The contribution of visual information to the perception of speech in noise with and without informative temporal fine structure

Paula C. Stacey[1] Pádraig T. Kitterick[2] Saffron D. Morris[3] Christian J. Sumner[3]

paula.stacey@ntu.ac.uk padraig.kitterick@ntu.ac.uk saffron90@live.co.uk chris@ihr.mrc.ac.uk

[1]Division of Psychology
Nottingham Trent University
Burton Street
Nottingham NG1 4BU

[2] NIHR Nottingham Hearing Biomedical Research Unit
Ropewalk House
113 The Ropewalk
Nottingham NG1 5DU

[3] MRC Institute of Hearing Research
University Park
Nottingham NG7 2RD

Abstract

Understanding what is said in demanding listening situations is assisted greatly by looking at the face of a talker. Previous studies have observed that normal-hearing listeners can benefit from this visual information when a talker's voice is presented in background noise. These benefits have also been observed in quiet listening conditions in cochlear-implant users, whose device does not convey the informative temporal fine structure cues in speech, and when normal-hearing individuals listen to speech processed to remove these informative temporal fine structure cues. The current study (1) characterised the benefits of visual information when listening in background noise; and (2) used sine-wave vocoding to compare the size of the visual benefit when speech is presented with or without informative temporal fine structure. The accuracy with which normal-hearing individuals reported words in spoken sentences was assessed across three experiments. The availability of visual information and informative temporal fine structure cues was varied within and across the experiments. The results showed that visual benefit was observed using open- and closed-set tests of speech perception. The size of the benefit increased when informative temporal fine structure cues were removed. This finding suggests that visual information may play an important role in the ability of cochlear-implant users to understand speech in many everyday situations. Models of audio-visual integration were able to account for the additional benefit of visual information when speech was degraded and suggested that auditory and visual information was being integrated in a similar way in all conditions. The modelling results were consistent with the notion that audio-visual benefit is derived from the optimal combination of auditory and visual sensory cues.

Keywords: audio-visual; visual speech; temporal fine structure; sine-wave vocoding; cochlear implants

1. Introduction

Speech perception in normal-hearing listeners is very resilient to distortions in the auditory signal and the presence of background noise. In contrast, understanding speech in background noise is difficult for adults with hearing impairment (Davis, 1989; Kramer *et al.*, 1998) and is particularly problematic for users of cochlear implants (CI) whose device degrades the spectral and temporal information in speech (Schafer and Thibodeau, 2004; Wolfe *et al*., 2009; Fu *et al.*, 1998; Skinner *et al.*, 1994). Shannon and colleagues (1995) showed that when signals were presented in quiet, listeners with normal hearing were able to tolerate a dramatic reduction in the amount of spectral and temporal information present in the speech signal before there was any appreciable effect on performance. The 'noise-vocoding' technique used by Shannon *et al.* (1995) involved: (1) dividing the speech signal into a limited number of frequency bands; (2) extracting the slow amplitude modulations or 'temporal envelope' within each frequency band; and (3) using these envelopes to modulate a wide-band random-noise carrier signal which was then filtered by the same filters used in stage (1). The use of a random-noise carrier has the effect of replacing the informative high-rate fluctuations in frequency near the centre-frequency of each band with non-informative fine structure. As the first two stages of this process mimic the processing stages implemented by a speech processor of a cochlear implant, vocoders have been widely used to investigate the difficulties experienced by users of cochlear implants.

The inability of cochlear implants to convey informative temporal fine structure cues has severe consequences for the ability of cochlear-implant users to perceive speech in the presence of background noise (e.g. Schafer & Thibodeau, 2004), and this difficulty has been replicated using noise-vocoding in normally-hearing individuals (Qin and Oxenham, 2003; Ihlefeld *et al.*, 2010, Rosen *et al.*, 2013). Qin and Oxenham (2003) investigated speech perception in noise with 4-, 8-, and 24-channel vocoders. Normal-hearing listeners were presented with IEEE sentences, and the signal-to-noise ratio (SNR) at which performance was 50% correct (known as the Speech Reception Threshold, SRT_{50}) was estimated by varying the relative levels of speech and noise. When speech was unprocessed and presented in single-talker background noise, participants could achieve 50% correct performance at an SNR of -10.3dB. When speech was then processed by an 8-channel vocoder, listeners required the level of the speech to be 6.4-dB higher than the noise to reach the same performance level. The addition of more spectral channels improved performance with the vocoder but a positive SNR (+0.7dB) was still required to report 50% of keywords correctly even in the 24 channel condition. Qin and Oxenham (2003) concluded that the reduction of pitch cues found in the temporal fine structure and low frequency harmonics of speech may be responsible for this performance detriment. Somewhat lower levels of susceptibility to the presence of noise have been reported for speech processed using a 'sine-wave vocoder' in which the informative temporal fine structure is replaced with sine waves rather than noise (Whitmal *et al.*, 2007). There is some evidence that sine-wave vocoders match the percept of cochlear-implant users more closely than noise-band vocoders (e.g. Dorman *et al.*, 1997) and are better at preserving the envelope fluctuations present in speech (e.g. Whitmal *et al.*, 2007; Dau *et al.*, 1999).

Although the impact of removing informative temporal fine structure cues has been studied extensively for audio-only situations, its impact on the audio-*visual* perception of speech in noisy conditions has received little attention, despite this being the more ecologically relevant problem. Sumby and Pollack's (1954) seminal work with normal-hearing adults showed that word recognition improved considerably under audio-visual conditions compared to listening to the audio alone. In fact, the addition of visual speech information was found to be equivalent to increasing the signal-to-noise ratio by +15dB compared with audio-only presentation. It is perhaps not surprising therefore that people with impaired hearing and users of cochlear implants gain considerable benefit from being able to see the faces of talkers (Erber, 1975; Kaiser *et al.*, 2003; Tyler *et al.*, 1997).

Kaiser *et al.* (2003) tested audio-only, visual-only, and audio-visual recognition of monosyllabic English words in both normal-hearing listeners and cochlear-implant users. Normal-hearing listeners were presented with words at -5dB SNR, and cochlear-implant users were presented with words in quiet. The results showed that both groups of listeners performed best in the audio-visual condition in which word recognition scores were similar in both groups. There was some evidence that cochlear-implant users made better use of visual information when listening conditions were more difficult, such as when they were required to identify lexically difficult words (low frequency words with many phonetic neighbours, Luce & Pisoni, 1998). More recent studies have added support to the idea that people with cochlear implants may be better at integrating auditory and visual information than normal-hearing listeners (Rouger *et al.*, 2007; Desai *et al.*, 2008).

A number of previous studies have found that benefits from visual speech information depend on the nature of the auditory signal. Grant *et al.* (1985, 1991, 1994) investigated the way in which different sorts of degraded speech signals combined with visual speech cues. More recently, McGettigan *et al.* (2012) demonstrated greater benefits from visual speech information for speech lacking in auditory clarity, such that visual speech information boosted performance more for 2- and 4-channel noise-vocoded speech than it did for 6-channel vocoded speech.

These studies lead logically to the idea that the value of any sensory input is not fixed, but can depend of the value or nature of another sensory input; i.e. the visual signal is of greater value when the auditory input is degraded. This is consistent with the 'Principle of Inverse Effectiveness' (Lakatos *et al.*, 2007, Tye-Murray *et al.*, 2010) which asserts that the value of one modality will increase as the value of another declines. A number of models have been proposed to try to explain the nature of multisensory integration (Massaro, 1987; Blamey *et al.*, 1989; Braida, 1991; Grant *et al.*, 1998; Kong and Carlyon, 2007; Rouger *et al.*, 2007; Micheyl and Oxenham, 2012). Models can be broadly categorised as to whether information is integrated in some raw sensory form before any decision is made ('pre-labelling') or after decision processes are applied separately to each modality ('post-labelling'; Braida, 1991; Peelle and Sommers, 2015).

Recently, Micheyl and Oxenham (2012) proposed a pre-labelling model based on Signal Detection Theory (SDT) to explain the capacity of normal-hearing listeners to integrate vocoded information in one ear with low-frequency acoustic information in the other ear. Their model and those applied in other similar studies suggested that the benefits of integrating electric and acoustic information can be explained as an additive interaction (Seldran *et al.*, 2011; Micheyl and Oxenham, 2012, Rader *et al.*, 2015) of the raw sensory information prior to any decision. Rouger *et al.* (2007) applied a post-labelling model to examine the properties of audio-visual integration, which assumes that decisions are made about individual cues prior to integrating these to make an overall decision. Their model is an extension of the 'probability summation model' (Treisman, 1998), which states that the probability of answering correctly is equal to the probability that either one or both of the modalities presented individually would result in the correct answer. Interestingly, Rouger *et al.*'s implementation of this model on their data suggested that integration across modalities operated differently in cochlear implantees and normal hearing subjects listening to noise-vocoded speech.

The current project systematically investigates the perception of sine-wave vocoded speech (labelled as ENV speech) at a range of SNRs, and compares this with performance in 'clear' speech conditions where informative temporal fine structure cues remain (labelled as TFS speech). The primary question of interest is whether the size of the benefit received from visual speech information depends on the presence of informative temporal fine structure information. This question was addressed using both open-set and closed-set tests of speech perception as we might expect to find differences between different types of speech tests (see Lunner *et al.*, 2012). Not only were we interested in whether any numeric improvement in performance with the addition of visual information depended on the presence of TFS, but also whether any observed differences implied a difference in the underlying integration process. Three experiments are presented below; in the first participants completed an open-set sentence test using a between participants design, the second reports an open-set sentence test using a mixed participants design, and the third reports a closed-set sentence test using a mixed participants design. Background noise consisted of multi-talker babble. In each experiment we expected to find that visual speech information contributed more to understanding vocoded speech in background noise than to understanding clear speech in background noise. These results were interpreted within the framework of a SDT model.

2. General methods

2.1 Apparatus

The presentation of stimuli and collection of responses was achieved using the EPrime software (Version 2.0, Psychology Software Tools Inc., Sharpsburg, US). Acoustic stimuli were presented over HD280pro headphones (Sennheiser, Wedemark, Germany) via a custom built digital-to-analogue converter. The presentation level of the acoustic stimuli was calibrated to achieve an average presentation level between 70-73 dB sound pressure level (SPL). Calibration was performed by coupling the headphones to an artificial ear (Brüel & Kjær Type 4153) using a flat-plate adaptor. Calibration measurements were made using a 0.5-inch pressure field microphone (Type 4192) connected to a sound level meter (Type 2260). Visual stimuli were presented on a computer-controlled visual display unit measuring 25.4cm high by 44.5cm wide positioned approximately 0.5m away from the participants and at head height.

