SOURCES AND CORRELATES OF PERFORMANCE ENHANCEMENT IN AUDIOVISUAL SPEECH PERCEPTION

by

Celina Nahanni

A thesis submitted to the Centre for Neuroscience Studies
In conformity with the requirements for
The degree of Doctor of Philosophy

Queen’s University
Kingston, Ontario, Canada
(July, 2014)

Copyright ©Celina Nahanni, 2014
Abstract

In a noisy environment, speech intelligibility is greatly enhanced by seeing the speaker’s face. This enhancement results from the integration of auditory and visual signals, but the underlying mechanisms remain largely unknown. This thesis describes the results from four studies investigating the enhancement of speech comprehension in an open-set, word-in-noise identification task. In the first study, we examined how the auditory signal-to-noise ratio impacts audiovisual performance in a large participant sample. Consistent with the majority of previous studies, we found high inter-individual variability, whose magnitude increased monotonically with increasing signal-to-noise ratio. We also found that audiovisual performance was highly variably across studies, even after normalizing with auditory-only performance, suggesting that experimental factors strongly affect performance. In the second study, we replicated the findings from the first study and developed a measure of audiovisual ‘integration enhancement’ that captures how much of the audiovisual performance cannot be accounted for by performances in visual-only and auditory-only tasks. Contrary to the principle of inverse effectiveness, this integration enhancement was found not to decrease with auditory signal-to-noise ratio but to peak at an intermediate ratio. In the third study, we developed a model based on the congruency of response errors (confusions) observed in auditory-only and visual-only task conditions to predict audiovisual performance enhancement. This relatively simple categorical model was found to account for audiovisual performance enhancement in 60% of our stimuli, lower level processing was likely responsible for the enhancement in the remainder of the words. In the fourth study, we investigated how much of the audiovisual performance enhancement can be predicted by the integration of auditory and visual signals, a process that we estimated from the susceptibility of one’s auditory perception to be altered by incongruent (McGurk) visual stimuli.
Surprisingly, we found no correlation between the McGurk illusion susceptibility for nonsense syllables and our measure of audiovisual integration enhancement for words. Altogether, these findings suggest that a large amount of performance enhancement in speech comprehension is attributable to categorical constraints rather than to a discrete integration process. This work also highlights the importance of developing standardized approaches for future investigations of audiovisual speech perception.
Acknowledgements

This project could not have come together without the help and support from a great number of people. First and foremost, I would like to thank my supervisors, Dr. Martin Paré and Dr. Kevin Munhall. Though this process has not always been easy, I have always known that your guidance was helping me to produce my best work. I am grateful for all the lessons you have taught me and I am a better scientist and person for them.

I would also like to thank members of the Paré and Munhall labs, past and present, for making my working environment so welcoming, getting to know you all has been such a pleasure.

To my wonderful family: Dad, Grandpa, Tamara, Taj, Adrienne, Sierra, Tristan, Russell, Karen, Kiana, Taylor. And to my soon-to-be family: Evelyn, John, Steven and all the Tators, Shumaks and Dantowitzs. Thank you all for celebrating with me in good times and supporting me in challenging times. I am so fortunate to be related to such wonderful, supportive and thoughtful people. I would also like to thank Dave Milstein for getting his PhD first and sharing of his experiences so I could avoid learning exclusively through trial-and-error. And to my good friend Seth Mendelson, you’re alright. Last but not least, I would like to thank Michael Warfe, your steadfast support and patience made all the difference and I cannot express how grateful I am to have you in my life. Thank you.

This work was supported by operating grants from the Canadian Institutes of Health Research (CIHR) and Queen’s University. In addition, to I held a postgraduate scholarship from the Natural Sciences and Engineering Council of Canada (NSERC).
Co-Authorship

Dr. Martin Paré and Dr. Kevin Munhall were the principal investigators and supervisors for the studies described in this thesis. Dr. Paré and Dr. Munhall conceived and designed the research protocols for Chapter 1 and I assisted them in designing the protocols described in Chapters 2, 3 and 4. Mr. Justin Deonarine collected a portion of the data and assisted in developing some of the DMDX display software used in Chapter 1. I was responsible for all the remaining data collection and analysis. I produced the first draft of this thesis and subsequent drafts included editorial submissions from Dr. Paré and Dr. Munhall.
# Table of Contents

Abstract ............................................................................................................................................... ii  
Acknowledgements ......................................................................................................................... iv  
Co-Authorship ................................................................................................................................. v  
List of Figures ..................................................................................................................................... vii  
List of Abbreviations ....................................................................................................................... xii  
Chapter 1: Introduction ..................................................................................................................... 1  
Chapter 2: Performance Variability in Audiovisual Speech-in-Noise Perception ......................... 29  
Chapter 3: Measuring Performance Gain in an Audiovisual Speech-in-Noise task ..................... 57  
Chapter 4: Lexical Congruency between Unimodal Confusions Accounts for Some (But not All) of the Performance Enhancement in Audiovisual Speech Perception ............................................. 87  
Chapter 5: Susceptibility to the McGurk Illusion is Negatively Correlated with Performance in an Audiovisual Speech-in-Noise Task ................................................................................................................. 114  
Chapter 6: Discussion ..................................................................................................................... 133  
References ........................................................................................................................................ 149  
Appendix A: Word List with Phonemic Transcription .................................................................. 173
List of Figures

Figure 1: All collected data. Error bars represent the 95% confidence interval for the mean as measured via bootstrapping. **Top panel:** Average performance curves are in bold and individual subjects are illustrated in fine lines. Both functions are consistent with smooth sigmoid functions with performance asymptotes at 0 and 1. The AV function hits the ceiling at +10 dB. **Bottom panel:** the performance difference (AV-AO) is illustrated in fuchsia. The shape of the difference exhibits a peak at the -6 dB due to floor effects on the AO function. The percent gain (AV-AO)/AV is illustrated in purple; the shape of the curve is monotonically decreasing. ............................................................... 41

Figure 2: The P(AV) vs. P(AO) averaged across subjects. Each point represents an SNR level. The correlation of the averaged data is $R^2 = 0.902$ (p = 0.001), representative of the high within-subject correlations between P(AO) and P(AV). ............................................................... 44

Figure 3: The width of the 95% confidence interval (AO in blue, AV in red) is related to SNR in the AV condition, but not in the AO condition. Dashed lines between points indicate measures that contained more than 15% saturation (Performance at 0 or 1). Performance saturation will lead to depression of the variability and therefore must be interpreted cautiously. The low CI for AO at -10 dB is due to floor effects. The levelling of the CI for high AV values is due to ceiling effects. ............................................................... 45

Figure 4: Estimates of the first derivatives of AV performance (calculated with Equation 1) for average AV data in the current study (left panel) and from Ross, Saint-Amour, Leavitt, Javitt, et al. (2007)(right panel). The presence of a local maximum (indicated with arrows) indicates points of inflection where the rate of change of performance peaks. The largest maximum is associated with the reported 'sweet spot' of integration. The smaller maximum is likely due to ceiling effects that depressed the performance at +14 dB. No such local maxima were observed in the current study....................................... 47

