible criterion (or decision boundary) for deciding whether the Stimulus 1 or 2 is present. (b) An example similar to (a), except $d' = 1$, which is here representing the two stimuli in the visual modality. (c) The outcome of adding the visual and auditory noisy-variables together (independent noise), optimally weighted to achieve the maximum possible performance. (d) The outcome of adding the two variables optimally if the source of noise in the representation arises not from the processing in each modality separately but from a later process postintegration (late noise). SDT = signal detection theory.

av_i are affected by this). This yields two new distributions separated by ~ 1.8 standard deviations, that is, a d' of 1.8. This corresponds to a ~82% chance of answering correctly. No other weighting can improve on this performance: It is the optimal linear sum. This optimal weighting will depend on the relative discriminability of the unisensory inputs. The intuitive reason for this is if one source of information were much more informative than the other, it would make sense to weight this more strongly, and in the extreme case to completely ignore one modality if it is of no use in the decision.

It is possible to compute this d' using many Monte Carlo simulations, to optimize the weights. However, SDT elegantly provides a shortcut to the same answer (Micheyl & Oxenham, 2012). The combined d' can be calculated as:

$$d'_{av} = \sqrt{\left(d'_a\right)^2 + \left(d'_v\right)^2} \tag{2}$$

This equation yields $(1 + 1.5^2)^{0.5}$ =1.8, just like our simulation. Thus, given d'_A and d'_V measured in the unimodal conditions, the d' calculation predicts the optimal multisensory performance, assuming that statistically independent, noisy information from two modalities are combined with a simple weighted sum.

The above description considers that the only sources of noise in the internal representation arise purely from the unimodal sources. Any processing that occurs after the integration is perfect, every time. We call this the *independent noise* condition. It is also possible that noise could instead arise *after* the integration of information. Examples of this might be noise due to inattention, decision processes, or perhaps other cognitive sources of noise such as ambiguous lexical information in a speech task. This is called the *late noise* model (Micheyl & Oxenham, 2012).

Figure 1(d) shows a Monte Carlo simulation where the noise is late, and of the same variance as the previous independent noise examples. In the unimodal conditions, discrimination is identical to the independent noise model. However, in the audio-visual condition, the variance does not increase from the unimodal conditions, since it occurs after the integration. Thus, the simulation delivers a separation of 2.5 standard deviations (SDs) between the two multimodal distributions (or ~89% correct).

The implication of this assumption of purely late noise is that d'_{av} now becomes a simple addition of the individual d' values:

$$d'_{\text{av}} = d'_{\text{a}} + d'_{\text{v}} \tag{3}$$

Equations 2 and 3 provide a simple but powerful pair of models to aid the interpretation of the observed pattern of performance data in multisensory experiments, with no free parameters. Given the unisensory performance (as d') in both modalities, we can predict two extremes of optimal discrimination performance in the multisensory case, d'_{av}, under two extreme assumptions about the source of noise in the processing. Importantly, for each model, the only change in the underlying processing between audio and visual conditions is the optimal reweighting of the evidence, which SDT assumes us to be very good at.

When there are more than two stimulus categories, such as is typically the case for speech, the relationship between d' (which is assumed not to be influenced by the number of categories) and the proportion of correct trials that a participant will achieve becomes more complex. However, when presented with stimuli in a single modality, it can be expressed as a function of the overall discriminability, d', of the m different stimulus categories:

$$P = \int_{-\infty}^{+\infty} \phi\left(z - \mathrm{d}'\right) \Phi^{m}(z)\mathrm{d}z \tag{4}$$

where $\phi(.)$ is the standard normal probability density function, and $\Phi(.)$ is the cumulative standard normal function. This can also be easily extended to the audio-visual case, by inserting the corresponding equations above in place of d'. If it is assumed that the variability arises purely from the independent representations of the audio and visual stimuli before they are integrated (i.e., independent noise), then the proportion of correct trials is given by:

$$P = \int_{-\infty}^{+\infty} \phi\left(z - \sqrt{(\mathrm{d}'_A)^2 + (\mathrm{d}'_V)^2}\right) \Phi^{m}(z)\mathrm{d}z \tag{5}$$

Alternatively, if the variability could arise solely due to processes occurring after integration, the proportion of correct trials is given by:

$$P = \int_{-\infty}^{+\infty} \phi\left(z - (\mathrm{d}'_A + \mathrm{d}'_V)\right) \Phi^{m}(z)\mathrm{d}z \tag{6}$$

Following previous studies that have suggested that open set speech perception is best modeled as dependent on vocabulary size (Müsch & Buus, 2001), the value of m was set to 8,000.

It is worth noting that both these formulations assume that auditory and visual information is independent. It seems counter-intuitive that this could be true of information from lip movements and the overall speech envelope. If the information were significantly correlated then the value of using both sources of information would be less than if it were independent, and the models would overestimate audio-visual performance. However, as we shall see, audio-visual performance always equals or exceeds the independent noise model. Thus, there is little indication that correlations across the modalities are affecting our results very much.