2.2 Signal processing

Audio-visual sentence materials (IEEE sentences, IEEE, 1969) were processed using the Matlab programming environment (Mathworks, Nantick MA). The desired signal-to-noise ratio (SNR) was achieved by attenuating the stimulus (for negative SNRs) or a multi-talker babble (for positive SNRs) and summing before normalising the RMS of the composite signal. The composite signal was then band-pass filtered into 8 adjacent frequency bands spaced equally on an equivalent rectangular bandwidth frequency scale between 100 Hz and 8 kHz (Glasberg and Moore, 1990) using Finite Impulse Response filters. In experimental conditions that included informative temporal fine structure (TFS), the auditory stimuli were constructed by summing the output of the eight band-pass filters. In all other conditions (referred to as ENV), the temporal envelope of each filter output was extracted using the Hilbert transform and used to modulate a sine wave at the centre 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 regardless of whether the fine structure was informative (TFS conditions) or uninformative (ENV conditions) (Eaves *et al.*, 2013).

2.3 Procedure

Participants sat in a quiet room in front of the computer-controlled visual display unit. On each trial, a stimulus was selected randomly from the corpus of audio-visual sentence materials and the acoustic stimulus was presented over headphones while the visual display unit remained blank. In audio-visual conditions, a video showing the animated face of the talker uttering the same sentence was displayed simultaneously with the acoustic stimulus.

Four experimental conditions were defined by whether or not the processing preserved informative TFS (*processing* manipulation) and whether visual information was presented or not (*modality* manipulation). Stimuli were presented at a range of SNRs in each condition. The specific range of SNRs in any particular condition was chosen according to the stimulus materials used and the type of signal processing applied based on pilot testing in order to span the widest possible range of performance levels. The order of trials within each condition was randomised so that the SNR varied unpredictably from trial to trial.

A summary performance level was calculated for each SNR within each condition. The method of calculating the summary performance level varied across the experiments according to the materials used. A three- or four-parameter logistic function was fit to each participant's data using Matlab to describe the relationship between SNR and accuracy:

$$f(SNR) = a_{min} + \frac{(a_{max} - a_{min})}{1 + e^{-(SNR - x_0)/b}}$$

Where a_{max} and a_{min} are the asymptotic values of the function, x_0 is the mid-point of the function, and b is the slope of the function. For Experiments 1 and 2, a_{min} was always set to 0 to reflect the open-set nature of the speech perception task that was used. As we show in Section 6.2, performance in visual-only conditions is non-zero but very poor. The relatively small total number of key-words for each participant at each SNR (experiment 1: 50; experiment 2: 25) mean that small percentage differences cannot be resolved. In addition allowing the a_{min} parameter to vary to fit the data results in poorer fits. The fitted function was used to determine the SNR at which the participant achieved an accuracy of 50% correct (the Speech Reception Threshold, SRT_{50}), as follows:

$$SRT_{50} = x_0 - \ln\left(\frac{(a_{max} - a_{min})}{(0.5 - a_{min})} - 1\right)$$

3. Experiment 1

This experiment used an open-set test of speech understanding to test the hypothesis that the benefit from visual speech when listening in noise is larger when informative temporal fine structure is not available, such as in those who hear using a cochlear implant alone, compared to when informative TFS is available.

3.1 Methods

3.1.1 Participants

Twenty-eight students (9 male, age range 18-29 years) from the Nottingham Trent University took part. All reported having normal hearing, normal or corrected-to-normal vision, and spoke English as their first language. Ethical approval was granted by the Nottingham Trent University.

3.1.2 Stimulus materials

The audio-visual materials were 80 IEEE sentences spoken by a single male talker with a British accent. Each sentence contained 5 key words. An example sentence with the key words underlined is "The <u>slang</u> <u>name</u> for <u>all</u> <u>alcohol</u> is <u>booze</u>." The auditory stimulus had a sample rate of 44100 Hz with 16-bits of quantization. The corresponding video stimulus was recorded at 25 frames per second and measured 19cm high by 24cm wide on the visual display unit. Each sentence was approximately 3 seconds long.

3.1.3 Procedure

Each participant completed one of the four experimental conditions defined by the factorial combination of *processing* and *modality* manipulations, resulting in seven participants per condition. Pilot testing had indicated that the full range of performance levels could be spanned in most conditions by presenting the sentences at SNRs between -20 dB and +8 dB in 4-dB intervals. In the condition with auditory-only presentation and ENV speech, the range was adjusted as pilot testing indicated that participants required more favourable SNRs to achieve highly-accurate performance levels. In that condition, auditory stimuli were presented between -12 dB and +16 dB, with the first three participants being presented with stimuli between -16 and +12 dB. On each trial, participants were instructed to listen carefully to the sentence and repeat any words they could hear out loud. The experimenter recorded which words were correctly identified and participants initiated the next trial. A total of 10 sentences were presented at each SNR with each containing 5 key words. Performance at each SNR was summarised as the percentage of the 50 key words that were identified correctly.

3.2 Results and discussion

Figure 1 (Panel A) shows the percentage of key words identified correctly as a function of SNR in the four conditions of the main experiment, with three-parameter logistic functions fit to the average data. The pattern of the data confirmed that the experiment had been successful in spanning the full range of performance levels and also that the data were well-described by a sigmoidal function. As expected, the location of the function varied as a function of the availability of TFS and visual information. Figure 2 (Panel A) shows the SRT_{50}s for all conditions. Participants were able to report 50% of key words correct (the SRT_{50}) at highly-adverse SNRs when both visual speech and TFS information were available (mean -8.8 dB, s.d. 1.8) but required more favourable SNRs to achieve the same performance level when neither type of information was available (mean 3.4 dB, s.d. 3.2).

[INSERT FIGURE 1]

The average SRT_{50}s were subjected to an analysis of variance with between-subject factors of *processing* (TFS vs. ENV) and *modality* (auditory only vs. audio-visual). The analysis confirmed that the SNR required to reach an accuracy of 50% correct was influenced by the presence of both visual information ($F(1,24)=48.19$, $p<.001$, $\eta_p^2=.69$) and informative TFS ($F(1,24)=66.16$, $p<.001$, $\eta_p^2=.73$). The presence of visual speech information improved performance by a similar magnitude as the presence of TFS information, with an overall difference of 5.6dB between audio-visual and audio-only conditions, and an overall difference of 6.6dB between TFS and ENV conditions (Table 1).

[INSERT TABLE 1]

The benefit gained from the addition of visual speech in each condition is shown in Figure 3. The data did not support the hypothesis that visual information is more valuable when informative TFS is not available as no significant interaction was observed ($F(1,24)=3.07$, $p=.092$, $\eta_p^2=.11$). An analysis of the gradients of the fitted sigmoidal functions revealed no significant main effect of *processing* and no interaction, but slopes were marginally steeper in the audio-only conditions (mean slope at the 50%-correct point 19.6%/dB, s.d. 22.0) than in the audio-visual conditions (mean slope at the 50%-correct point 8.3%/dB, s.d. 2.8) ($F(1, 24) = 3.99$, $p=0.057$, $\eta_p^2=.14$).

The results are compatible with the idea that seeing the face of the talker provides additional cues that can aid speech understanding when acoustic information is degraded, whether by the presence of a background noise or by the unavailability of informative TFS. However, the lack of a significant interaction meant that the results did not support the hypothesis that visual benefit when listening in noise is larger for those listeners who do not have access to informative TFS information such as cochlear-implant users.

Post-hoc power analyses indicated that the experiment had sufficient power to detect the main effects of *processing* and *modality* (power > .99) but may have been underpowered to detect the interaction effect (power = .27). An additional experiment was therefore conducted which was powered prospectively to detect the interaction effect using a mixed experimental design in which the effect of modality was assessed within rather than between participants.

4. Experiment 2

This experiment sought to replicate the main effects of manipulating the availability of informative TFS and visual information observed Experiment 1 but was prospectively designed and powered to detect an interaction between the two manipulations. The experiment therefore tested the hypothesis that visual information is more beneficial in the absence of informative TFS than when it is present

4.1 Methods

4.1.2 Power calculation

An analysis of the results of Experiment 1 suggested that the size of the interaction effect, expressed in terms of number of standard deviations, was 0.38. Presuming a within-subjects correlation between auditory-only and audio-visual performance of 0.5, detecting an interaction effect of this size in a mixed experimental design with a power of .80 and α=.05 would require 16 participants (Faul *et al.*, 2007).

4.1.3 Participants

Sixteen students from the Nottingham Trent University, who had not participated in Experiment 1 (3 male, age range 18-23 years) took part. All reported having normal hearing, normal or corrected-to-normal vision, and spoke English as their first language.

4.1.4 Procedure

The procedure was similar to that of Experiment 1. The SNR of the sentences was varied between -20 and +8 dB in 4-dB intervals except in the condition without either informative TFS or visual information, in which the SNR was varied between -12 dB and +16 dB in 4-dB intervals for all participants. Participants were presented with 5 sentences at each SNR rather than 10 as used in Experiment 1. The factorial combination of *processing* (TFS vs ENV) and *modality* (auditory-only vs audio-visual) defined four conditions. The modality of the stimuli was varied within participants while the type of processing applied was varied across two groups of eight participants. The scoring of responses and analysis of performance was identical to that used in Experiment 1.

4.2 Results and discussion

The overall pattern of results was found to be very similar to that of Experiment 1 (Figure 1, panel B). The manner in which average performance varied as a function of SNR was well-described by a sigmoidal function, whose place was similarly affected by both the type of processing applied to the auditory stimulus and the availability of visual information. An analysis of variance on SRT_{50}s (Figure 2) confirmed a significant effect of both *modality* ($F(1,14)=100.21$, $p<.001$, $\eta_p^2=.88$) and *processing* ($F(1,14)=105.30$, $p<.001$, $\eta_p^2=.88$). As in Experiment 1, visual speech information and TFS cues impacted on SRT_{50}s to a similar degree (Table 1).