Figure 5: Illustration of the variability in reported performances on speech-in-noise tasks from 8 studies of single words presented in noise – thick lines represent the results from the current study. AV performance is in red; AO performance is in blue. The results from the current study are bolded. **Top Panel:** The reported variability due to different SNR scales is great. **Bottom Panel:** Data from the same studies on a normalized SNR scale
such that the AO performance at two points (0.26 and 0.70) is locked. Variability in the AV performances is reduced, but still remains large. The variability across studies is due to factors outside of AO performance and differing SNR measurements. Figure 6: Combined analysis for 8 studies measuring performance in a speech-in-noise task for single words. The results from the current study are in bold (red). Due to a high degree of variability in SNR measures, is it more meaningful to present AV data as a function of AO performance. However, variability is not fully explained by SNR effects. Inter-study variability due to methodological differences and presentation effects remain substantial. Figure 7: Performance on 3 of the 14 counterbalancing conditions. In this subset the AV performance appears to be enhanced at −6 dB, however this effect is not present when all the counterbalancing conditions are averaged together. Figure 8: If the two sets (AO and VO) have an area of overlap (R) then the total area is calculated as (AO + VO - R). This is because the simple addition (AO + VO) will lead to the overlap region (R) being counted twice as it occurs in both the AO and VO conditions. Figure 9: Mean Performance in the VO, AO and AV conditions. Auditory quality was varied across 6 SNRs. Error bars indicate the 95% confidence interval calculated using bootstrapping (70 samples with repeats, and 10,000 trials). Figure 10: Estimated Redundancy between the AO and VO channels at each SNR. This value will never exceed VO performance (0.17). Error bars represent the 95% confidence intervals measured with bootstrapping (N=70 with repeats; 10,000 trials). Figure 11: Different Gain metrics applied to the data from this study. When a metric is normalized to the total possible performance increase (visual enhancement (VE) and normalized integration enhancement (IEnh)) it will not follow the PoIE. Metrics that are not normalized (difference score (DS) and integration enhancement (IE)) as well as %Gain (which is normalized to the total performance) are consistent with the PoIE. **Left Panel:** Gain Metrics that estimate contributions from VO plus Integration. These are calculated using AV and AO values. **Right Panel:** Gain Metrics measuring the deviation of AV from a linear estimate. These metrics estimate the contribution due to
the integration mechanism that is outside of the unimodal streams; this includes integration gain modulation, intermodal information and lexical constraints.  

Figure 12: Proportion of the performance due to each element in bimodal speech perception. The sum of these elements will always equal 1. The contributions are from auditory information (blue), visual information (green) and bimodal-specific performance gain (pink). The %Gain Integration curve (pink) is the same as the one illustrated in the right panel of Figure 11.  

Figure 13: The magnitude of the three components associated with AV performance. These curves result from taking each point in Figure 12 and multiplying the point by the AV performance at a given SNR. As a result, the sum of the three curves at any SNR will equal the AV. The Contribution of VO decreases as the R between AO and VO increases. The contribution from the bimodal-specific integration process (pink curve) is lowest at higher SNRs, increasing as the SNR is decreased. At -6 dB the contribution begins to fall off as information necessary for whole word identification in the AO channel is critically degraded by noise.  

Figure 14: Overlap in the unimodal error confusions can lead to a prediction that is used to constrain the bimodal solution. In Case 1 the AO and VO error confusions overlap over the correct solution (Prediction = Target) and performance is boosted. In Case 2 there is no overlap between the unimodal confusions and therefore there is no constraint and the system will perform at the rate of the sum of AO and VO. In Case 3 the AO and VO confusions overlap over an incorrect solution (Prediction = Distractor) and the system is biased away from the target word, depressing performance.  

Figure 15: The initial AO and VO inputs activate a set of possible solutions during the co-activation stage. Competition between multiple overlapping solutions leads to a correct prediction for the word “bush”. High co-activation in both the AO and VO conditions means that “bush” out-competes “push” and “porch”.  

Figure 16: Competition between multiple overlapping solutions leads to an incorrect prediction of the word “gross” for the presented word “roast”. High probability of selection in the AO condition and co-activation with the probability of selection in the VO condition leads to maximal activation of a distractor.
Figure 17: Overview of the Confusion Constraints (CC) Model. Co-activation of the AO and VO channels is used to constrain the solution set from which the final bimodal response is chosen. When the model predicts the target word performance is maximally enhanced. When the model predicts a distractor the performance is depressed. When the model is not able to make a prediction the performance is estimated to be the sum of the unimodal performances.

Figure 18: **Top Panel:** Group results for the observed AV performance (red line) and the linear estimate $AV_{\text{LIN}}$ (yellow line). Error bars are the 95% confidence interval for the mean as measured by bootstrapping (N=350 with repeats; 10,000 trials). **Bottom Panel:** Integration Enhancement (IE) is the difference between the observed performance ($AV_{\text{obs}}$) and the predicted performance. This figure depicts the linear model of AV performance ($AV_{\text{LIN}}$).

Figure 19: The IE for the Confusion Constraints model ($IE_{\text{CC}}$) is lower than for that of the linear model ($IE_{\text{LIN}}$). This means that the CC model accounts for some, but not all, of the super-additivity.

Figure 20: The $IE_{\text{CC}}$ broken down into the three CC cases (see Figure 14). Showing that super-additivity is accounted-for in Case 1 (60% of the wordlist), where the presented word and the predicted word coincide (Prediction = Target). The IE remains high for words with no Prediction (18% of the wordlist) and for words whose prediction was incorrect (Prediction = Distractor) (22% of the wordlist).

Figure 21: Behavioural results. Panel A: Average susceptibilities to the McGurk illusion for nonsense VCV syllables presentations, the flanking vowel was /a/. Stimuli are labeled as [auditory, visual] consonant combinations. Results were measured as the average responses that were not the presented audio or video. The four significantly different token pairs are labeled. Error bars indicate the variance. Panel B: Average performance on three speech perception tasks – AO at -5 dB, VO in silence and AV at -5 dB. Stimuli were single monosyllabic high frequency words. Performance in AO and VO were not significantly different from each other. Performance in AV was significantly different from AO and VO and was substantially greater than the sum of the unimodal performances. Error bars indicate the standard deviation.
Figure 22: Scatterplot illustrating the negative correlation ($\beta_1 = -0.189$) between susceptibility to the [K, P] McGurk stimulus and VO.