From these equations, given the audio-only and visual-only performance observed in the present study, it was possible to calculate a value of d'_{A} and d'_{V} and use them to predict performance for the AV conditions. These calculations were performed on the percentage correct at each SNR in each condition (reconstructed from the adaptive tracks), averaged across participants, for both Experiments 1 and 2, and compared with the observed data.

Thresholds were calculated from the model output by fitting a logistic function to the predicted AV psychometric functions, and the SRT was calculated as the 50% point on this function. Due to the nature of the adaptive track, psychometric functions were poorly sampled much below each participant's threshold. Therefore, fitting was limited to SNRs where model performance was 30% or greater. Confidence intervals were estimated for model SRTs by bootstrapping, simulating variability in performance at each SNR according to binomial statistics. Psychometric functions were each generated 1,000 times, assuming 300 trials for each SNR (consistent with the data across all participants), logistic functions were fitted, and SRTs were calculated for each iteration.

Results

Experiment 1

Panel A of Figure 2 shows the SRT50s for the audio-only condition for clear and vocoded speech, and Panel B shows the audio-visual SRT50s. A repeated-measures analysis of variance (ANOVA) indicated that SRTs were affected by the number of background talkers, $F(2,\ 44) = 161.01$, $p < .001$, $\eta^2_{\mathrm{p}} = 0.88$, and that this effect was mediated by whether speech was clear or vocoded, $F(2, 44) = 84.11$, $p < .001$, $\eta^2_{\mathrm{p}} = 0.79$. The interaction appeared to arise because performance with clear speech was affected to a greater degree by the number of background talkers, $F(2,\ 44) = 221.05$, $p < .001$, $\eta^2_{\mathrm{p}} = 0.91$, than performance in vocoded speech, $F(2,\ 44) = 3.34$, $p = .045$, $\eta^2_{\mathrm{p}} = 0.13$, as indicated by the difference in the observed effect sizes. As expected, SRTs were lower for clear speech (mean: −11.6 dB, *SD*: 6.6) than for vocoded speech (5.2 dB, *SD*: 4.6; $F(1, 22) = 881.14$, $p < .001$, $\eta^2_{\mathrm{p}} = 0.98$).

Average SRTs were lower (better) in audio-visual conditions than in audio-only conditions, $F(1, 22) = 184.70$, $p < .001$, $\eta^2_{\mathrm{p}} = 0.89$. The benefit received from the visual information was calculated by measuring difference in SRT50s between audio-visual and audio-only conditions. Figure 3 shows the average visual speech benefit for clear and vocoded speech, when there were 1, 2, or 16 background talkers. There was a significant effect of the number of background talkers on visual benefit, $F(2,\ 44) = 4.17$, $p < .05$, $\eta^2_{\mathrm{p}} = 0.16$. There was no overall significant difference between the level of visual speech benefit between clear and vocoded speech, $F(1,\ 22) = 0.34$, $p = .57$, $\eta^2_{\mathrm{p}} = 0.02$, and the interaction between these two main effects did not reach significance, $F(2,\ 44) = 1.17$, $p = .32$, $\eta^2_{\mathrm{p}} = 0.051$. The average visual-only score (not shown) was 1.56% correct (*SD*: 2.17).

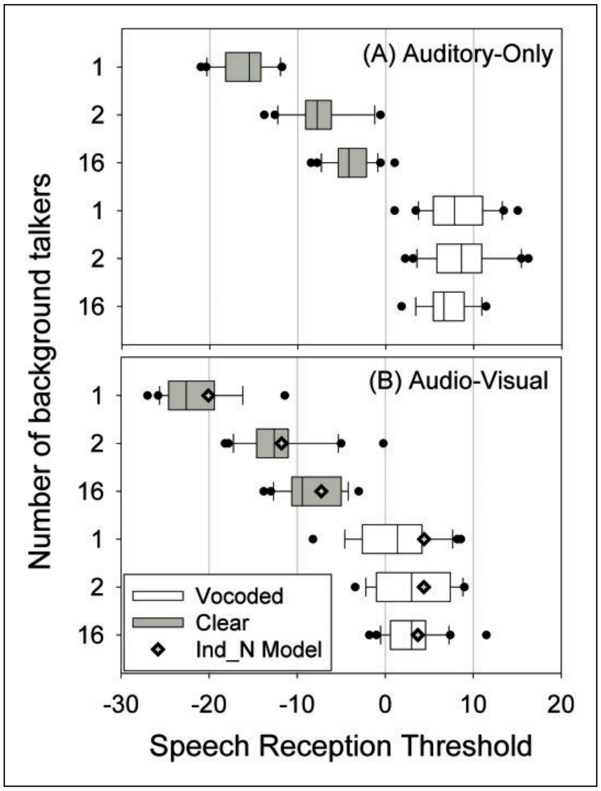

Figure 2. Audio-only (Panel A) and audio-visual (Panel B) speech reception thresholds for clear and vocoded speech. The rectangular boxes show the lower (25%) and upper (75%) quartiles of the data, with the solid line showing the median. The whiskers show the 10% to 90% range, and the black dots show outlier data which fall outside that range. Diamonds show the audio-visual thresholds predicted by the independent noise model.