[INSERT FIGURE 2]

Unlike in Experiment 1, the interaction term was found to be significant ($F(1,14)=5.30$, $p=0.038$, $\eta_p^2=.27$; Figure 2). Inspection of the data confirmed that the effect of providing visual information was larger when informative TFS was not available (Figure 3). SRT_{50} decreased from -5.6 dB to -9.2 dB with the provision of visual information in the TFS condition (mean change 3.5 dB, s.d. 1.7), and from 1.3 dB to -4.4 dB with the provision of visual information in the ENV condition (mean change 5.7 dB, s.d. 2.0). An analysis of the gradients of the logistic functions revealed no significant main effects or interactions.

[INSERT FIGURE 3]

The results of Experiment 2 supported the hypothesis that the benefits of visual information are larger when speech is lacking in informative TFS. This finding is compatible with the idea that visual information may be more beneficial for those who listen exclusively through a cochlear implant. When listening in noise, the absence of informative TFS can hinder the ability to identify the target talker based on vocal characteristics and also to segregate speech from background noise based on cues such as periodicity (Moore, 2008). Listeners who cannot access TFS cues experience severe difficulties with understanding speech in noise are therefore more likely to benefit from exploiting the additional information and redundancy provided through visual cues.

5. Experiment 3

Using an open-set test of speech perception, Experiments 1 and 2 demonstrated that the visual information provided by a talker's face can aid speech perception both when speech is degraded by background noise and when it is processed to remove informative TFS cues. It is possible that the contribution of TFS and visual speech cues may vary between open and closed-set tests due to differences in the predictability of the target stimuli. For example, Lunner *et al.* (2012) found larger benefits from TFS information for their young normal-hearing participants when they were presented with open-set tests of speech perception than when they completed a closed-set test. Therefore, the current experiment sought to establish whether the effects observed in Experiments 1 and 2 generalise to a closed-set test of speech perception using stimuli recorded by a different talker.

5.1 Method

5.1.2 Power calculation

No data were available with which to conduct a power calculation to determine how many participants would be required to detect the interaction between modality and processing on a closed-set test. The previous power calculation for experiment 2 indicated that 16 participants would be required for an open-set test where the effect size for the interaction was estimated to be 0.38. As it was unclear whether this effect size would be larger or smaller for a closed-set test, twenty participants were recruited which was sufficient to detect an effect as small as 0.34 with a power of .80 and α=.05.

5.1.3 Participants

Twenty students (2 male, age range 18-25 years) from the Nottingham Trent University took part. All reported having normal hearing, normal or corrected-to-normal vision, and spoke English as their first language.

5.1.4 Stimulus materials

The closed-set materials were 160 sentences from the GRID corpus produced by the University of Sheffield (Cooke *et al.*, 2006). Each sentence took the form "Put <u>Colour</u> at <u>Letter Number</u> now." An example sentence is "Put <u>Blue</u> at <u>G 9</u> now". A single female talker with a northern British accent was selected from the set of available talkers; this talker was of average intelligibility according to the audio-only intelligibility tests carried out by Cooke *et al.* (2006). The auditory stimulus was recorded at a sample rate of 25,000Hz with 16-bits of quantization. The corresponding video stimulus was recorded at 25 frames per second. Each sentence was approximately 3 seconds long. The 160 sentences selected incorporated the 10 most difficult letter words to identify based on pilot testing.

5.1.5 Procedure

The procedure was similar to that of Experiments 1 and 2. The SNR of the sentences was varied between -24 and +4 dB in 4-dB intervals except when neither visual information nor informative TFS was available. In that condition, the SNR was varied between -16 and +12 dB to avoid floor effects at multiple SNRs. Ten sentences were presented at each of the 8 SNRs providing 80 trials in both the auditory-only and audio-visual conditions. After a set of 10 practice trials, participants were presented with the 160 sentences in a random order. The type of processing (TFS or ENV) was varied between two groups of 10 participants.

On each trial, participants were instructed to listen carefully to the sentence and to use a computer mouse to select the correct letter word from a matrix of possible options. The matrix was shown on the visual display unit after the stimulus had ended. They were also asked to identify the number word from 5 alternatives. Pilot testing had indicated that performance on this secondary task approached ceiling and it was included to ensure that participants were attending and listening to the sentences throughout. The experiment took approximately 20 minutes to complete. Performance was summarised as the percentage of sentences on which the correct letter word was identified at each SNR.

5.2 Results and discussion

In general terms, the results of Experiment 3 were similar to those of Experiments 1 and 2. Figure 1(Panel C) shows the average performance at each SNR for the auditory-only and audio-visual materials in the TFS and ENV groups. An analysis of variance on SRT_{50}s confirmed the main effects of *modality* ($F(1,18)=16.61$, $p<.001$, $\eta_p^2=.48$) and *processing* ($F(1,18)=34.80$, $p<.001$, $\eta_p^2=.66$) but the interaction failed to reach significance ($F(1,18)=3.63$, $p=.073$, $\eta_p^2=.17$). Table 1 shows that the overall difference between audio-visual and audio-only conditions was numerically smaller (2.5dB) than the difference between TFS and ENV conditions (4.36dB). While performance in all conditions was well-described by a logistic function, as in Experiments 1 and 2, the slope of the function was less steep in conditions where visual information was provided (mean audio-visual slope 6.3%/dB, s.d. 5.6; mean auditory-only slope 16.7%/dB, s.d. 20.6) ($F(1,18)=7.59$, $p<.05$, $\eta_p^2=.30$). Further analyses of the function gradients revealed no other main effects or interactions.

The contribution of TFS and visual speech information was calculated individually for each of the 10 letter words participants were presented with. Data were collapsed across -16 to +4 dB SNRs (as these were used in all conditions) in order to give the overall proportion of letter words correct. The top panel of Figure 4 shows that TFS information benefitted the recognition of all the letter words, with particularly large benefits for 'D', 'G', 'L', and 'Z'. A 10 (*letter word*) x 2 (*processing*) mixed ANOVA on overall performance in the Auditory-Only condition revealed a significant main effect of *letter word* ($F(9,162)=15.13$, $p<.001$, $\eta_p^2=.46$) confirming that some words were easier to identify than others, a main effect of *processing* ($F(1,18)=86.98$, $p<.001$, $\eta_p^2=.83$) such that overall performance was better with informative TFS, and a marginally significant interaction ($F(9,162)=1.93$, $p=.051$, $\eta_p^2=.097$). Post-hoc t-tests with a False Discovery Rate (FDR) correction for multiple comparisons revealed that performance was better in the TFS condition for all letter words except 'I', 'N', and 'Q'.

[INSERT FIGURE 4]

The bottom panel of Figure 4 shows the visual benefit for each letter word in TFS and ENV conditions. For the TFS condition, there was significant visual benefit for 'J' and 'N', while for the ENV condition there was significant visual benefit for 'D', 'I', 'J', 'S', and 'U'. A 10 (*letter word*) x 2 (*processing*) mixed ANOVA on visual speech benefit revealed a significant main effect of *letter word* ($F(9,162)=4.40$, $p<.001$, $\eta_p^2=.20$) confirming that some words benefitted more from visual speech than others, a main effect of *processing* ($F(1,18)=4.42$, $p<0.05$, $\eta_p^2=.20$) such that there was overall more benefit from visual speech for the ENV condition, and a marginally significant interaction ($F(9,162)=1.90$, $p=.055$, $\eta_p^2=.096$). Post-hoc t-tests with FDR correction revealed that the only significant difference in visual speech benefit between TFS and ENV was for the letter word "L", where performance was poorer with visual speech information in the TFS condition.

The results of Experiment 3 were broadly similar to the previous experiments in confirming the beneficial nature of visual information and informative temporal fine structure when reporting words embedded in sentences spoken in the presence of background noise. The benefit from visual information was also found to be numerically greater in ENV than in TFS conditions. To examine the consistency of this interaction effect and to better estimate the true size of the additional benefit of visual information without informative TFS, the results from the three experiments were subject to a random-effects meta-analysis. The analysis indicated that heterogeneity, expressed in terms of the ratio between the total heterogeneity and total variance, was low ($I^2=0\%$) and not significant (Cochran's $Q(2)=0.16$, $p>.05$), indicating that the size and variability of the effect was similar across the three experiments. The pooled estimate of the size of the additional benefit that visual information provides in the ENV compared to TFS condition was 2.3 dB and was found to be significantly greater than zero (95% confidence interval 1. to 3.6 dB; Figure 5). This meta-analysis suggests that visual information contributes significantly more to speech understanding in noise when informative TFS information is not available, akin to the input to cochlear-implant users, compared to when informative TFS cues are available as in normal-hearing listeners.

[INSERT FIGURE 5]

6. Modelling the audio-visual interaction

The meta-analysis of Experiments 1 to 3 suggests that there is a modest but consistent increase in benefit from visual information when acoustic signals are degraded: introducing visual information lowers (improves) SRT_{50}s to a greater degree when informative TFS information is not available compared to when it is available. One possible explanation for the increased utility of visual information when auditory information is degraded is that listeners integrate information more efficiently in some way under these adverse conditions. An alternative explanation is that performance differences arise naturally from the way that the two sources of information are combined. The plausibility of these differing explanations was explored by re-analysing the data from Experiments 1 to 3 using two different types of decision models based on signal detection theory, and a model based on probability-summation.

6.1 Methods

Signal detection theory (SDT) considers that a sensory decision must be made on the basis of one or more noisy sensory variables (Green and Swets, 1966). In SDT, the discriminability of two different signals depends on the both the mean difference between sensory variables for the two stimuli and the trial-to-trial variability (or 'noise'). The proportion of correct trials that an observer will achieve when presented with stimuli in a single modality can be expressed as a function of the *overall discriminability*, *d'*, of the *m* different stimulus categories that are presented:

$$P = \int_{-\infty}^{+\infty} \phi(z - d')\, \Phi^{m}(z) dz \qquad \text{(eqn. 1)}$$

where $\phi(.)$ is the standard normal probability density function and $\Phi(.)$ is the cumulative standard normal function. This approach can be extended to multiple sources of information such as auditory and visual speech used in the present experiments. There are many ways information could be combined. Here we adopt a previously described model for combining such information (see Micheyl and Oxenham, 2012).