Figure 23: Performance on vowels. A different scoring method was applied to the data from Chapter 3. Performance was measured according to whether the response word had the same vowel as the presented word. The high performance values show how whole word identification performance underestimates the actual informational content of the signals.
List of Abbreviations

ANCOVA: Analysis of covariance
ANOVA: Analysis of variance
AO: Auditory only
AV: Audiovisual
$AV_{\text{predicted}} / AV_{\text{pred}}$: Predicted audiovisual performance
CC: Confusion constraints model
CMU: Carnegie Melon University
CI: Confidence interval
dB: Decibel
FLMP: Fuzzy logic model of perception
fMRI: Functional magnetic resonance imaging
IE: Integration Enhancement
MANCOVA: Multivariate analysis of variance
MSE: Mean-squared-error
MRC: Medical Research Council
PoIE: Principle of inverse effectiveness
PRE: Pre-labeling model
R: Redundancy
SNR: Signal-to-Noise ratio
STS: Superior temporal sulcus
VCV: Vowel-consonant-vowel
VO: Visual-only
WF: Word-frequency
Chapter 1: Introduction

The study of audiovisual (AV) speech perception is a category of multisensory perception research. Although the current research addresses specific questions related to the comprehension of speech, the questions are couched in the larger framework of multisensory perception aimed towards understanding more generally the concept of information binding. Information binding is the process that allows sensory signals to be joined into a unified percept resulting in a richer representation of the stimulus. The processing of audiovisual speech is a special case of integration because there exist specialized processing structures in the brain for speech as opposed to non-speech stimuli. Investigators have described specific cortical structures being devoted to speech processing (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Moore, 2000), and the perception of audiovisual speech is arguably processed by a specialized integration mechanism maximally tuned for such stimuli (Calvert, Spence, & Stein, 2004; Tuomainen, Andersen, Tiippana, & Sams, 2005; Vatakis, Ghazanfar, & Spence, 2008).

AV integration manifests as improved performance on speech comprehension tasks. For example, when a person is listening to a talker in a noisy environment their visual system collects information on the place-of-articulation speech feature, which is degraded in noise in the auditory channel, and the auditory signal carries more detailed information regarding the voicing and nasality. These signals are eventually merged into a coherent percept that carries lexical information. The merging of these signals results in improved performance in a speech comprehension task, and the enhanced performance is
super-additive, meaning it is greater than the sum of the performances in auditory-only (AO) and visual-only (VO) conditions.

There are many questions regarding how the brain is able to accomplish this process. First, the neural encoding of each sensory signal must carry information that the brain is able to interpret as a “binding cue” that marks the two signals for integration; the precise nature of this cue is unknown (Engel, König, Kreiter, Schillen, & Singer, 1992; Senkowski, Schneider, Foxe, & Engel, 2008; Singer & Gray, 1995). Second, once the two signals are flagged for integration, the mechanism through which they are combined is also unknown. To account for super-additivity, the integration mechanism may be either amplifying the strength of the speech signal by employing a gain modulatory process, or by depressing the noise in the signal enforcing constraints to the processing of the sensory information. Such constraints would be any factor that helps the brain narrow down the set of possible solutions from which it chooses its response. This thesis focuses on the integration process that occurs after the signals are flagged for binding. Specifically, this thesis addresses questions related to super-additivity in a word-identification task, and how it is related to the integration process more generally.
Characteristics of AV Speech Perception

Super-additivity

Previous work in audiovisual speech perception research has shown that the intelligibility of single words presented in noise is substantially better when presented as a bimodal signal as opposed to an auditory-only signal (Erber, 1969, 1971a, 1971b; Ewertsen & Nielsen, 1971; Ma, Zhou, Ross, Foxe, & Parra, 2009; O'Neill, 1954; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sommers, Tye-Murray, & Spehar, 2005; Sumby & Pollack, 1954). The difference between AV and AO performance has been observed to be super-additive, that is to say, greater than the sum of the performance in the AO and VO conditions (Dodd, 1977; Ewertsen & Nielsen, 1971; Ma et al., 2009; Risberg & Lubker, 1978; Ross, Saint-Amour, Leavitt, Javitt, et al., 2007; Tye-Murray, Sommers, & Spehar, 2007a, 2007b; Wright, Pelphrey, Allison, McKeown, & McCarthy, 2003). The magnitude of this super-additivity varies across the literature, ranging from modest (+6% for Tye-Murray et al. (2007b)) to large (40% for Ewertsen and Nielsen (1971) and Ma et al. (2009)). The super-additivity in the performance gain indicates that the integration process must either employ a gain modulatory process, a noise depression process, or apply constraints to reduce the number of candidates for lexical selection.

There are three prevailing theories for explaining integration, and therefore super-additive performance. These theories will be revisited in greater detail in later sections of this thesis. The first theory, developed for defined signals (neural spiking rates), is that of neural convergence, which states that integration results when specialized multisensory neurons receive inputs from unisensory neurons. The multisensory neurons amplify the
converging input and the result is a stronger signal than the sum of the inputs (Meredith & Stein, 1986). This approach – despite persisting controversies that these findings are not easily replicated in awake animals (A. Bell, Corneil, Munoz, & Meredith, 2003; A. H. Bell, Meredith, Van Opstal, & Munoz, 2005; Populin & Yin, 2002) – forms the basis for much of the research regarding AV speech integration.

The second theory is that of independent processing, wherein each sensory stream produces an independent estimate of the stimulus and these estimates are combined into a composite estimate (Blamey, Cowan, Alcantara, Whitford, & Clark, 1989; Braida, 1991; Massaro & Stork, 1998; Tye-Murray et al., 2007b). The fact that the integration process occurs after a token estimate of the solution is made for each input means that these integration models are also known as “late-stage” integration models. These models can also be conceptualized as a late stage constraints process, wherein each unimodal estimate is scrutinized based on some criterion that restricts the set of candidate solutions, effectively decreasing the noise in the speech signal.

The third theory is that of dependent processing, wherein the unisensory signals interact and are modified prior to integration. In these models there is an interaction between the sensory signals before they are combined and before an estimate of the stimulus is made (van Wassenhove, 2007). This is an “early-stage” integration model. One or both of the sensory signals is subject to constraints prior to integration, where the constraints are derived from correlations of the unimodal inputs. For example, the temporal correlation of an auditory feature and a visual feature could provide information about the speech element, narrowing the number of possibilities for the solution. This
leads to the pruning of the unimodal signal(s), boosting signal clarity, which in turn leads to super-additive performance after integration. Temporal dynamics is the key difference between this type of model and that of independent processing. The dependent processing structure implies that the speech signal is modified online according to bimodal-specific information, such as temporal correlations between the auditory and visual signals. Models formalizing this kind of processing are complicated by the time-dependence, as there are many questions regarding how such a time-varying mechanism could be implemented.

**Features of Super-Additivity**

Multisensory research describes multi-sensory signal enhancement as being an inverse function of the input clarity. That is to say, as the unimodal inputs weaken, the signal enhancement increases. This phenomenon is often referred to as the principle of inverse effectiveness (PoIE). PoIE originates from neurophysiological research by Meredith and Stein (1986) on anaesthetized cats; in their experiment spiking rates were recorded from neurons in the superior colliculus responding to audiovisual stimuli (flashes and noise burst combinations). This result was not successfully replicated in awake cats (Populin & Yin, 2002). The original conception of PoIE is based on the theory of neural convergence and states that the percentage gain of a multisensory neuron’s response is greatest when the unisensory neural signal(s) is/are weakest. That is to say, the degree of response enhancement to a bimodal presentation decreases as the quality of the unimodal signal(s) increases. As a consequence, the bimodal gain in the
response may fall below super-additive, and sometimes below additive, at higher signal clari
ties (Stanford & Stein, 2007).