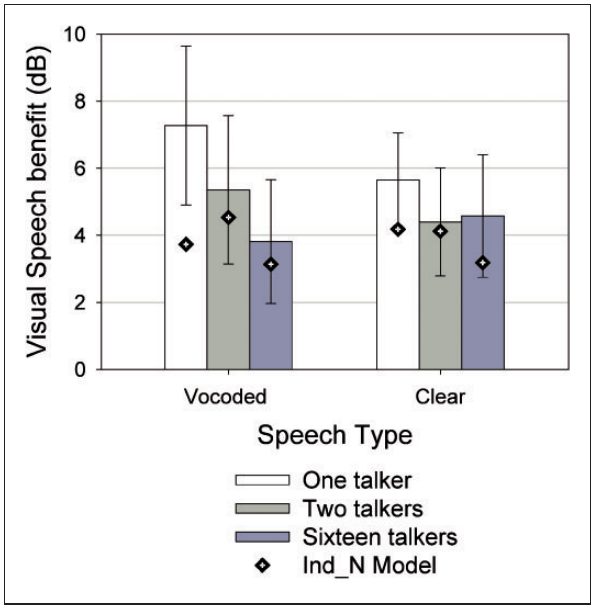

Figure 3. Mean visual speech benefit for each condition. Error bars denote 95% confidence intervals. Diamonds indicate audio-visual performance predicted from the independent noise model.

Figure 4 shows the predicted performance of two different SDT models (see Methods section) plotted over the raw percentage correct data, reconstructed from the adaptive tracks. The independent noise model, which assumes that each sensory modality provides a separate "noisy" source of information about the identity of the speech, closely approximated the data. The late noise model, in contrast, assumes that the information contained in each modality is noiseless (i.e., perfect), and that the limiting factor on performance occurs somewhere in the brain *after* the audio-visual integration. The late model predicted that audio-visual performance would be far higher than observed. The late noise model actually overpredicted performance to the extent that estimating SRTs from it was problematic (predicted AV performance was > 50%), so it was not considered further. The data appeared consistent with the expectations of SDT and a fairly optimal integration of information where the variabilities are independent unimodal sources. SDT predicts that the visual speech benefit corresponds to a minimum gain in performance of approximately 10% at many SNRs, even though visual performance alone is ~ 1.6%.

Figures 2 and 3 also show the audio-visual SRTs predicted by the independent noise model, and the model results in terms of the visual speech benefit. In all but two of the conditions, the model predictions for SRT50s lay within the interquartile ranges of the data, and in one case, they were only marginally outside. The independent noise model predicted the visual speech benefit from 2 and 16 background talkers, with the predictions falling within the 95% confidence intervals for the data. However, it underpredicted the larger visual speech benefit received for vocoded speech with a single background talker by a substantial margin (3.5 dB). This would imply that there is some change in the way that audio and visual information is integrated in this condition.

Experiment 2

Inspection of each participant's responses identified that seven participants found the auditory intelligibility of Talker 4 so poor that the adaptive tracks failed to converge on the 50% point in the audio-only vocoded condition. This clearly indicated that we had identified talkers with a range of intelligibility. However, it was problematic for statistical analysis. To maximize statistical power, but also to exclude thresholds from tracks that did not converge, we analyzed clear and vocoded speech separately. For clear speech, the analysis included all participants for all talkers, whereas the analysis of the vocoded conditions necessarily excluded data from