In SDT, the variability of the sensory representation is in part considered to be due to 'internal' noise. In the case of multiple sources of information, noise can arise both before ('independent noise') and after ('late noise') integration (but still prior to any decision; i.e. pre-labelling). These different sources of noise affect the integration process in different ways. The equation below assumes that raw sensory information is combined prior to arriving a decision (Braida 1991), and that noise arises in the observer's internal representation of both the auditory and visual stimuli independently before the sources of information are integrated (Micheyl and Oxenham, 2012) [1]:

$$P = \int_{-\infty}^{+\infty} \phi\left(z - \sqrt{(d'_A)^2 + (d'_V)^2}\right) \Phi^m(z)dz \qquad \text{(eqn. 2) independent noise model}$$

where d'_A and d'_V represent the overall discriminability of the auditory and visual stimuli respectively. An alternative assumption is that noise arises in the observer's internal representation of the audio-visual stimulus after the information in the two modalities has been combined (the so-called 'late noise' model). This 'late noise' model can be expressed through a further revision of Equation 2, as follows:

$$P = \int_{-\infty}^{+\infty} \phi(z - (d'_A + d'_V))\, \Phi^m(z)dz \qquad \text{(eqn. 3) late noise model}$$

Following Micheyl and Oxenham (2012), Equations 2 and 3 represent the extreme cases where one source of internal noise dominates; i.e. all noise is assumed to arise before (Equation 2) or after (Equation 3) integration. Following previous studies that have suggested that open set speech perception is best modelled as dependent on vocabulary size (Musch and Buus 2001), the value of *m* in Experiments 1 and 2 was set to 8000. For Experiment 3, *m* was set to 10 to reflect the number of possible response options on the closed-set test of speech discrimination.

To examine the capacity of the SDT noise models to explain the pattern of performance observed across the three experiments, Equations 2 and 3 were used to generate predictions for performance in the AV conditions. Predictions with and without informative TFS information were generated at each SNR and independently for each experiment. As equations 2 and 3 require data on Visual-only (VO) performance, an additional 10 participants (age range 21-71 years, 7 male) from the MRC Institute of Hearing Research were recruited in a supplemental experiment. They completed both the open-set sentence test (from Experiments 1 and 2) and the closed-set test (from Experiment 3) in an order counterbalanced across participants. For the open-set test, participants were asked to attend

[1] We chose this model because it is often superior to the alternative late-integration ('post-labelling') models, whereby a decision of sorts is arrived at for each modality independently, and then subsequently combined for a final decision. We will also only consider the case where sensory variables from the two modalities are only combined additively. In other words, a decision will be made on the basis of a linear (potentially weighted) sum of the noisy sensory variables from both modalities.

carefully to each sentence and report any words they could perceive. Participants were presented with 80 IEEE sentences, leading to a total of 400 key-words per participant. For the closed-set test, participants were presented with 80 GRID sentences, which incorporated 8 of each of the 10 consonant sounds that were used.

The value of the parameters d'_A and d'_V in Equations 2 and 3 were therefore computed directly from the AO and VO conditions using Equation 1, with the performance level P at a particular SNR set to the observed mean performance level in the data. The ability of one model to generate accurate predictions of AV performance within a single experiment could be interpreted as evidence that a particular model of audio-visual integration better reflects the underlying decision processes adopted by listeners. Performance intermediate to the two models would suggest a mix of unisensory and crossmodal noise sources. Performance outside of the extremes of the two models would imply either a supra-additive, or sub-additive combination of sensory information.

The results were also modelled using Rouger *et al.*'s (2007) extension of the 'probability summation model' (Treisman, 1998). The probability summation model states that the probability of answering correctly is equal to the probability that either one or both of the modalities presented individually would result in the correct answer. Formally this can be written:

$$P = P_{AO} + P_{VO} - P_{AO}P_{VO} \qquad \textit{(eqn. 4)}$$

where P_{AO} and P_{VO} are the probability of answering correctly in the AO and VO conditions. Rouger *et al.* generalised this model to one in which there were an arbitrary number of independent unisensory 'cues' and that overall probability of answering correctly was equal to the probability that T or more of those cues would be correctly identified. The case where $T=1$ corresponds to equation 4, and provides the lower bound for this kind of model. They term this the 'minimal integration' model since it assumes that auditory and visual information are evaluated as independent single sources of information. This family of models fall into the post-labelling category since integration is modelled as the combination of the probability of correct decisions. Note that this model cannot work with a closed set. For eqn. 4, in Experiment 3 chance performance is 10% and it predicts 19.9%.

The goodness of fit of each model to each experiment was assessed using a X^2 test between the data and each of the models (Table 2). To indicate whether the data was significantly different from a resulting model, we performed bootstrap simulations of a simple version of the fitted model (Langeheine *et al.* 1996). In a single simulation, for each AV condition (SNR, TFS vs. ENV), numbers were drawn from a binomial distribution with a probability corresponding to the fitted model value and sample size corresponding to that point in the data. From the number of correct and incorrect trials in each condition we computed X^2 of these simulated values against the mean model output. This gave the goodness of fit for a single simulated run of the model against mean model values. Repeating this simulation of the model many (5000) times yielded a distribution of X^2 values, and the likelihood (i.e. p-value) of observing a given goodness of fit under the assumption that the model was correct. From this were able to compute the likelihood of observing the data if the model were correct.

[INSERT TABLE 2]

6.2 Results and discussion

The average visual-only performance for the open-set IEEE test was 2.85% key-words correct (s.d. 3.20), and was 10.8% (s.d. 3.5) letter-words correct in the closed-set GRID test.

The two variants of SDT models were evaluated by their ability to predict the AV condition, given the performance in the AO and VO conditions. The results of applying the models revealed that the observed AV performance for ENV and TFS conditions in Experiments 1 and 2 lay between the 'independent' and 'late' noise SDT models (Figure 6, Panels A and B, see Table 2 for mean signed errors and X^2). The Rouger model, applied directly to the data with no fitting of the parameters (T=6, as in Rouger *et al.* 2007), provided a reasonable qualitative fit to all the conditions in Experiments 1 and 2.

[INSERT FIGURE 6]

Both models under predicted AV performance in Experiment 3 for both the ENV and TFS conditions by ~8% (Figure 6, Panel C and Table 2). This result stemmed from the fact that performance in the VO condition of Experiment 3 did not exceed chance levels. Therefore, no further evaluation of modelling Experiment 3 was conducted (see discussion).

Figure 2 shows the fits of the models to the data in terms of SRT_{50}s. Table 2 provides X^2 goodness of fit and estimates of the likelihood of the model being correct. Both SDT models are significantly different from the data, implying an intermediate model would be required to explain both TFS and ENV data. Thus, the data in both TFS and ENV conditions appear to be consistent with the optimal combination of auditory and visual information, and may result from a mixture of independent and late noise sources. The visual benefit varied from -0.6dB to -3.1dB (see Table 2) and the size of the observed visual benefit did not exceed that predicted by the purely-additive SDT models of integration. The data are also reasonably consistent with the post-labelling model proposed by Rouger *et al.*, even using the exact same model parameters as they did, although this model is nevertheless not a perfect fit to the data ($p<0.05$, Table 2). Thus, overall no models can account completely for the data. However, qualitatively they suggest that the way in which acoustic and visual information is combined is similar for acoustic input with and without informative TFS, whether assessed in the light of pre-labelling or post-labelling models.

7. General discussion

The current series of experiments investigated the benefits obtained from visual speech information when listening to degraded speech in background noise. The results show that the availability of visual speech information improves the understanding of speech with and without informative TFS; i.e. listeners were able to tolerate more noise in the signal when visual speech information is present. In addition, the present results suggest that the size of the benefit from visual speech information is greater, by roughly double the amount, when informative TFS is not available. This pattern of results was found to be consistent across different experimental designs (between or mixed groups), speech tasks (open vs closed set), and stimuli.

7.1 Effects of visual speech and TFS information

In the open-set experiments reported in Experiments 1 and 2, the size of the benefit received from TFS and visual speech information are similar in magnitude. In Experiment 1, when combined across AV and AO modalities, the SRT_{50} was 6.6dB lower for TFS than for EVV speech. This compares with a difference of 5.6dB between audio-visual and audio-only conditions when combined across TFS and ENV speech types. For Experiment 2 the speech processing difference was 5.8dB compared with 4.6dB for the modality difference. These figures reinforce the importance of visual speech information when processing speech in background noise. The difficulties faced by cochlear-implant users are well documented, and many studies have demonstrated the poor performance of normal-hearing participants when TFS information is removed in vocoder simulations, especially when listening in background noise (Qin and Oxenham, 2003; Ihlefeld *et al.*, 2010, Rosen *et al.*, 2013). However, the importance of visual speech information when listening to degraded speech in background noise has received little investigation. Therefore, in order to truly reflect the performance of listeners in demanding situations, the role of visual speech information needs to be taken into account.

A strength of the current series of experiments is that we have demonstrated similar effects of visual speech and TFS information across open- and closed-set tests of speech perception. This is important as some research (e.g. Lunner *et al.*, 2012) has shown that the importance of TFS information may vary according to the type of speech test used. Consistent with the predictions from Lunner *et al.* (2012) we did find numerically smaller benefits of visual speech information and TFS cues in Experiment 3, where the choices presented to participants reduced uncertainty, and perhaps also reduced the usefulness of TFS cues and visual speech information.

The closed-set test also allowed us to look more closely at which stimuli in particular benefitted from visual speech and TFS information, with some letter words being more affected than others. Specific letter words that benefitted from TFS information included ‘D’, ‘G’, ‘L’, and ‘Z’, and the letter words ‘J’ and ‘N’ benefitted most from visual speech information. However, due to limitations in the nature of the stimuli (being letter words and not consonant sounds), a full phonetic analysis was not possible. Future research with consonant sounds would allow an information transfer analysis (Miller & Nicely, 1995) to be performed, which would enable an analysis of the extent to which different speech sounds (e.g. place, manner, and voicing) were transmitted to the listener. This would reveal further insights into the way in which visual speech and TFS cues interact for different features under noisy speech conditions that were not possible to perform using data from the current study.