The PoIE has also been reported in behavioural research (Calvert, Brammer, & Iversen, 1998; Serino, Farne, Rinaldesi, Haggard, & Ladavas, 2007; Stein, Huneycutt, & Meredith, 1988), wherein investigators reported the greatest behavioural performance enhancement occurring with the weakest unimodal performances. There is a long-history of AV speech research describing the modulation of AV performance enhancement being a function of auditory signal clarity (Binnie, Montgomery, & Jackson, 1974; Erber, 1969, 1975; McCormick, 1979; Neely, 1956; Sumby & Pollack, 1954), and a number of investigators have since studied the PoIE through non-invasive population response studies (Callan et al., 2003; Calvert, Campbell, & Brammer, 2000; Calvert, Hansen, Iversen, & Brammer, 2001; Stevenson et al., 2012; Stevenson & James, 2009).

Strict adherence to the PoIE has been challenged by several recent studies that propose it is either not applicable to AV speech-in-noise (Ross, Saint-Amour, Leavitt, Javitt, et al., 2007; Ross, Saint-Amour, Leavitt, Molholm, et al., 2007; Tye-Murray, Sommers, Spehar, Myerson, & Hale, 2010) or only sometimes applicable to behavioural data (Holmes, 2007). Moreover, imaging studies have indicated that the PoIE may not be applicable to the BOLD signal associated with integration processes (Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Laurienti, Perrault, Stanford, Wallace, & Stein, 2005). These studies highlight well-founded concerns regarding the applicability of the PoIE beyond its original description in neurophysiological research.
A fundamental limitation regarding the application of the PoIE to behavioural identification data is that these data, as compared to neurophysiological or detection data, are likely measuring different signals. This difference is well illustrated by comparing performances on detection and identification tasks. For example, while neural responses and behaviour on a detection task are closely associated (Stein et al., 1988), the signal required to perform correctly on a whole word identification task is different from the signal required to perform correctly on a detection task. This is because identification requires the convergence of multiple streams of information to categorize and identify a stimulus, which is a more complex process than that which registers the presence or absence of a stimulus. Performance on an identification task therefore underestimates the information available in a signal, and for this reason, the performance on an identification task cannot be assumed to be a direct measure of integration per-se. Rather, it is a composite of integration as well as number of lexical elements, such as access and facilitation. Nevertheless, there is a consistent finding that transcends this controversy: the degree of performance enhancement is a function of the strength of the inputs.

**Floor and Ceiling Effects**

Behavioural results for an open-set word identification task, with performance correctness on the ordinate axis and the manipulated variable on the abscissa (usually signal-to-noise ratio (SNR)), is bound between 0 and 1, corresponding to completely inaccurate performance (the performance “floor”) and completely accurate performance (the performance “ceiling”), respectively. Because performance cannot go beyond these
values, the floor and ceiling will exert effects on average performance (across subjects) as soon as some of the subjects start to hit the floor or ceiling.

Subtler floor/ceiling effects are exerted before the functions saturate at 0 or 1 when the functions first begin to show non-linear behaviour. The result is that the performance curve is de-linearized and the variance is reduced near the asymptotes. To confidently correlate performance with the manipulated parameter, analysis should be restricted to the intermediary region of the psychometric curve. The confusion arising from this concept makes the case for the collection of the full performance curve so that the saturation point of each curve can be estimated – as was done in this thesis. Alternatively, the Speech Reception Threshold (SRT) measures (MacLeod & Summerfield, 1990) also avoids these effects by manipulating SNR to achieve a standardized performance. For example, the experimenter will present stimuli and manipulate the SNR level until a performance of 50% is achieved. Because the investigator chooses the performance level he or she wishes to measure, a performance level can be chosen that is likely free from floor and ceiling effects.

**Variability**

AV performance is highly variable across subjects. This variability, which is greater than that associated with AO or VO performances, has been reported for single words (Blamey et al., 1989; Erber, 1969; Grant, Walden, & Seitz, 1998; Middelweerd & Plomp, 1987), sentences (Ewertsen, Nielsen, & Nielsen, 1970; MacLeod & Summerfield, 1987, 1990), syllables (Grant & Seitz, 1998; Middelweerd & Plomp, 1987), and
consonants (Blamey et al., 1989; Grant & Seitz, 1998). This variability could be due to properties of the stimuli, such as word-frequency or speaker effects. Alternatively, it could be due to properties of the subject, such as the efficacy of the multimodal integration process, lexical access, or working memory capacity (Akeroyd, 2008).

Speaker characteristics exert profound effects on the intelligibility of the speech signal. Evidence points to contributions from the fundamental frequency (Bradlow, Torretta, & Pisoni, 1996; Laures & Bunton, 2003; Lu & Cooke, 2009), duration cues (Bond & Moore, 1994), lexical stress (Field, 2005) and vowel space (Bradlow et al., 1996), all of which can vary across speakers. These results indicate that speech quality is not a scalar value, but rather is complex time-varying signal. And although some studies have looked at the impact of different speakers on the intelligibility of speech-in-noise (Barker & Cooke, 2007; Grant & Braida, 1991; O'Neill, 1954), there has not been a consistent effort to control such speaker effects. The failure to systematically quantify the complex attributes of speech quality has likely contributed to the wide variety of results reported from AV speech-in-noise performance studies.

The quantification of masking noise is also complex with the efficacy of the noise mask depending on a number of parameters, such as the frequency of the mask, whether or not the noise is stationary, glimpsing effects, and temporal correlates with the target signals. In AV speech-in-noise studies, one of six noise types is typically used: white noise (Blamey et al., 1989; Dodd, 1977; Erber, 1969; Ewertsen & Nielsen, 1971; MacLeod & Summerfield, 1987, 1990; O'Neill, 1954; Robert-Ribes, Schwartz, Lallouache, & Escudier, 1998; Sumby & Pollack, 1954; Watson, Qiu, Chamberlain, &
Li, 1996), pink noise (Ma et al., 2009; Ross, Saint-Amour, Leavitt, Javitt, et al., 2007;
Ross, Saint-Amour, Leavitt, Molholm, et al., 2007), broadband noise (Binnie et al., 1974;
Rosenblum, Johnson, & Saldaña, 1996), low-frequency noise (Erber, 1971a), speech-
shaped noise (Barker & Cooke, 2007; Grant & Braida, 1991; Grant & Seitz, 1998; Grant
et al., 1998; Smith & Bennetto, 2007) or multi-talker babble noise (Callan et al., 2003;
Munhall, Kroos, Jozan, & Vatikiotis-Bateson, 2004; Sommers et al., 2005; Tye-Murray
et al., 2007a, 2007b; Tye-Murray et al., 2010). The precise interaction between talker and
noise-type is not defined in these studies. This is further complicated by the fact that
talkers’ speech may vary in relative amplitude over time, meaning the interaction
between noise and intelligibility can also vary over time. This additional degree of
freedom further complicates between-study comparisons. As such, the reported effect of
SNRs on intelligibility is not easily interpreted.

Previous attempts have been made to account for this variability through subject
effects by examining the correlation between AV and the unimodal performances across
subjects (Blamey et al., 1989; Grant et al., 1998; Middelweerd & Plomp, 1987; Smith &
Bennetto, 2007; Tye-Murray et al., 2007a; Watson et al., 1996). A significant correlation
between AV performance and VO performance has been observed for sentence stimuli
(Grant et al., 1998; MacLeod & Summerfield, 1987), although this was not found to be
the case for consonants (Grant et al., 1998). Research has also attempted to weigh model
predictions based on how different individual subject process AO and VO information
(Schwartz, 2010) to remove the between-subject variability. Despite these efforts, there
remains a high degree of residual variance in AV performance that is unaccounted for.
Speculation regarding the missing variance has led investigators to conclude that the integration mechanism could have its own source of variance that cannot be measured by performance in the unimodal experiments (Watson et al., 1996).