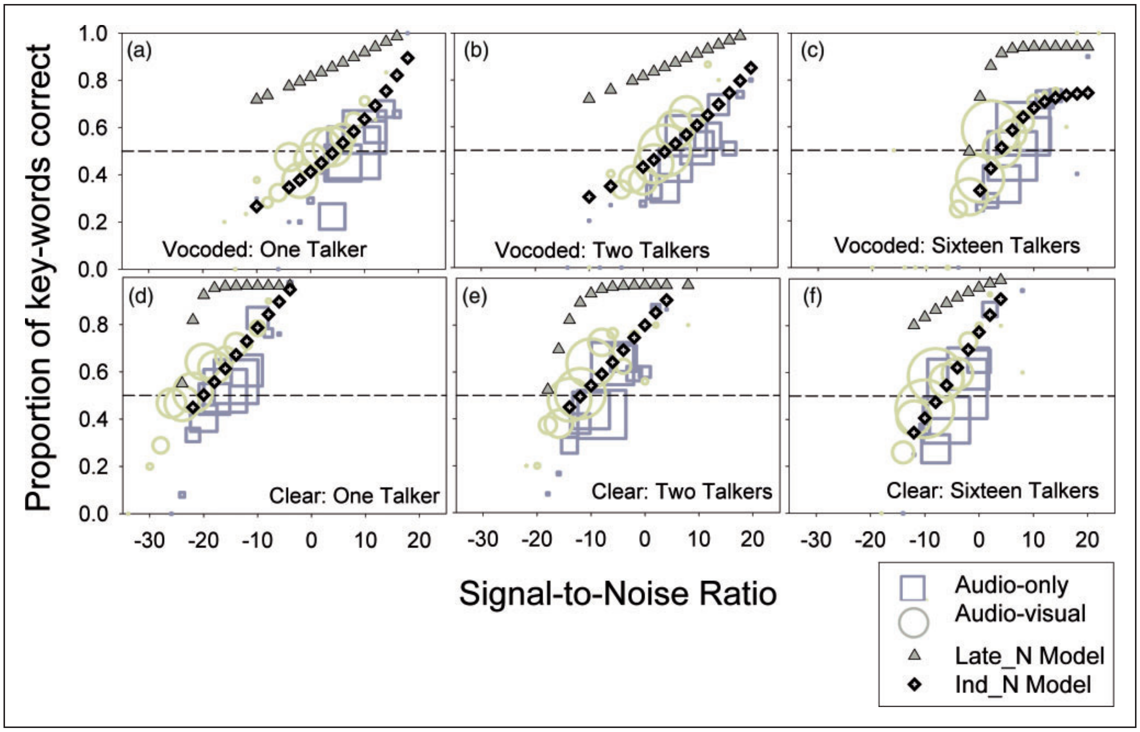

Figure 4. Bubble plots showing the proportion of key words correctly identified according to signal-to-noise ratio, collapsed across all participants. The larger the bubble, the more trials were presented at that particular SNR. Predicted audio-visual performance according to independent noise model is shown as diamonds, and predictions from the late noise model are shown as triangles.

Talker 4. A subsequent comparison of clear and vocoded performance was conducted by excluding Talker 4 from all conditions. A separate analysis of the SRTs from Talker 4, which excludes participants who had failed tracks, is presented as Supplementary Material.

The clear speech performance for each talker is shown in Figure 5 Panel A, and the vocoded speech performance is shown in Panel B. An ANOVA for the clear speech SRTs revealed a significant main effect of talker, $F(3,\ 69) = 251.77$, $p < .001$, $\eta_{\mathrm{p}}^2 = 0.92$. In line with the results from the pilot study, audio-only performance was best with Talker 1 (mean $\mathrm{SRT}_{50} = -14.7\,\mathrm{dB}$, *SD*: 2.0), and much more favorable SNRs were needed to understand Talker 4 (mean $\mathrm{SRT}_{50} = 0.3\,\mathrm{dB}$, *SD*: 5.2). A significant main effect of modality was observed, $F(1,\ 23) = 167.28$, $p < .001$, $\eta_{\mathrm{p}}^2 = 0.88$, such that performance was better in audio-visual than audio-only conditions. There was also a significant interaction between modality and talker, $F(3,\ 69) = 11.17$, $p < .001$, $\eta_{\mathrm{p}}^2 = 0.33$.

A similar pattern of results was found when the vocoded SRTs were subjected to a repeated-measures ANOVA. Performance levels varied according to the talker, $F(2,\ 46) = 117.88$, $p < .001$, $\eta_{\mathrm{p}}^2 = 0.84$: Performance was better with Talker 1 (mean SRT50 = −3.31 dB, *SD* = 3.37), and Talker 3 (mean SRT50 = −3.73 dB, *SD* = 3.84) than with Talker 2 (mean SRT50 = 1.95, *SD* = 3.76). Performance was also better in audio-visual than audio-only conditions, $F(1,\ 23) = 152.41$, $p < .001$, $\eta_{\mathrm{p}}^2 = 0.87$, but unlike clear speech conditions, the interaction between modality and talker was not significant.

The amount of visual speech benefit for each talker is shown in Figure 6. There was a significantly larger visual speech benefit when speech was vocoded (mean = 4.68, *SD*: 3.28) than when speech was clear (mean = 3.74, *SD*: 2.67; $F(1,\ 23) = 5.58$, $p < .05$, $\eta_{\mathrm{p}}^2 = 0.20$). As intended, there was a significant main effect of talker, $F(3,\ 69) = 10.55$, $p < .001$, $\eta_{\mathrm{p}}^2 = 0.32$, but no significant interaction between talker and speech type, $F(3,\ 69) = 0.84$, $p = .44$, $= 0.04$.

For clear speech conditions only, a significant main effect of talker was found, $F(3,\ 69) = 11.17$, $p < .001$, $\eta_{\mathrm{p}}^2 = 0.33$. Post hoc *t* tests with a Bonferroni correction revealed that significantly more benefit was obtained from Talker 2 than Talker 1, $t(23) = 5.40$, $p < .001$ or Talker 3, $t(23) = 3.49$, $p < .05$, and more benefit was obtained from Talker 4 than from Talker 1, $t(23) = 3.85$, $p < .01$, or Talker 3, $t(23) = 3.19$, $p < .05$. An analysis on vocoded speech SRTs revealed that the main effect of talker failed to reach significance, $F(2,\ 46) = 2.21$, $p = .12$, $\eta_{\mathrm{p}}^2 = 0.09$.