7.2 Visual-only performance

Visual-only (VO) performance was also tested for the open-set IEEE sentences used in Experiments 1 and 2, and for the closed-set GRID test used in Experiment 3. The average VO performance was 2.85% keywords correct for the IEEE sentences and was 10.8% consonants correct for the GRID sentences. The average performance levels for the IEEE sentences demonstrates the fact listeners were on average able speechread some information from the sentences, although to a limited extent. Altieri, Pisoni, and Townsend (2011) found much higher levels of performance for a group of young normal-hearing participants when given the CUNY sentence test (Boothroyd *et al.*, 1988); participants reported an average of 12.4% of words correct (standard deviation 6.67%). Higher levels of performance are however to be expected for CUNY sentences as they are semantically and syntactically more predictable than IEEE sentences. The average VO performance of 10.8% on the closed-set GRID sentences reflects the fact that participants were not able to lipread the target letters at a level above chance (given that there were ten response options). Part of the difficulty with these tasks is that visual speech reading performance is challenging and participants may well have struggled to maintain motivation. In all experiments VO conditions were performed as a separate block. For the open-set task, verbal responses were recorded by an experimenter present in the sound booth, and we can be sure that the participants were engaged appropriately in the task. For the closed-set task, responses were made via a computer in isolation in a sound booth, making it difficult to monitor task engagement. Motivation was less likely to be a problem in AO or AV conditions, since the overall performance was higher. Consistent with this interpretation, asymptotic performance at the lowest SNRs in the AV conditions was considerably higher than chance, whilst AO conditions were not.

7.3 The nature of multisensory integration

Although there is a significant numerical advantage of visual speech information for ENV speech, this advantage is consistent with models which assume that visual information is integrated in a consistent way and regardless of whether TFS is available or not.

The results from the SDT models are consistent with previous research that has modelled the advantages that arise from receiving combined electrical and residual acoustic stimulation (Seldran *et al.*, 2011, Micheyl and Oxenham 2012, Rader *et al.*, 2015). In fact, the diversity in the balance between independent and late noise is also seen across other experiments (Micheyl and Oxenham 2012). In addition, using Braida's (1991) pre-labelling model of integration, Grant *et al.* (2007) showed that normal-hearing and hearing-impaired listeners exhibited a similar degree of integration efficiency of auditory and visual information. These findings therefore imply that the larger body of data on audio-visual integration in conditions of normal, undegraded speech (e.g. Tye-Murray *et al*., 2010; Sumby and Pollack, 1954), and studies of audio-visual integration in hearing impaired listeners (e.g. Grant *et al.*, 1998; Grant *et al.*, 2007) may well apply to degraded speech conditions and perhaps to users of cochlear implants.

Our data for both ENV and TFS speech were also well explained by the model used by Rouger *et al.* (2007). The finding that Rouger *et al.*'s model fit our data for the vocoded speech condition is inconsistent with their data which suggested that compared with cochlear-implant users, normal-hearing participants integrated sub-optimally when listening to noise-vocoded speech. However, given that Rouger's model fits our data well, it is clear that the differences in conclusions reflect differences between their data and ours; while the normal-hearing participants who listened to vocoded speech integrated sub-optimally in Rouger's study, our normal-hearing participants displayed optimal integration of auditory and visual information.

The models failed to predict the data for Experiment 3. However, performance in the VO condition here was very close to chance. Since d'~0, we would not expect *any* model of integration to predict the AV performance, which was improved over AO conditions, albeit only slightly overall. This could indicate some fundamental limitation of such models. However we think it more likely that it reflected poor motivation for the AO conditions in Experiment 3, as discussed above.

Finally, we note that although our data are consistent with a mixed noise source additive-SDT model, we do not know of an analytical equation similar to Equations 2 and 3 that can parameterise such a mix of noise sources, which would allow a quantitative fit to the data to be assessed. The lack of a more precise fit of the SDT models cannot be taken as evidence in favour of post-labelling models such as proposed by Rouger *et al.* We refer the reader to Micheyl and Oxenham (2012) for a discussion of the theoretical merits of different models.

7.4 Limitations & future research

The current work provides a starting point for investigations of the benefits obtained through visual speech information when listening to degraded speech in noise, and there are several avenues through which the work can be extended upon. One such avenue is to consider the type of background noise which is used. We have used multi-talker babble here, but it is possible that maximum visual speech benefit will occur with only a few competing talkers (e.g. 2, 4), when informational masking causes difficulties for speech perception (Freyman *et al.*, 2004; Brungart *et al.*, 2009). These are situations when additional listening strategies such as 'dip-listening' are possible and TFS cues might be particularly important (Lorenzi *et al.*, 2006; Moore, 2014; see also Bernstein *et al.*, 2009). Thus, it is difficult to predict whether estimates from the current experiment will generalize to situations with small numbers of background talkers. However, it should be noted that Rosen *et al.* (2013) found very small effects of the number of masking talkers when the speech and noise were both noise vocoded. It should also be acknowledged that only a single talker recorded the speech materials in Experiments 1 and 2, and a different talker was used in Experiment 3. Extending this work to different talkers is important as the utility of visual speech cues may differ according to the individual characteristics of different talkers (see Yakel *et al.*, 2000).

One question arising is to what extent degrading the speech stimuli generally led to a greater reliance on the visual signal, rather than the removal of information in the stimulus TFS *per se*. Two audio manipulations were used in these experiments: variation in SNR and removal of cues from the stimulus TFS. All the variants of models presented here are relatively successful in accounting for both of these manipulations. They assume that the interaction with the visual stimulus is exactly the same whether TFS or SNR are manipulated. Thus the modelling suggests that, at least for these two manipulations, it is intelligibility that matters and not the nature of the degradation. This could be logically tested further with, for example, manipulations of the spectral resolution, or stimuli that preserve TFS cues at the expense of ENV cues.

Limitations of vocoding as a simulation of the performance of cochlear-implant users also need to be acknowledged. The acoustic simulation used here simulates only the consequences of removing TFS from the speech signal and filtering the speech into a discrete number of frequency bands. Many other factors, such as the spread of electrical current along and across the cochlea (Cohen *et al.*, 2003), are not simulated, and the primary sources of stochasticity (normal hearing: inner haircell/auditory nerve synapse, Sumner *et al*. 2003; cochlear implant: spiral ganglion cell excitability, Horne *et al*. 2016) are very different. Thus, the encoding of speech on the auditory nerve is expected to be very different between electrical and tone-vocoded inputs. One potential difference in the nature of encoding has been highlighted recently by Shamma and Lorenzi (2013), who applied a model of early auditory processing explain the auditory nerve responses to Amplitude Modulated (AM) and Frequency Modulated (FM) vocoded speech. The AM conditions were the same as the ENV condition described here; the FM component was replaced by a tone with frequency equal to the central frequency of the analysis band. Shamma and Lorenzi's (2013) modelling suggested that regardless of vocoder manipulations, both ENV and TFS cues are expressed in the auditory nerve for vocoded speech, and both of these cues contribute to speech intelligibly. Thus, they argue that processing the speech to filter out TFS or ENV cues is not reflected in auditory nerve responses to these speech stimuli. They argue further that this is contrary to the auditory nerve responses for users of cochlear implants. It is therefore important to make the distinction between ENV and TFS cues present in the stimulus, which are similar for tone vocoding and cochlear implants, and the nature of the encoding on the auditory nerve which for the numerous reasons outlined is likely to be very different.

Another concern is that vocoder simulations in normal-hearing listeners cannot account for any adaptation to electrical stimulation over extended periods of time. Therefore, one must exercise caution in generalising the current findings related to the effects of informative TFS in normally-hearing listeners to users of cochlear implants. Future work with users of cochlear implants will establish whether the same pattern of results is observed. In addition, testing users of cochlear implants with the ENV conditions will allow us to test whether this manipulation introduces distortions that are additional to those attributable to their implants.

7.5 Conclusion

Visual information appears to be integrated in a similar way whether or not TFS cues are present in speech. However in practice this results in slightly better SNR advantages in the absence of TFS cues. Regardless, it suggests that visual information is at least as valuable when the auditory signal is degraded and this corresponds to a very valuable gain (4-7dB advantage in SNR). The results from the current studies suggest that the role of visual speech information needs to be given greater emphasis when evaluating people's ability to understand speech in noise, especially when faced with degraded speech input.

8. Acknowledgements

Work supported by the Nottingham Trent University, the NIHR Nottingham Hearing Biomedical Research Unit, and the MRC Institute of Hearing Research.

9. References

Altieri, N. A., Pisoni, D. B., & Townsend, J. T. (2011). Some normative data on lip-reading skills. *Journal of the Acoustical Society of America, 130*(1), 1-4. doi:10.1121/1.3593376

Bernstein, J. G. W., & Grant, K. W. (2009). Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, *125*(5), 3358-3372. doi:10.1121/1.3110132

Blamey, P. J., Cowan, R. S. C., Alcantara, J. I., Whitford, L. A., & Clark, G. M. (1989). Speech perception using combinations of auditory visual and tactile information. *Journal of Rehabilitation Research and Development, 26*(1), 15-24.

Boothroyd A., Hnath-Chisolm T., Hanin L., & Kishon-Rabin L. (1988). Voice fundamental frequency as an auditory supplement to the speechreading of sentences. *Ear and Hearing, 9,* 306–312.

Braida, L. (1991). Crossmodal integration in the identification of consonant segments. *Quarterly Journal of Experimental Psychology Section A-Human Experimental Psychology*, *43*(3), 647-677.

Brungart, D. S., Chang, P. S., Simpson, B. D., & Wang, D. (2009). Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and number of talkers. *Journal of the Acoustical Society of America*, *125*(6), 4006-4022. doi:10.1121/1.3117686

Cohen, L. T., Richardson, L. M., Saunders, E. & Cowan, R.S.C. (2003). Spatial spread of neural excitation in cochlear implant recipients: comparison of improved ECAP method and psychophysical forward masking. *Hearing Research*, *179,* 72-87. doi:10.1016/S0378-5955(03)00096-0

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition (L). *Journal of the Acoustical Society of America*, *120*(5), 2421-2424. doi:10.1121/1.2229005

Dau, T., Verhey, J., & Kohlrausch, A. (1999). Intrinsic envelope fluctuations and modulation-detection thresholds for narrow-band noise carriers. *Journal of the Acoustical Society of America, 106*(5), 2752-2760. doi:10.1121/1.428103

Davis, A. C. (1989). The prevalence of hearing impairment and reported hearing disability among adults in great-britain. *International Journal of Epidemiology, 18*(4), 911-917. doi:10.1093/ije/18.4.911

Desai, S., Stickney, G., & Zeng, F. (2008). Auditory-visual speech perception in normal-hearing and cochlear-implant listeners. *Journal of the Acoustical Society of America, 123*(1), 428-440. doi:10.1121/1.2816573

Dorman, M., Loizou, P., & Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *Journal of the Acoustical Society of America, 102*(4), 2403-2411. doi:10.1121/1.419603

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

Erber, N. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders, 40*(4), 481-492.