**Theories of AV Speech Integration**

*Neural Convergence*

The theory of neural convergence, based on the work of Meredith and Stein (1986) in the cat superior colliculus, describes signals from sensory neurons converging on multisensory neurons. Multiple inputs are summed in the multisensory neuron, which amplifies the signals and outputs an enhanced multisensory signal. Integration therefore occurs at a specific discrete location. Neurophysiological and anatomical studies describe multisensory neurons in the superior-temporal-sulcus (Benevento, Fallon, Davis, & Rezak, 1977; Bruce, Desimone, & Gross, 1986), the orbito and pre-frontal cortices (Benevento et al., 1977; Fuster, Bodner, & Kroger, 2000), as well as the posterior parietal cortex (Leinonen, Hyvärinen, & Sovijärvi, 1980). Evidence from functional imaging studies also describes multisensory convergence associated with speech occurring in the superior-temporal-sulcus (STS) (Callan et al., 2003; Lee & Noppeney, 2011; Murase et al., 2008; Reale et al., 2007; Wright et al., 2003), making the STS the front-running candidate for the site of this integration process. However, additional studies have indicated that the integration process involves a more distributed network of systems (Skipper, Nusbaum, & Small, 2005; Szycik, Jansma, & Munte, 2009; Wilson, Molnar-
szakacs, & iacoboni, 2008), suggesting the STS might only be a component of a larger process.

Independent Processing

In an independent model, the unimodal streams are processed separately, each producing their own estimates of the stimulus, at which point the estimates are integrated into the final solution. Models that assume this kind of processing are the PRE-labeling model (braida, 1991), the Fuzzy-Logic-Model-of-Perception (FLMP) model (massaro & friedman, 1990; massaro & stork, 1998), the PROB model (blamey et al., 1989) and the Neighbourhood Activation Model (NAM) (Luce & Pisoni, 1998; Mattys, Bernstein, & Auer, 2002; Tye-Murray et al., 2007b).

In the PRE-labeling model each unimodal input differentially activates a feature vector in speech “feature space” which is matched against an internally stored speech element representation. For example, in the auditory stream the signal may carry strong information on nasality and voicing while the visual stream may carry strong information regarding the place of articulation. Each of these elements, nasality, voicing and place of articulation, correspond to an axis in the speech feature space. Each internal representation - be they sub-phonemic representations, phonemes or even whole words - is also associated with a particular point in the space. The estimate of the stimulus will be the representation that is the shortest Euclidean distance to the point in feature space.

This model accounts for super-additivity by using the complementary information to restrict the set of possible solutions. While the AO performance (in a whole word
identification task) might be very low, adding visual information effectively “fills in the blanks” in the deficient auditory signal, allowing for identification in the AV condition. Since a given estimate is assigned to an internal representation, the noise that knocks the estimate off the “true” value is eliminated – provided the noise is not so great that it knocks it closer to another internal representation. This model requires a common neural coding between the AO and VO streams such that they can be combined into a common feature space representation. It should be emphasized that the features chosen in the aforementioned example (nasality, voicing and place of articulation) have been shown to predict the intelligibility of consonants (Miller & Nicely, 1955), but they were chosen here for illustrative purposes only. The true nature of the speech encoding dimensions is unknown.

In the FLMP models each unimodal stream processes its respective sensory input and presents a series of estimates of all possible solutions to the integration mechanism which then combines these estimates into the final solution. The estimates vector will be composed of a set of probabilities associated with each possible solution. For example, if a person hears the word “foot” presented in noise, their estimates might resemble [foot (0.75), hood (0.20), put (0.05)] where values in brackets represent the probability associated with each solution. In this case there are 3 possible solutions and the most likely solution is the word “foot” with 75% likelihood. When there are two channels of information, the integration of these estimates is an element-by-element multiplication of the two vectors followed by a normalization process based on Maximum-likelihood estimation (Massaro & Friedman, 1990). The final decision is made based on the final
solution with the greatest probability. This model accounts for super-additive performance by using a likelihood estimate as a constraint to the possible solutions. Since the ‘most likely’ solution is chosen, the system can handle noise and still produce the correct solution provided the noise isn’t so great that it will attenuate the likelihood of the correct solution below that of an incorrect solution. This model’s structure necessarily requires two things. First, the task must be closed-set because the AO and VO estimate vectors must be finite in length (equal to the number of possible solutions). Second, the estimate vectors must be the same length, meaning the number of elements in the AO estimate vector must be equal to the number of elements in the VO estimate vector so that they can be multiplied. The AO and VO estimates are therefore associated with the same set of possible solutions. This is a tenuous assumption because the lexical estimates of AO and VO speech can be quite different (Tye-Murray et al., 2007b), indicating that each channel could provide very different estimates of the set of possible solutions. This problem is avoided in implementations of FLMP by using a closed-set task, usually consonant identification. Again, it is important to note that the elements that make up the estimate vectors are not defined. In our aforementioned example they were estimated as words, however, the actual estimate vectors would encode likelihoods of the speech elements at the integration stage, which could be at the phonemic or pre-phonemic level.

The Neighbourhood Activation Model (NAM) proposes that the integration process is not an explicit mechanism, but rather that it is a consequence of the co-activation of lexical prototypes by the auditory (Luce & Pisoni, 1998) and visual stream (Mattys et al., 2002; Tye-Murray et al., 2007b). In this model, words stored in the lexicon
are the solution prototypes. The auditory and visual streams automatically activate regions of the lexicon consistent with their inputs that will be confusable with the stimulus. These activated sets of word prototypes are called “lexical neighbourhoods”. The overlap of these neighbourhoods will lead to the AV percept.

The auditory neighbourhoods are distinct from the visual neighbourhoods because the auditory and visual channels encode largely complementary, rather than redundant, information. For instance, the auditory signal is known to carry information on voicing and nasality, while signals for place-of-articulation are weak. Conversely, the visual signal carries robust information for place-of-articulation but weak information regarding voicing and nasality. This distinctiveness leads to small intersections between the neighbourhoods from which the solution is selected. For example, if the word “cat” were presented to the subject, the visual neighbourhood could include the words: “cat”, “hat” and “gnat”, all of which are visemically equivalent. The auditory neighbourhood could include words that are phonemically similar, such as: “cat”, “cast”, and “kit”. The only overlap between the auditory and visual neighbourhoods in this case is the word “cat”, which is the response model prediction, and the correct solution. This model therefore accounts for super-additivity by inducing lexical constraints on the set of possible solutions such that solutions that are either only supported by auditory information (but not visual information) or solutions that are only supported by visual information (and not auditory information) are automatically eliminated. This reduces the number of alternatives from which a response must be chosen, boosting the likelihood that the system will choose the correct solution. By forcing the system to consider only
those solutions that intersect with both the visual and auditory solution sets, the set of possible solutions from which the AV solution is chosen is greatly reduced. This process therefore effectively depresses the noise in the speech signal.