Visual-only performance levels were generally low, with some variability between the four talkers. Average

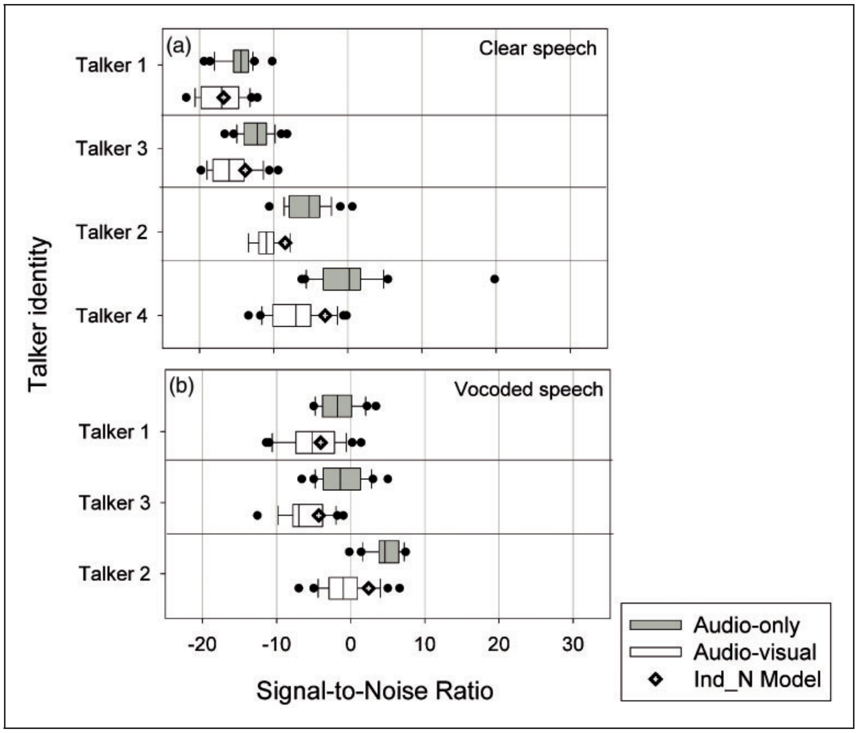

Figure 5. Audio-only and audio-visual speech reception thresholds for clear (Panel A) and vocoded (Panel B) speech for the four different talkers. Talkers have been ordered according to their intelligibility in the audio-only condition for clear speech. Diamonds indicate audio-visual performance predicted from the independent noise model. Three talkers are shown for vocoded speech due to the failure of the adaptive tracks for Talker 4 in the vocoded speech condition.

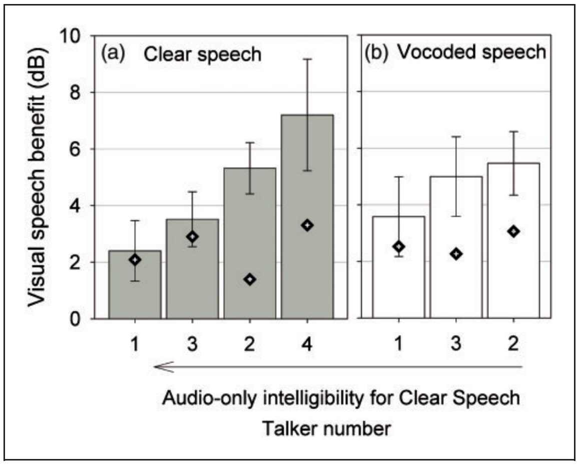

Figure 6. Levels of visual-speech benefit for each talker. Talkers have been ordered by their audio-only intelligibility for clear speech—Talker 1 was the most intelligible, and Talker 4 was the least intelligible. Diamonds indicate audio-visual performance predicted from the independent noise model.

performance using visual speech information alone varied from 2.0% for Talker 1 (*SD*: 2.7), 1.8% for Talker 2 (*SD*: 2.6), 1.4% for Talker 3 (*SD*: 2.1), and 0.6% for Talker 4 (*SD*: 1.1).

The predicted SRTs and visual benefit for the independent noise model are shown in Figures 5 and 6, respectively. As with Experiment 1, the late noise model generously overpredicted performance and so its predictions are not shown. The independent noise model again captured a qualitative benefit of visual speech across the conditions (Figure 4), but as before showed a tendency to underpredict the benefit. Predicted performance was within confidence intervals for Talker 1 in both conditions. However, the additive model underpredicted performance for the other three talkers, with nonoverlapping confidence intervals in at least one of the audio-processing conditions.

Discussion

The aim of the current experiments was to establish how the benefit that listeners receive from seeing a talking face varies in response to two manipulations that alter the nature of the task demands. It was hypothesized that degrading the speech signal by increasing the number of background talkers and selecting target talkers of decreasing auditory intelligibility would increase the benefit obtained from access to visual speech information, in line with the PoIE. Overall, the data present a mixed picture suggesting that audio-visual integration is dependent on several factors rather than following this simple principle.

In Experiment 1, for vocoded speech, there was significantly more visual speech benefit in the 1-talker background noise than the 16-talker noise. This was the only condition in which the SDT model, which assumed unimodal sources of internal noise, could not account for audio-visual performance. Overall, the visual benefit in vocoded conditions decreased as the number of talkers increased, which was also not predicted by the SDT model. This effect was not driven by any apparent change in task difficulty related to the choice of background "noise" as thresholds were similar regardless of the number of background talkers. In contrast, no difference was observed in the visual speech benefit for the clear speech stimuli as the number of background talkers was increased and those conditions were reasonably well accounted for by the same SDT model. This result was observed despite the fact that for clear speech, audio-only thresholds were affected strongly by the nature of the background noise in line with previous results (Rosen et al., 2013). Taken together, the data and model results suggest that the background noise and stimulus processing altered the demands of the task in such a way as to drive a change in the way audio and visual information is integrated, but not in the way predicted by the PoIE.