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191. doi:10.3758/BF03193146

Freyman, R., Balakrishnan, U., & Helfer, K. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *Journal of the Acoustical Society of America, 115*(5), 2246-2256. doi:10.1121/1.689343

Fu, Q., Shannon, R., & Wang, X. (1998). Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing. *Journal of the Acoustical Society of America, 104*(6), 3586-3596. doi:10.1121/1.423941

Glasberg, B., & Moore, B. (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., Ardell, L., Kuhl, P., & Sparks, D. (1985). The contribution of fundamental-frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects. *Journal of the Acoustical Society of America, 77*(2), 671-677. doi:10.1121/1.392335

Grant, K., Braida, L., & Renn, R. (1991). Single band amplitude envelope cues as an aid to speechreading. *Quarterly Journal of Experimental Psychology Section A-Human Experimental Psychology, 43*(3), 621-645. doi:10.1080/14640749108400990

Grant, K., Braida, L., & Renn, R. (1994). Auditory supplements to speechreading - combining amplitude envelope cues from different spectral regions of speech. *Journal of the Acoustical Society of America, 95*(2), 1065-1073. doi:10.1121/1.408468

Grant, K. W., Tufts, J. B., & Greenberg, S. (2007). Integration efficiency for speech perception within and across sensory modalities by normal-hearing and hearing-impaired individuals. *Journal of the Acoustical Society of America, 121*(2), 1164-1176. doi:10.1121/1.2405859

Grant, K., Walden, B., & Seitz, P. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America, 103*(5), 2677-2690. doi:10.1121/1.422788

Green, D.M., Swets J.A. (1966) *Signal Detection Theory and Psychophysics*. New York: Wiley.

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

IEEE. (1969). *IEEE Recommended Practice for Speech Quality Measurements*. New York: Institute for Electrical and Electronic Engineers.

Ihlefeld, A., Deeks, J. M., Axon, P. R., & Carlyon, R. P. (2010). Simulations of cochlear-implant speech perception in modulated and unmodulated noise. *Journal of the Acoustical Society of America, 128*(2), 870-880. doi:10.1121/1.3458817

Kaiser, A., Kirk, K., Lachs, L., & Pisoni, D. (2003). Talker and lexical effects on audio-visual word recognition by adults with cochlear implants. *Journal of Speech Language and Hearing Research, 46*(2), 390-404. doi:10.1044/1092-4388(2003/032)

Kong, Y., & Carlyon, R. P. (2007). Improved speech recognition in noise in simulated binaurally combined acoustic and electric stimulation. *Journal of the Acoustical Society of America, 121*(6), 3717-3727. doi:10.1121/1.2717408

Kramer, S., Kapteyn, T., & Festen, J. (1998). The self-reported handicapping effect of hearing disabilities. *Audiology, 37*(5), 302-312.

Lakatos, P., Chen, C., O'Connell, M. N., Mills, A., & Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron, 53*(2), 279-292. doi:10.1016/j.neuron.2006.12.011

Langeheine R., Pannekoek J., & Van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods and Research, 24(4),* 492-516. Doi: 10.1177/0049124196024004004

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., & Moore, B. C. J. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences of the United States of America, 103*(49), 18866-18869. doi:10.1073/pnas.0607364103

Luce, P., & Pisoni, D. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*(1), 1-36. doi:10.1097/00003446-199802000-00001

Lunner, T., Hietkamp, R. K., Andersen, M. R., Hopkins, K., & Moore, B. C. J. (2012). Effect of speech material on the benefit of temporal fine structure information in speech for young normal-hearing and older hearing-impaired participants. *Ear and Hearing, 33*(3), 377-388. doi:10.1097/AUD.0b013e3182387a8c

Massaro D. W. (1987). Speech perception by ear and eye. In *Hearing by Eye: The Psychology of Lip-Reading*, eds Dodd B., Campbell R., editors. (Hillsdale, NJ: Lawrence Erlbaum), 53–83.

McGettigan, C., Faulkner, A., Altarelli, I., Obleser, J., Baverstock, H., & Scott, S. K. (2012). Speech comprehension aided by multiple modalities: Behavioural and neural interactions. *Neuropsychologia, 50*(5), 762-776. doi:10.1016/j.neuropsychologia.2012.01.010

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

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, 27, 338-352.

Moore, B. (2014). *Auditory processing of temporal fine structure: effects of age and hearing loss*. Singapore: World Scientific.

Moore, B. C. J. (2008). The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *Jaro-Journal of the Association for Research in Otolaryngology, 9*(4), 399-406. doi:10.1007/s10162-008-0143-x

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

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

Rader, T., Adel, Y., Fastl, H., & Baumann, U. (2015). Speech perception with combined electric-acoustic stimulation: A simulation and model comparison. *Ear and Hearing, Epub ahead of print*

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. *Journal of the Acoustical Society of America, 133*(4), 2431-2443. doi:10.1121/1.4794379

Rouger, J., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O., & Barone, P. (2007). Evidence that cochlear-implanted deaf patients are better multisensory integrators. *Proceedings of the National Academy of Sciences of the United States of America, 104*(17), 7295-7300. doi:10.1073/pnas.0609419104

Schafer, E. C., & Thibodeau, L. M. (2004). Speech recognition abilities of adults using cochlear implants with FM systems. *Journal of the American Academy of Audiology, 15*(10), 678-691. doi:10.3766/jaaa.15.10.3

Seldran, F., Micheyl, C., Truy, E., Berger-Vachon, C., Thai-Van, H., & Gallego, S. (2011). A model-based analysis of the "combined-stimulation advantage". *Hearing Research, 282*(1-2), 252-264. doi:10.1016/j.heares.2011.06.004

Shamma, S., & Lorenzi, C. (2013). On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system. *Journal of the Acoustical Society of America*, 133(5), 2818-2833. doi:10.1121/1.4795783

Shannon, R., Zeng, F., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*(5234), 303-304. doi:10.1126/science.270.5234.303

Skinner, M., Clark, G., Whitford, L., Seligman, P., Staller, S., Shipp, D., *Et al.* (1994). Evaluation of a new spectral peak coding strategy for the nucleus 22 channel cochlear implant system. *American Journal of Otology, 15*, 15-27.

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

Sumner, C.J., Lopez-Poveda, E.A., O'Mard, L.P., & Meddis R. (2002). A revised model of the inner-hair cell and auditory-nerve complex. *Journal of the Acoustical Society of America, 111*(5), 2178-88. doi: 10.1121/1.1453451

Treisman, M. (1998). Combining information: probability summation and probability averaging in detection and discrimination. *Psychological Methods 3*, 252–265. doi: 10.1037/1082-989X.3.2.252

Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J., & Hale, S. (2010). Aging, audio-visual integration, and the principle of inverse effectiveness. *Ear and Hearing, 31*(5), 636-644. doi:10.1097/AUD.0b013e3181ddf7ff

Tyler, R., Parkinson, A., Woodworth, G., Lowder, M., & Gantz, B. (1997). Performance over time of adult patients using the ineraid or nucleus cochlear implant. *Journal of the Acoustical Society of America, 102*(1), 508-522. doi:10.1121/1.419724

Whitmal,N A. III, Poissant, S. F., Freyman, R. L., & Helfer, K. S. (2007). Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience. *Journal of the Acoustical Society of America, 122*(4), 2376-2388. doi:10.1121/1.2773993

Wolfe, J., Schafer, E. C., Heldner, B., Muelder, H., Ward, E., & Vincent, B. (2009). Evaluation of speech recognition in noise with cochlear implants and dynamic FM. *Journal of the American Academy of Audiology, 20*(7), 409-421. doi:10.3766/jaaa.20.7.3

Yakel, D., Rosenblum, L., & Fortier, M. (2000). Effects of talker variability on speechreading. *Perception & Psychophysics, 62*(7), 1405-1412. doi:10.3758/BF03212142

Table 1: Average SRT_{50}s for each of the experiments, including the overall differences in SRT_{50}s according to modality and processing; for *modality* the Audio-visual and Audio-only SRT_{50}s have been averaged across both types of processing (TFS and ENV) and for *processing* the TFS and ENV SRT_{50}s have been averaged over both modalities (Audio-visual and Audio-only). All values show dBs, and standard deviations are shown in brackets.

	Experiment 1	**Experiment 2**	**Experiment 3**
Audio-visual	-6.3 (3.2)	-6.8 (3.0)	-9.1 (2.9)
Audio-only	-0.6 (4.8)	-2.2 (3.6)	-6.6 (3.9)
Modality Difference	*5.7*	*4.6*	*2.5*
TFS	-6.7 (2.6)	-7.4 (2.4)	-10.0 (1.6)
ENV	-0.1 (4.4)	-1.6 (3.2)	-5.7 (3.2)
Processing Difference	*6.6*	*5.8*	*4.3*

Table 2: For each experiment the results of fitting the different models. The goodness of fit is expressed as the X^2 statistic between the AV data conditions and model, *p* represents that probability that these are indistinguishable, and the mean signed error (in % correct) between the data and model indicates where the real performance is greater than or less than the models. The bottom row gives the SRT advantage of adding visual information for the ENV condition over the TFS condition.