The PROB model is based on the seminal work of Miller and Nicely (1955), and uses an information theoretic approach to estimate AV performance. First, the information content in the auditory and visual channels is estimated using unimodal confusions (Blamey et al., 1989; Ronan, Dix, Shah, & Braida, 2004). Like FLMP and PRE, this approach requires a closed-set task for the calculation of the probability of a given response and it also requires that the auditory and visual streams be statistically independent. Because the auditory and visual channels are assumed to be independent sources of information, this approach does not address super-additivity in performance. The predicted informational content of the AV signal is equivalent to a corrected linear sum of the unimodal performances – and since AV performance is known to be super-additive, this model therefore underestimates performance (Blamey et al., 1989). However, this approach could provide an estimate of the degree that AV performance deviates from a linear sum model, thereby providing a measure for the amount of super-additivity.

**Class-Conditional Independence**

All independent models share a common assumption: that intermodal interactions are minimal for the purpose of mathematical tractability. This is called the assumption of “class conditional independence”. Mathematically, it can be summarized as follows:
\[ P(R_{AV}|c) = P(R_{AO}|c) \times P(R_{VO}|c) \]

This is also a fundamental assumption of Bayesian statistics wherein the joint probability of an event happening is equal to the product of the individual probabilities that each event will happen. Conceptually, this formula means that the probability of the subject giving response \( R_{AV} \) when they are presented with the stimulus category “\( c \)” is equal to the product of the probabilities of the subject giving responses \( R_{AO} \) and \( R_{VO} \) when presented with the unimodal presentations of stimulus category “\( c \)”. For example, in the AV presentation of the word “bat” the subject might be 75% likely to correctly identify the word as “bat”. For the auditory and visual streams to be independent, the product of the probabilities for the unimodal presentations of the word “bat” in the AO and VO condition must also be 0.75, one possible solution is AO being 93% likely and VO being 83% likely. This might seem counter-intuitive since the joint probability is less than the input probability. This is because any probability less than 1 introduces the possibility of errors and therefore the joint probability must reflect union of the likelihood for errors in either condition. In a Bayesian estimate this decrease in selection probability is offset by a normalization term – the sum of the likelihood of all possible responses (in a closed-set task).

The appeal of a model based on this assumption is that the unimodal behavioural data can be combined to produce numerical predictions of bimodal performance (Braida, 1991; Grant, 2002; Grant & Walden, 1996; Massaro & Cohen, 2000). This provides a readily testable means of evaluating the quality of the model. Unfortunately, as we will now discuss, there is a growing body of physiological and behavioural evidence that the
integration process involves cortical and subcortical interactions that would violate this assumption.

Cortical Interactions

In this section we will review evidence that VO signals elicit activity in different cortices and facilitate speech perception. Consequently, it has been proposed that VO speech information has access to auditory cortex, meaning that there is an opportunity for the visual speech information to modify the auditory signal prior to integration. Interactions between the sensory cortices would violate the assumption of class-conditional-independence required for independent processing models.

Anatomical studies in non-human primates have found connections between primary visual and auditory cortex through which these interactions could be occurring. Auditory cortex has been found to receive inputs from the peripheral visual field representations of V2 (Falchier, Clavagnier, Barone, & Kennedy, 2002; Falchier et al., 2010; Groh, Trause, Underhill, Clark, & Inati, 2001; Rockland & Ojima, 2003). This connectivity with the peripheral visual field could facilitate localizations of sounds in the peripheral visual field. Evidence of the interaction of the STS and the auditory belt and parabelt regions has also been demonstrated through a phase coherence between visual and auditory responses (Ghazanfar, Chandrasekaran, & Logothetis, 2008; Ghazanfar, Maier, Hoffman, & Logothetis, 2005).

There is also evidence that visual responses can be modified by sound. Falchier et al. (2002) identified projections to primary visual cortex from core and parabelt regions
of auditory cortex, as well the multisensory region of the polysensory region of the temporal lobe in primate cortex. And it has been shown that the visual response can be modified by sound for the integration of audiovisual non-speech stimuli (Watkins, Shams, Tanaka, Haynes, & Rees, 2006). Further, the early onset of the auditory speech presentation during incongruent audiovisual presentations has been shown to modulate visual activity (Jones & Callan, 2003).

The presentation of VO speech alone has been shown to activate or modify activity in auditory cortex as well as multisensory cortex (Calvert et al., 1997; Calvert & Campbell, 2003; Mottonen, Schurmann, & Sams, 2004; Paulesu et al., 2003; Pekkola et al., 2005; Sams et al., 1991) – though this is not always the case (Bernstein et al., 2002). This modification does not hold for meaningless lip movements (Campbell et al., 2001), indicating that visual speech in particular could have access to auditory and multisensory cortex. It has also been proposed that the visual signals modify the auditory signal by putting the neurons in auditory cortex into a “high excitability” state – wherein the baseline neural activity is raised closer to the firing threshold, making the neural population more responsive to incoming excitatory signals - via long-range oscillations (Schroeder, Lakatos, Kajikawa, Partan, & Puce, 2008). Provided that the phases of the “high excitability” state coincides with phases of the auditory signals that aid in identification of the stimulus, this mechanism amplified the signal, thus boosting intelligibility.

Cortical interactions are also consistent with observed intermediate activity, such as facilitative processing (van Wassenhove, Grant, & Poeppel, 2005). Strictly speaking,
the addition of information should increase the cognitive processing load. If AV speech were being processed strictly via integration of the auditory and visual streams, the addition of visual information should increase processing time compared to auditory-only processing. However, the addition of concurrent visual information has been shown to speed up processing (van Wassenhove et al., 2005), indicating that the addition of concurrent visual information is facilitative to speech perception. It has also been found that functional activity during AV speech perception can be less than the sum of unimodal activity (Kayser & Logothetis, 2007; Wright et al., 2003). Further, the interactions between sensory cortices occur before the signal reaches the STS (Mottonen et al., 2004) and hence before feedback modulation could be occurring (Foxe & Schroeder, 2005). These results suggest that the visual signal has direct access to the auditory signal and could be decreasing the processing load associated with speech perception. This could occur via pruning of the auditory signal to reduce noise.

A possibly mechanism through which these cortical interactions might be occurring is via the coherent oscillations of neural signals. The coherence of oscillatory activity in disparate regions of cortex could serve to form and sever functional connectivities between populations of neurons. This has been referred to as the ‘integration through coherence’ hypothesis (Engel, Fries, & Singer, 2001; Fries, 2005). This framework presents a flexible and specific mechanism for the transmission of information via the modulation of functional connectivities. This is an important advantage over the classical model because goals and perceptual contexts are constantly changing, and this behaviour is consistent with a dynamic and flexible network.
Additionally, this hypothesis provides an explanation for how disparate signals are flagged for integration (Senkowski et al., 2008).