Although no benefit was found for having visual information in vocoded speech over clear speech for the target talker in Experiment 1, this was not the case for Experiment 2, where participants received greater visual speech benefit when speech was vocoded compared with when speech was clear. The results from Experiment 2 are consistent with the results from Stacey et al. (2016) who also found larger benefits for vocoded over clear speech. It is possible that the particular characteristics of the talker used in Experiment 1 meant that no effect was found. It is also possible that different talkers offer differing levels of visual speech benefit depending on whether speech is vocoded or not. For example, in Experiment 2, there was a difference of 1.8 dB between levels of visual speech benefit in clear speech for Talkers 2 and 3, but in vocoded speech, the amount of visual speech benefit differed between the same talkers by only 0.47 dB.

Experiment 2 demonstrated that the value of visual speech depends on the auditory intelligibility of the target talker: The most intelligible talker provided the least amount of visual speech benefit for both clear and vocoded speech and the least intelligible the greatest amount of benefit. These results were more in line with the predictions of the PoIE than those of Experiment 1, since poorer intelligibility was associated with greater visual benefit. As in Experiment 1, the independent noise SDT model could not predict this pattern of benefit and underpredicted performance overall.

The underprediction itself does not represent a fundamental failure of the models. Recall that the late noise model in both experiments predicted much higher levels of performance. Participants' performance was therefore overall intermediate between the two model extremes. The simplest interpretation of the data is therefore that subjects' performance was limited by a mix of independent and late-noise sources (although the data are closer to the independent case). This suggests that a model incorporating both independent and late-noise would account for the data slightly better, which was also the case in our previous study (Stacey et al., 2016). Unfortunately, as far as we are aware that there is no formulation for such an intermediate model.

In any case, the lack of any systematic difference in the independent noise model across conditions suggests that no single SDT model, with any particular mix of independent and late noise, could account for the pattern of results observed in Experiment 2. A mixed-noise model would increase the overall predicted visual speech benefit. The only way to model the present data is if the relative proportions of late and independent noise were to be different in different conditions. Therefore, the logical interpretation is that audio-visual integration is operating somewhat differently depending on the target talker.

The auditory intelligibility of the target talker may not have been the only contributory factor to the level of speech benefit received in Experiment 2. It is also possible the amount of visual information provided by each target talker varied (see Conrey & Gold, 2006) and would therefore contribute to variation in the level of visual benefit received. While participants tend to focus on the mouth area when trying to understand visual-only information (Lansing & McConkie, 2003), Conrey and Gold (2006) argue that it may be more useful to attend to other areas of the face to increase understanding of some target talkers. Therefore, the perceptual strategy of each participant for each talker may have affected the level of visual speech benefit they received. This idea is supported by the fact that visual-only intelligibility also varied across the target talkers. An additional benefit of evaluating the results in the light of the SDT model is that it takes account of these differences in visual performance. Accordingly, the pattern of visual benefit predicted by the model in clear listening conditions follows the differences in performance with only visual information available (Figure 6 cf. VO performance described in Experiment 2). This pattern is quite different to that of the visual benefits and suggests that differences in intelligibility based on visual information alone are unlikely to underlie the effects of talker observed under audio-visual conditions.

The results of Experiments 1 and 2 suggest that there are limitations to applying the PoIE to speech perception.

The principle did not apply equally across all experimental conditions, for the types of background noise used, and whether speech was clear or degraded by the use of vocoding. However, this observation may have resulted in part due to the use of adaptive procedures that equalized performance across conditions at 50%. Inspection of performance as a function of SNR (Figure 4) suggested that the benefit of visual information was maximal at negative SNRs where audio-only performance was poor (∼10%). Thus, a robust PoIE-compatible effect of varying task difficulty on visual speech benefit was observed and was also largely accounted for by simple SDT models. Further studies using nonadaptive procedures such as the method of constant stimuli could help to confirm these observations.

Overall, visual speech benefit is robustly dependent on intelligibility of the auditory stimulus, is maximal when performance is low (but not negligible), and a large part of this is consistent with a simple, optimal integration of information. However, if intelligibility is kept constant by measuring at a fixed performance level, some auditory conditions can still influence integration, and in a way that implies that the integration process is changing. We speculate that, consistent with the concepts of unimodal independent and later multimodal and potentially higher level sources of internal noise (Micheyl & Oxenham, 2012), this could reflect changes in the task demands between relying on low-level sensory or high level more cognitive or linguistic information.