	Experiment 1			Experiment 2			Experiment 3	
	Ind. noise	Late noise	Rouger model	Ind. noise	Late noise	Rouger model	Ind. noise	Late noise
α	0	1	-	0	1	-	0.29	1
g.o.f (X^2)	1042	1072	339	385	1626	137	220	141
p	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
M.S.E. (%)	11.86	-5.34	2.66	7.44	-9.48	-1.87	0.82	7.7
AV SRT advantage ENV–TFS	-0.7dB	-3.1dB	-1.9dB	-0.6dB	-1.4dB	-1.1dB		

Figure captions

Figure 1: Speech perception performance (in % correct) as function of Signal-to-Noise ratio. The plots on the left show data for ENV speech, while the plots on the right show performance for TFS speech. The filled triangles show data from the Audio-visual conditions, and the open triangles show Audio-only performance. Error bars indicate sample 95% confidence intervals. Sigmoidal curves have been fit to the averaged data. The red dashed line shows 50% correct performance.

Figure 2: Speech Reception Thresholds: The Signal-to-Noise ratio at which performance was 50% correct. Calculated from 3-parameter sigmoidal functions fit for each participant. Error bars indicate sample 95% confidence intervals. The dashed and dotted lines show the three models' (SDT Independent Noise, SDT Late Noise, and Rouger *et al*.'s model) predictions of the audio-visual (AV) data.

Figure 3: Visual speech benefit. The benefit (in dB) gained from the addition of visual speech information. For Experiment 1, this is calculated from the overall difference in SRT50s between the Audio-visual and Audio-only conditions for Vocoded and Clear Speech, and therefore represent the between-groups effect. For Experiments 2 and 3, the benefit was derived by averaging the difference between Audio-visual and Audio-only SRTs for each participant, and therefore represent the within-groups effect. Error bars indicate 95% confidence intervals; the confidence for Experiment 1 are expected to be wider than the confidence intervals for Experiments 2 and 3 as they include both within and between-subject variance.

Figure 4: Proportion of letter words correct. The top panel shows auditory-only accuracy for TFS and ENV conditions, and the bottom panel shows Visual Benefit. Error bars indicate 95% confidence intervals.

Figure 5: Meta-analysis of size of the additional visual benefit observed when information TFS was not available compared to when it was available across Experiments 1, 2, and 3. Filled circles plot the effect size (in dB) in each individual experiment and error bars plot the 95% confidence intervals for the effects. The filled diamond represents the pooled effect size across the three experiments from a random-effects meta-analysis.

Figure 6: The results of fitting the independent late noise models, along with Rouger *et al.*'s model to the three experiments. The points show the observed data, and the dotted and dashed lines show the predictions from models. Shaded regions show the standard errors for the data.

Figure 1

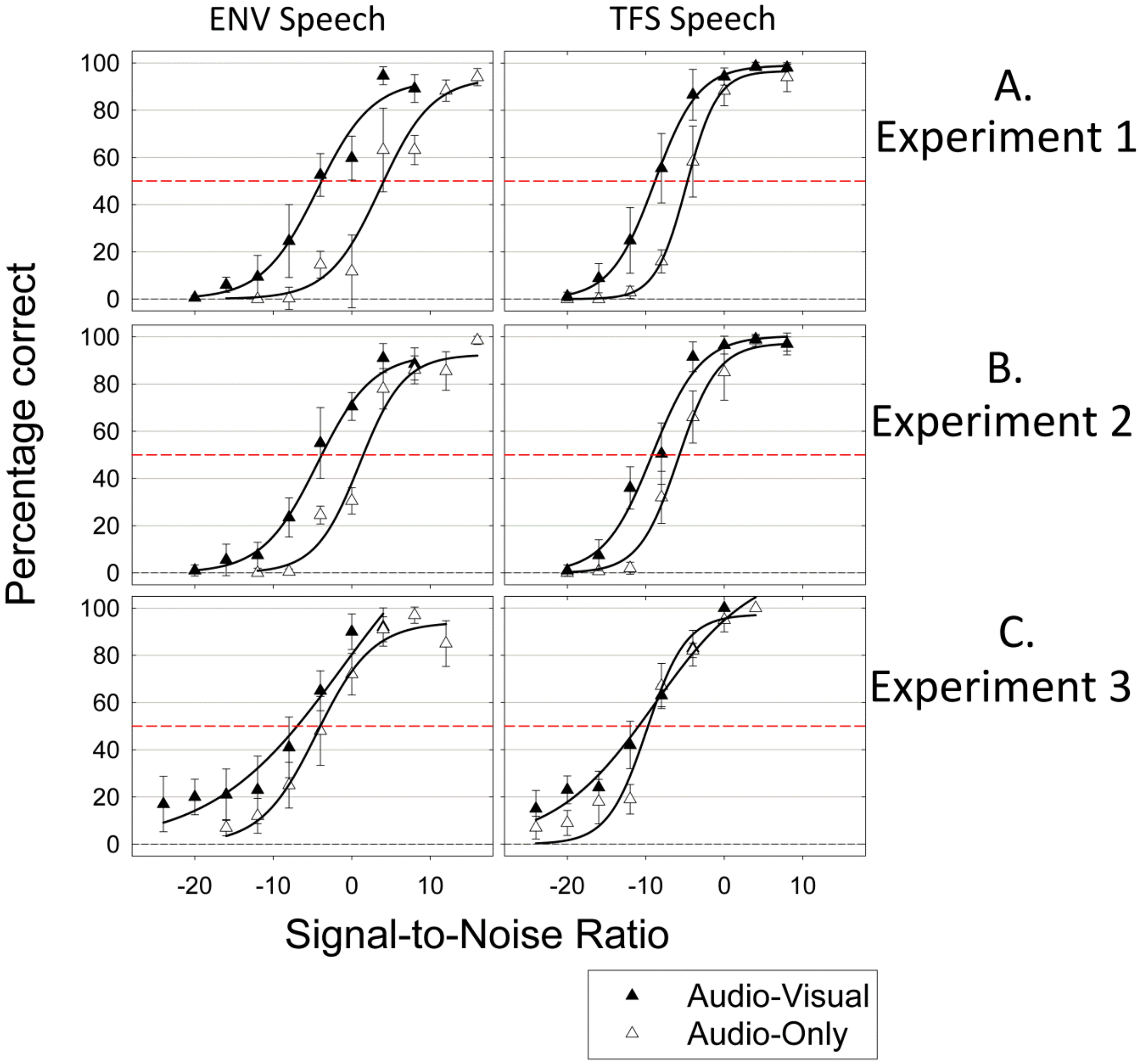
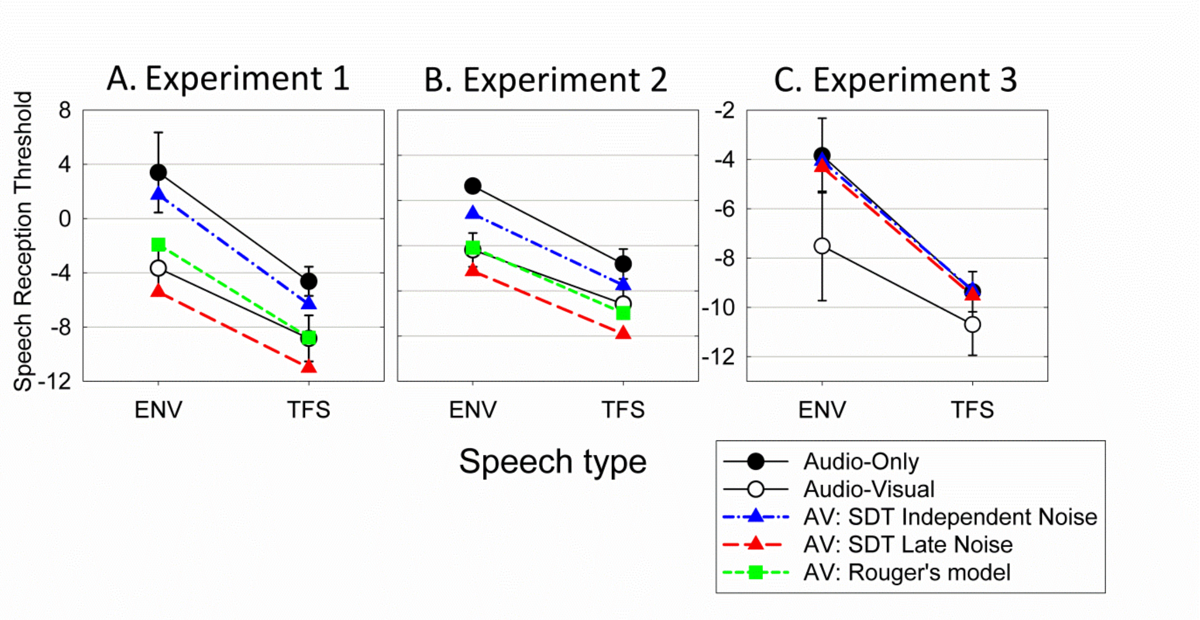

ENV Speech
TFS Speech
A.
Experiment 1
B.
Experiment 2
C.
Experiment 3
Percentage correct
100
80
60
40
20
0
-20
-10
0
10
Signal-to-Noise Ratio
Audio-Visual
Audio-Only

Figure 2



A. Experiment 1
B. Experiment 2
C. Experiment 3
Speech Reception Threshold
8
4
0
-4
-8
-12
-2
-4
-6
-8
-10
-12
ENV
TFS
ENV
TFS
ENV
TFS
Speech type
Audio-Only
Audio-Visual
AV: SDT Independent Noise
AV: SDT Late Noise
AV: Rouger's model