**Dependent Processing**

The dependent processing model is predicated on the assumption that cortical interactions lead to the modification of one or both unimodal signals. That is to say, the processing of one unimodal stream – usually the auditory stream – is shaped by the processing of the other unimodal stream – usually the visual stream (Macaluso, 2006; Schroeder et al., 2008; van Wassenhove, 2007). The processing of one stream therefore depends on the processing of the other stream. An unappealing implication of this kind of processing is that there is an unknown, non-trivial relationship between unimodal and bimodal performances as the brain is either pruning the information in the unimodal streams or supplementing the unimodal signals with intermodal cues. Both of these elements are inaccessible during unimodal presentations, which means that bottom-up models that use the unimodal performances to predict bimodal performance will under-predict performance.

Currently, there is only one model that accounts for modification of the processing of the auditory signal during AV presentations. The Analysis-by-Synthesis model developed for AV speech integration by van Wassenhove (2007) was based on previous work by Halle and Stevens (1962) as well as the Motor Theory of Speech perception (Liberman & Mattingly, 1985; Liberman & Whalen, 2000). In this model, the visual signal precedes the audio signal and ‘sets-up’ the system to selectively process
auditory information – this process is termed “predictive coding”. Super-additive performance is therefore due to constraints applied by the visual information on the subsequent processing of the auditory signal. This reduces the noise in the auditory signal. The AV percept is therefore congruent with both the visual and auditory signals and the overall processing load is reduced. This processing would explain why auditory-specific ERP signals are attenuated during AV presentation (van Wassenhove, 2007) and why bold activity is sometimes reduced in AV presentations (Kayser & Logothetis, 2007; Wright et al., 2003).

**Phonemic and Pre-phonemic integration**

A reoccurring question in the previously described models is the lexical stage at which integration occurs. There is substantial evidence that speech signals are encoded via a sub-phonemic amodal intermediate that maps to the lexicon. Skipper, van Wassenhove, Nusbaum, and Small (2007) indicated that these prototypes could be stored as gestural speech code elements in frontal motor areas. Poeppel (2001) examined patients suffering from auditory verbal agnosia and suggested that speech processing is mediated through bilateral amodal representations before interfacing with the lexicon. Conversely, work in the field of language acquisition in infants indicates that the discovery of statistical and prosodic patterns leads to phoneme and word recognition - see Kuhl (2004) for a review. If language development mirrors the construction of a hierarchical language-processing scheme then these results are indicative of sub-phonemic encoding preceding phonemic and word encoding. Further evidence for sub-
phonemic processing was provided by Green (1998), who observed that cross-modal speech influences occur at the feature level, such as voice-onset time, which is more fundamental than phoneme representations. Other studies have found that visual information could modify the perception of the voicing boundary (Green & Miller, 1985), the decision boundary between adjacent phonemes (Green & Norrix, 2001), as well as the place of articulation and voiced formant transitions (Green & Norrix, 1997). Overall, most studies are consistent with incoming sensory information being first assembled into sub-phonemic representations before it is assigned a lexical or phonemic token.

**The McGurk Illusion**

The McGurk Effect, discovered by McGurk and MacDonald (1976), is a perceptual illusion used to measure the degree of AV speech integration. In this illusion, concurrent but incongruent AV stimuli are presented to a subject. For example, when a subject is presented with an auditory /aba/ utterance played over a visual /aga/ utterance, they will often hear an illusory utterance, for this example common illusory responses are /ada/ and /atha/. Because the illusory response is clearly heard, this illusions was originally described as a an example of how the visual signal was able to modify the auditory signal (MacDonald & McGurk, 1978), but the definition has evolved and it is frequently used as a proxy measure of the AV speech integration mechanism (Massaro & Cohen, 1993). In McGurk studies, the frequency of illusory responses – called McGurk susceptibility – is used as a measure of the subject’s binding efficiency. Experiments that
vary parameters that modulate the susceptibility – such as stimulus-onset-asynchrony (Munhall, Gribble, Sacco, & Ward, 1996; van Wassenhove, Grant, & Poeppel, 2002) – are therefore assumed to vary parameters that modulate the integration process.

The susceptibility to the McGurk illusion is influenced by age (McGurk & MacDonald, 1976; Rosenblum, Schmuckler, & Johnson, 1997), cochlear implant usage (Skipper et al., 2007), rate of the visual presentation (Green & Miller, 1985), stimulus onset asynchrony (Jones & Callan, 2003; Massaro & Cohen, 1993; Munhall et al., 1996; Tomaskovic, Wiersinga-Post, Slabu, Renken, & Duifhuis, 2008; van Wassenhove et al., 2002), different speakers (Cienkowski & Carney, 2002), noise masking (Sekiyama, Kanno, Miura, & Sugita, 2003), attention effects (Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Alsius, Navarra, & Soto-Faraco, 2007; Soto-Faraco, Navarra, & Alsius, 2004), lexical effects (Brancacio, 2004), context effects (Windmann, 2004), and priming (Kilian-Hutten, Vroomen, & Formisano, 2011). It has also been found that susceptibility requires the subject be processing the visual stimulus as a face (Munhall, ten Hove, Brammer, & Paré, 2009) and the audio as speech (Vroomen & Stekelenburg, 2011). Despite extensive work using the McGurk as a proxy measure for speech integration, its correlation with speech performance has not been tested.
Scope of Studies:

Four studies were conducted on three data collections (corresponding to three different groups of subjects): a word identification task with static noise masking (pink noise) and a McGurk effect study. The same video recordings of the stimulus words were used in the word identification studies. Further, the McGurk stimuli were collected from the same speaker as the word stimuli. In the first data collection – used for the first study – performance data was collected for AO and AV presentations of words at 7 SNR levels (-10 to +14 dB). In the second data collection – used for the second and third study – performance data was again collected for AO and AV presentations of words, but at 6 SNR levels (-10 to +10 dB) and a speechreading condition was also included (presented in silence). In the last data collection – used for the fourth study – performance in AO and AV presentations of words was collected at a single SNR level (-5 dB), VO performance was collected in silence and responses to five different incongruent CVC nonsense syllables (McGurk stimuli) were recorded.

The goal of the first study was to assess the viability of the stimuli created for this project and to investigate sources of variability in the AV signal. The wordlist and video recordings that were created were confirmed to produce significantly greater performance in the AV as compared to the AO condition. Both the AO and AV performance functions were consistent with smooth sigmoids that monotonically increased with SNR. The AV performance exhibited a significant amount of variance that could not be accounted for by AO performance or interactions with word-frequency effects. Further, the variability in the AV stream was found to be SNR dependent,
greatest at lower SNRs, and decreasing as the auditory signal improved. This result is consistent with a process that is variably recruited in different people at low SNRs. These results were consistent with the AV process containing an independent source of variance outside of the auditory stream, making the between-subject variability in performance high.

The second study involves developing a quantitative measure of AV performance gain that was specifically associated with the bimodal presentation. Building on previously proposed measures for single cell responses (Stein et al., 1988) and information theoretic approaches to AV speech perception (Blamey et al., 1989), a measure for integration enhancement (IE) was developed for an open-set task. This measure is the difference between the actual AV performance and the corrected sum of the unimodal performances. The IE is therefore a measure of the magnitude of the AV performance that is specific to bimodal presentations. This enhancement could be due to a gain modulatory integration, or from the congruency of the auditory and visual information. That is to say, if the auditory information is insufficient to identify whole words, it might be sufficient if visual information fills in the missing piece, bringing the total informational content of the speech signal up to threshold for identification.