Given the large variation in CI users' outcomes (Pisoni, Kronenberger, Chandramouli, & Conway, 2016) and the importance of implants for these patients' long-term health outcomes (Hilly et al., 2016; Vermeire et al., 2005) and quality of life (e.g., Hawthorne et al., 2004), it is important to ensure that performance is measured accurately and consistently in the presence of background noise and when visual information is available. An important implication of the results is that careful attention should be paid to the selection of stimuli to be used in research. This applies to the type of background noise and the intelligibility of the target talker. Replication of results across studies and comparison of performance across conditions within studies appear problematic if these factors are not controlled for. This concern may apply equally to assessments of CI users if it can be assumed that the pattern of results obtained using vocoding in these experiments approximates their performance.

Vocoder studies provide a valuable first step in understanding the benefits of visual speech information that might be obtained in a range of circumstances by users of implants. However, limitations of the vocoding technique used here should be acknowledged, as the technique simply simulates the consequences of removing TFS from the signal and filtering it into a number of discrete frequency bands. Other difficulties faced by CI users, such as spiral ganglion excitability (Horne, Sumner, & Seeber, 2016), among others, are not simulated. In addition, the tone vocoding used in the present study did not limit the range of modulations extracted from each channel, meaning that there will be F0 related modulations in the extracted envelopes. These stimuli will therefore have provided more information about F0 than would be accessible to typical CI users (Souza & Rosen, 2009). Future studies could remove these F0 cues by low-pass filtering the envelopes at 30 Hz, or by using a noise-excited vocoder to make these cues less salient.

Conclusions

The current experiments have shown that the amount of visual speech benefit gained varies according to the task demands. Specifically, the number of talkers in the background noise and the auditory intelligibility of the target talker have an impact on the extent to which people benefit from seeing a talking face. These effects were not predicted fully by a simple SDT model and suggest that the nature of audio-visual integration differs as task demands are varied. Overall, the results highlight the complexity of assessing and interpreting audio-visual speech perception abilities. Clinicians and researchers should consider the characteristics of their stimuli carefully when assessing audio-visual speech perception abilities. Further study of influencing factors and mechanisms of integration is required if we are to maximize the benefit of access to auditory and visual information in hearing-impaired people.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Catherine L. Blackburn http://orcid.org/0000-0003-0805-1059
Pádraig T. Kitterick http://orcid.org/0000-0001-8383-5318

Supplemental material

Supplemental material is available for this article online.

References

Bhargava, P., Gaudrain, E., & Başkent, D. (2016). The intelligibility of interrupted speech: Cochlear implant users and normal hearing listeners. *Journal of the Association for*

Research in Otolaryngology, *17*(5), 475–491. doi:10.1007/s10162-016-0565-9.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1101–1109. doi:10.1121/1.1345696.

Callan, D. E., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2001). Multimodal contribution to speech perception revealed by independent component analysis: A single-sweep EEG case study. *Cognitive Brain Research*, *10*(3), 349–353. doi:10.1016/s0926-6410(00)00054-9.

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport*, *14*(17), 2213–2218. doi:10.1097/00001756-200312020-00016.

Calvert, G. A. (2001). Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex*, *11*(12), 1110–1123. doi:10.1093/cercor/11.12.1110.

Conrey, B., & Gold, J. M. (2006). An ideal observer analysis of variability in visual-only speech. *Vision Research*, *46*(19), 3243–3258. doi:10.1016/j.visres.2006.03.020.

Dorman, M. F., Liss, J., Wang, S., Berisha, V., Ludwig, C., & Natale, S. C. (2016). Experiments on auditory-visual perception of sentences by users of unilateral, bimodal, and bilateral cochlear implants. *Journal of Speech, Language, and Hearing Research*, *59*(6), 1505–1519. doi:10.1044/2016_jslhr-h-15-0312.

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., & Kidd, G. (2003). Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. *The Journal of the Acoustical Society of America*, *114*(1), 368–379. doi:10.1121/1.1577562.

Eaves, J. M., Summerfield, A.Q., & Kitterick, P. T. (2011). Benefit of temporal fine structure to speech perception in noise measured with controlled temporal envelopes. *The Journal of the Acoustical Society of America, 130*(1), 501–507. doi: 10.1121/1.3592237.

Erber, N. P. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Research*, *40*, 481–492. doi:10.1044/jshd.4004.481.

Gagne, J. P., Masterson, V., Munhall, K. G., Bilida, N., & Querengesser, C. (1994). Across talker variability in auditory, visual, and audio-visual speech intelligibility for conversational and clear speech. *Journal-Academy of Rehabilitative Audiology*, *27*, 135–158.

Glasberg, B. R., & Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*(1–2), 103–138. doi:10.1016/0378-5955(90)90170-t.

Grant, K., & Walden, B. E. (1996). Evaluating the articulation index for auditory-visual consonant recognition. *Journal of the Acoustical Society of America*, *100*, 2415–2424. doi:10.1121/1.417950.

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

Hawthorne, G., Hogan, A., Giles, E., Stewart, M., Kethel, L., White, K., & Taylor, A. (2004). Evaluating the health-related quality of life effects of cochlear implants: A prospective study of an adult cochlear implant program. *International Journal of Audiology*, *43*(4), 183–192. doi:10.1080/14992020400050026.