Figure 3

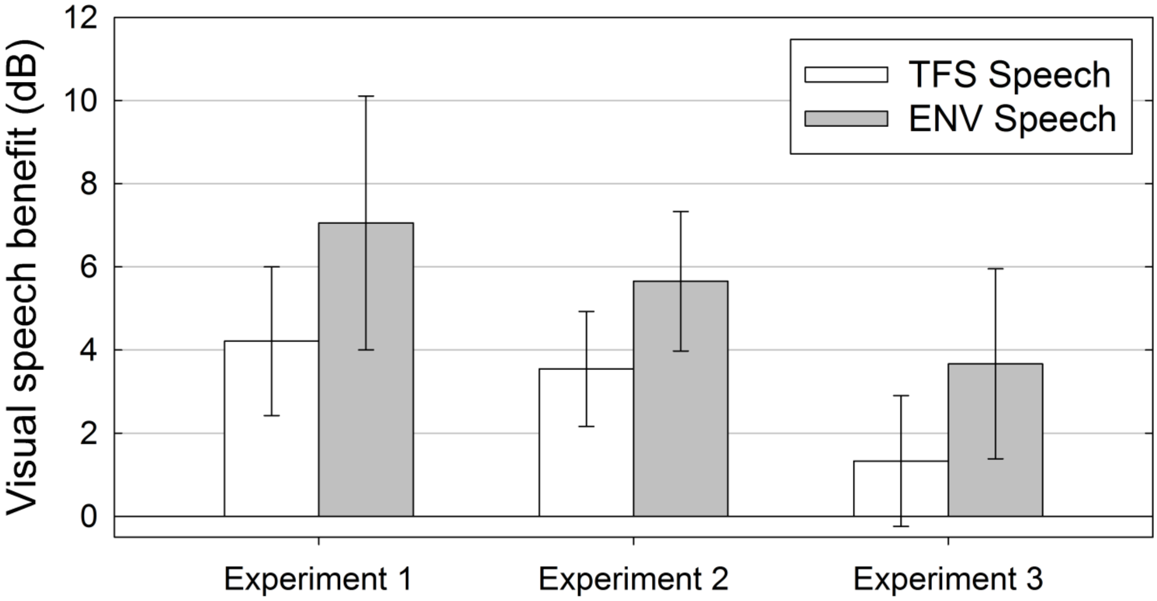
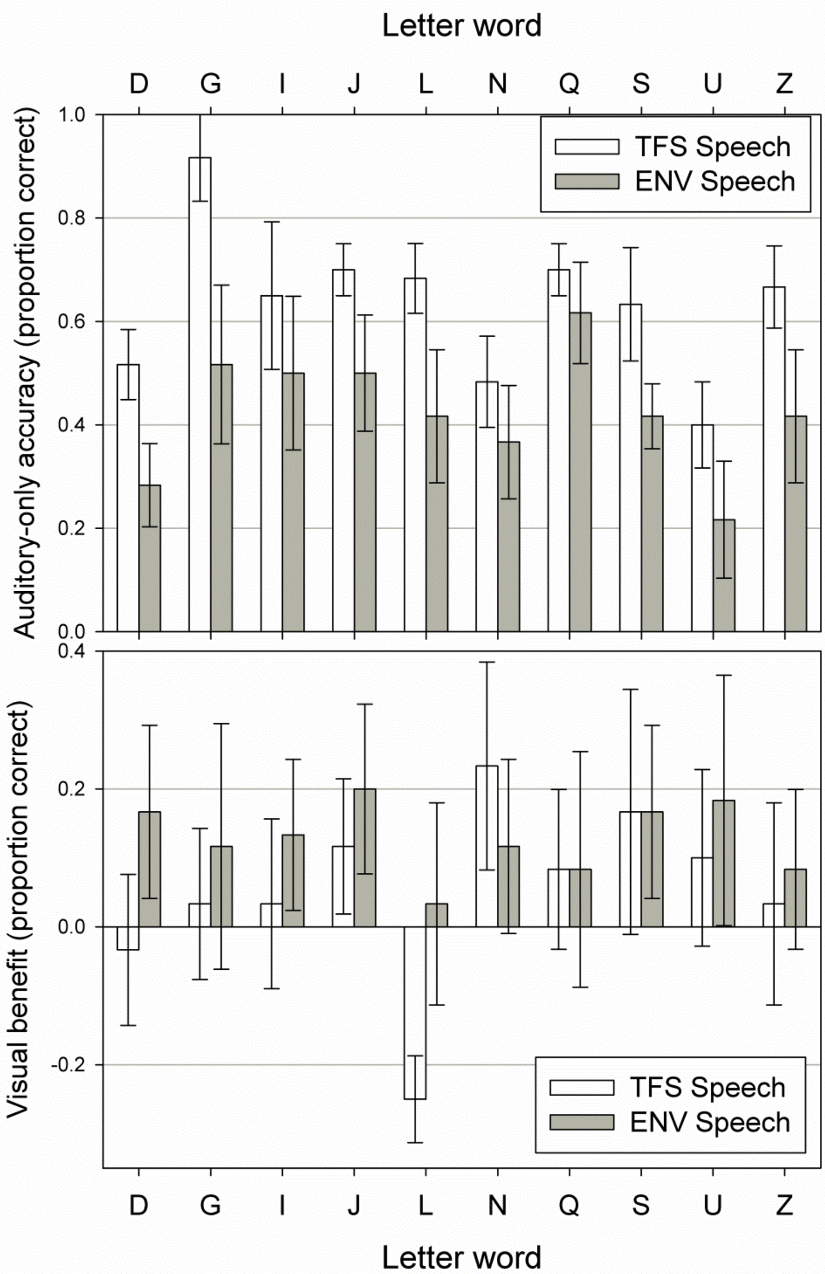
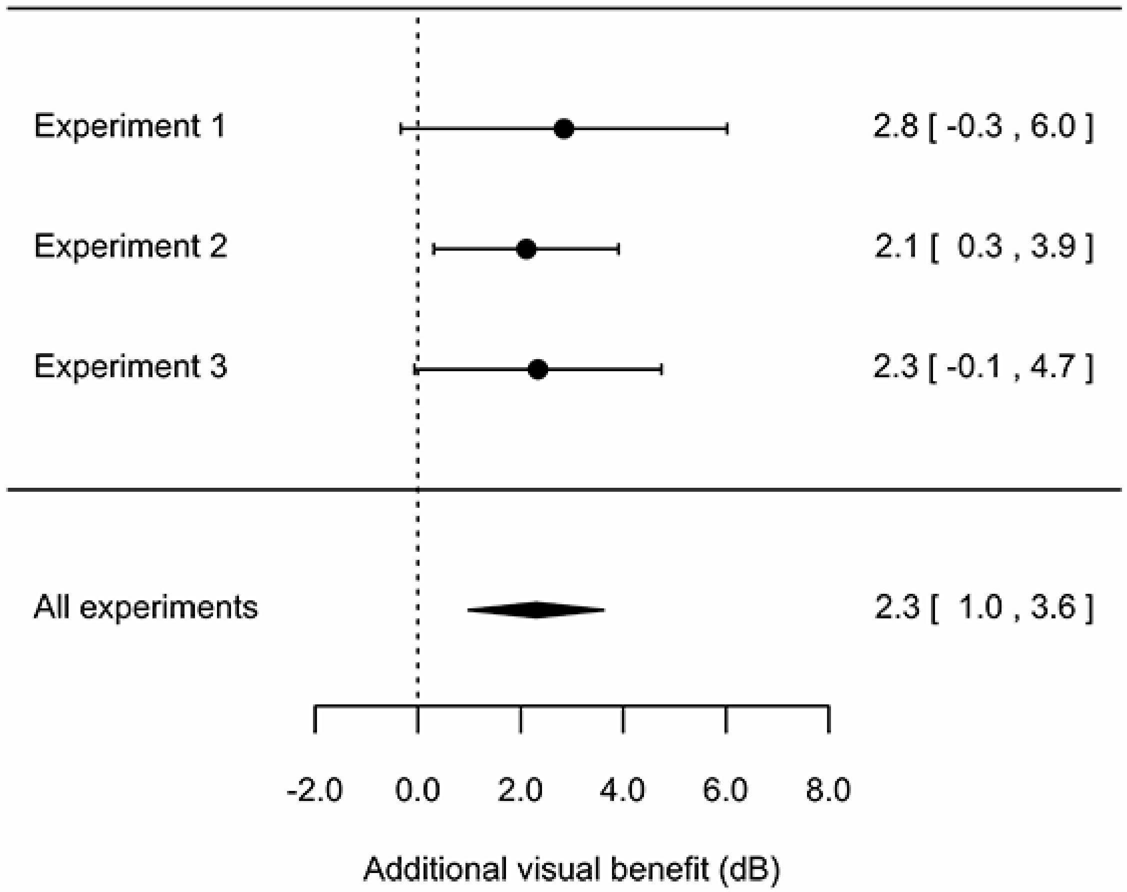

Visual speech benefit (dB)
12
10
8
6
4
2
0
TFS Speech
ENV Speech
Experiment 1
Experiment 2
Experiment 3

Figure 4



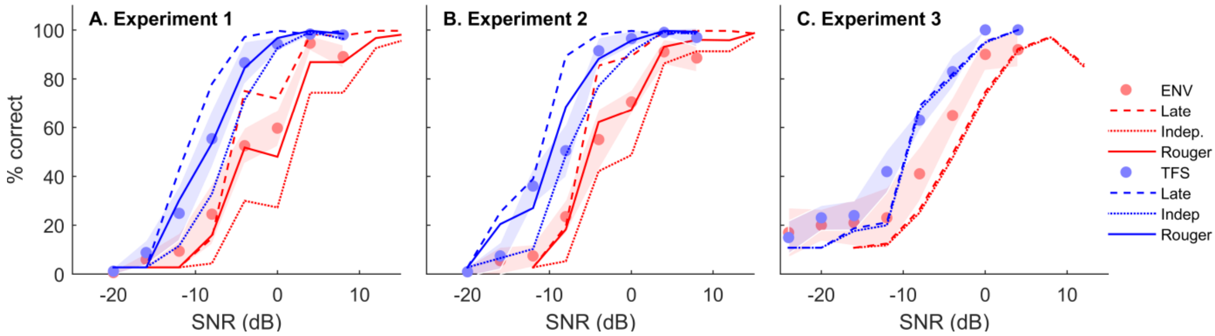
Letter word
D
G
I
J
L
N
Q
S
U
Z
Auditory-only accuracy (proportion correct)
1.0
0.8
0.6
0.4
0.2
0.0
TFS Speech
ENV Speech
Visual benefit (proportion correct)
0.4
0.2
0.0
-0.2
TFS Speech
ENV Speech
D
G
I
J
L
N
Q
S
U
Z
Letter word

Figure 5



Experiment 1
2.8 [-0.3 , 6.0]
Experiment 2
2.1 [0.3 , 3.9]
Experiment 3
2.3 [-0.1 , 4.7]
All experiments
2.3 [1.0 , 3.6]
-2.0
0.0
2.0
4.0
6.0
8.0
Additional visual benefit (dB)

Figure 6


A. Experiment 1
B. Experiment 2
C. Experiment 3
% correct
SNR (dB)
ENV
Late
Indep.
Rouger
TFS
Late
Indep
Rouger