Hilly, O., Hwang, E., Smith, L., Shipp, D., Nedzelski, J. M., Chen, J. M., . . . Lin, V. W. (2016). Cochlear implantation in elderly patients: Stability of outcome over time. *The Journal of Laryngology and Otology*, *130*, 706–711. doi:10.1017/s0022215116008197.

Horne, C. D., Sumner, C. J., & Seeber, B. U. (2016). A phenomenological model of the electrically stimulated auditory nerve fiber: Temporal and biphasic response properties. *Frontiers of Computational Neuroscience*, *10*, 8. doi:10.3389/fncom.2016.00008.

Kaiser, A. R., Kirk, K. I., Lachs, L., & Pisoni, D. B. (2003). Talker and lexical effects on audio-visual word recognition by adults with cochlear implants. *Journal of Speech, Language, and Hearing Research*, *46*, 390–404. doi:10.1044/1092-4388(2003/032).

Lander, K., & Davies, R. (2008). Does face familiarity influence speechreadability? *The Quarterly Journal of Experimental Psychology*, *61*(7), 961–967. doi:10.1037/e512682013-188.

Lansing, C. R., & McConkie, G. W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Perception & Psychophysics*, *65*(4), 536–552. doi:10.3758/bf03194581.

London, S., Bishop, C. W., & Miller, L. M. (2012). Spatial attention modulates the precedence effect. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(6), 1371–1379.

Lusk, L. G., & Mitchel, A. D. (2016). Differential gaze patterns on eyes and mouth during audio-visual speech segmentation. *Frontiers in Psychology*, *7*, 52. doi:10.3389/fpsyg.2016.00052.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. Cambridge, England: Cambridge University Press.

Markham, D., & Hazan, V. (2002). The UCL talker database. *Speech, Hearing and Language: UCL Work in Progress*, *14*, 1–17.

Meredith, M. A., & Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of Neurophysiology*, *56*(3), 640–662. doi:10.1152/jn.1986.56.3.640.

Micheyl, C., & Oxenham, A. J. (2012). Comparing models of the combined-stimulation advantage for speech recognition. *The Journal of the Acoustical Society of America*, *131*(5), 3970–3980. doi:10.1121/1.3699231.

Middelweerd, M. J., & Plomp, R. (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. *The Journal of the Acoustical Society of America*, *82*(6), 2145–2147. doi:10.1121/1.395659.

Miller, G. A. (1947). The masking of speech. *Psychological Bulletin*, *44*(2), 105. doi:10.1037/h0055960.

Müsch, H., & Buus, S. R. (2001). Using statistical decision theory to predict speech intelligibility. I. Model structure. *The Journal of the Acoustical Society of America*, *109*(6), 2896–2909. doi:10.1121/1.1371971.

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audio-visual speech perception. *Cortex*, *68*, 169–181. doi:10.1016/j.cortex.2015.03.006.

Pisoni, D. B., Kronenberger, W. G., Chandramouli, S. H., & Conway, C. M. (2016). Learning and memory processes following cochlear implantation: The missing piece of the puzzle. *Frontiers in Psychology*, *7*, 493. doi:10.3389/fpsyg.2016.00493.

Qin, M. K., & Oxenham, A. J. (2003). Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. *The Journal of the Acoustical Society of America*, *114*(1), 446–454. doi:10.1121/1.1579009.

Rosen, S., Souza, P., Ekelund, C., & Majeed, A. A. (2013). Listening to speech in a background of other talkers: Effects of talker number and noise vocoding. *The Journal of the Acoustical Society of America*, *133*(4), 2431–2443. doi:10.1121/1.4794379.

Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., & Weinstock, M. (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, *17*(3), 225–246. doi:10.1109/tau.1969.1162058.

Saki, N., Yadollahpour, A., Moniri, S., Karimi, M., Bayat, A., Abshirini, H., & Nikakhlagh, S. (2017). Investigating the impacts of cochlear implantation on the happiness and self-esteem of mothers of children with severe hearing loss. *International Journal of Mental Health and Addiction*, *15*(2), 288–294. doi:10.1007/s11469-016-9672-4.

Souza, P., & Rosen, S. (2009). Effects of envelope bandwidth on the intelligibility of sine- and noise-vocoded speech. *Journal of the Acoustical Society of America*, *126*(2), 792–805. doi:10.1121/1.3158835.

Stacey, P. C., Kitterick, P. T., Morris, S. D., & Sumner, C. J. (2016). The contribution of visual information to the perception of speech in noise with and without informative temporal fine structure. *Hearing Research*, *336*, 17–28. doi:10.1016/j.heares.2016.04.002.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215. doi:10.1121/1.1907309.

Turner, C. W., Gantz, B. J., Vidal, C., Behrens, A., & Henry, B. A. (2004). Speech recognition in noise for cochlear implant listeners: Benefits of residual acoustic hearing. *The Journal of the Acoustical Society of America*, *115*(4), 1729–1735. doi:10.1121/1.1687425.

Vermeire, K., Brokx, J. P., Wuyts, F. L., Cochet, E., Hofkens, A., & Van de Heyning, P. H. (2005). Quality-of-life benefit from cochlear implantation in the elderly. *Otology & Neurotology*, *26*(2), 188–195. doi:10.1097/00129492-200503000-00010.