In the third study a model is developed to test whether unimodal lexical confusions could account for IE. This model tested the proposal of Tye-Murray et al. (2007b) who asserted that integration occurs at the full lexical level. The response confusions from AO and VO responses to each of the presented words were assembled and compared. When the AO and VO responses overlap over the correct word the model predicted
performance enhancement, and when they overlapped over an incorrect word the model predicts performance depression. The results indicated that the performance enhancement could be accounted for in 60% of the words. However, performance on 18% of the words was predicted to only be additive, and 22% of the words were predicted to have sub-additive performance (the AO and VO responses overlapped over an incorrect solution). Comparatively, the actual AV performance for the latter two groups was super-additive at all SNRs, though the last group did display response depression as compared to the mean. This result could indicated that integration occurs at the sub-lexical level such that auditory information that is insufficient for whole word identification is being supplemented by visual information to produce performance gains that are beyond those predicted by the performance in an identification task.

The final study compares the IE measure with the susceptibility to a commonly used proxy of AV integration, the McGurk illusion. The McGurk illusion refers to the illusory auditory percept elicited by the presentation of certain incongruent AV stimuli – meaning the auditory and visual tracks are mismatched. The frequency with which a person perceives this illusion has been proposed as being either a measure of the modulatory capacity of visual information (MacDonald & McGurk, 1978) on the auditory percept or a measure of the person’s integration ability (Massaro & Cohen, 1993). However, the explicit link between McGurk susceptibility and performance enhancement for AV speech has not been thoroughly investigated. The results from this experiment indicate that McGurk susceptibility is a negative correlate of VO performance and not the bimodal process involved in AV performance enhancement. These results indicate that it
is dependent on unimodal processes and is therefore not a suitable proxy for the integration mechanism involved in degraded bimodal speech perception.

Research regarding multi-sensory integration is well developed but is nonetheless based on a number of assumptions, and involves competing theories regarding how integration is achieved. To address these assumptions and to use human behaviour to examine how integration occurs, and specifically how super-additive performance results from such a process, we undertook to evaluate, in detail, the variability in the AV signal, performance enhancement, and McGurk susceptibility. The results from these studies provide insight into how integration occurs, and how the assumptions described in the literature may be addressed. As well, the results help further clarify which of the above-described theories and processes best account for the integration process and performance enhancement in audio-visual speech perception.
Chapter 2: Performance Variability in Audiovisual Speech-in-Noise Perception

**Keywords:** Audiovisual speech, super-additive performance, word identification task, sweet-spot theory, Principle of Inverse Effectiveness

**Abstract:**

In this study we present the results from an analysis of performance data collected in an identification task of audiovisual speech monosyllabic English nouns in noise. Our results show strong within-subject correlations between auditory-only and audiovisual performance but low between-subject correlations. Interactions with word-frequency effects do not account for the variability, indicating that it is most likely due to subject differences. The variability of performance was analyzed across six signal-to-noise ratios (SNR) and the performance boost associated with audiovisual integration was found to be SNR dependent. Contrary to recent reports of audiovisual performance, no intermediate zone of maximal integration was observed. A cross-study comparison found a large degree of variability in performance, indicating that the experimental methodology also affects performance in bimodal speech perception.

**Introduction:**

Performance in an audiovisual (AV) speech identification task in noise is variable in two ways. First, it is variable across subjects – meaning different subjects, presented with the same stimuli will perform significantly differently. And second, it is variable across studies – meaning the reported performances for AV speech-in-noise tasks differ.
Understanding the sources of this variability is critical to understanding how bimodal speech is perceived. Because variability in performance is high, reproductions of studies are challenging and may require the acquisition of large data sets. Here we examine the within study variability observed in AV performance and compare it to between study variability.

*Between-subject Correlations – Subject Variability:*

Large variability in subject performance in AV speech perception tasks as compared to auditory-only (AO) or visual-only (VO) performance tasks has been reported for single words (Blamey et al., 1989; Erber, 1969; Grant et al., 1998; MacLeod & Summerfield, 1987; Middelweerd & Plomp, 1987), sentences (Ewertsen et al., 1970; Grant & Braida, 1991; MacLeod & Summerfield, 1987, 1990; Mattys et al., 2002), syllables (Grant & Seitz, 1998; Middelweerd & Plomp, 1987) and consonants (Grant & Seitz, 1998). Previous attempts have been made to account for this variability through correlations with AO and VO performances (Blamey et al., 1989; Grant et al., 1998; Middelweerd & Plomp, 1987; Smith & Bennetto, 2007; Tye-Murray et al., 2007a; Watson et al., 1996) as well as lexical neighbourhood densities in the AO and VO conditions (Mattys et al., 2002; Tye-Murray et al., 2007b). A significant correlation between AV performance and VO performance has been observed for sentence stimuli (Grant et al., 1998; MacLeod & Summerfield, 1987), though this was not found to be the case with consonant stimuli (Grant et al., 1998), indicating that VO and AV performances could be linked via high-order lexical facilitation or that there is a
threshold amount of visual data required to facilitate comprehension in the AV condition. Work has also been done to attempt to weigh model predictions based on subject differences (Schwartz, 2010) to remove between-subject variability.

Although a correlation has been found between the performance in the AV and VO tasks, the relationship between audiovisual performance and auditory-only performance is difficult to quantify. Previous studies have suggested that performance in an AV speech perception task cannot be indiscriminately correlated to performance in an AO speech perception task. Grant and Braida (1991) investigated the mapping of the AO Articulatory Index (AI<sub>AO</sub>) to the now seldom used AV Articulatory Index (AI<sub>AV</sub>) for sentences. The AI is a measure that quantifies the intelligibility of speech based on a set of speech features. The values range from 0 (unintelligible) to 1 (perfectly intelligible). They were unable to clearly identify the elements of visual speech that connected AI<sub>AO</sub> to AI<sub>AV</sub> and concluded that a wide array of auditory signals with varying spectral and temporal characteristics contributed to the AV percept. Grant and Walden (1996) revisited this investigation with consonants and once again, a clear mapping between AO and AV was not identified and it was concluded that intelligibility measures in AO could not be used to predict AV performance. Therefore there remains a large amount of residual variability in AV performance that is unaccounted for by the unimodal performances. This variability could be due properties of the stimuli, such as word effects, speaker effects or volume effects—wherein intelligibility varies with differing volume levels at the same signal-to-noise ratio (SNR; Duquesnoy, 1983). The variability
could also be due subject differences, such as the efficacy of the subject’s multimodal integration process, lexical access or working memory capacity (Akeroyd, 2008).

Speculation over the unaccounted for variance has led investigators to conclude that the integration mechanism must have its own source of variance that cannot be measured by performance in the unimodal experiments (Watson et al., 1996). It has also been proposed that visual speech information has access to auditory cortex - meaning that there is an opportunity for the visual speech information to modify the auditory signal. The modification of the auditory signal could introduce another source of variance that could not be accounted for with the unimodal performances. The presentation of VO speech has been shown to activate or modify activity in auditory cortex as well as multisensory cortex (Calvert et al., 1997; Calvert & Campbell, 2003; Mottonen et al., 2004; Paulesu et al., 2003; Pekkola et al., 2005; Sams et al., 1991). This modification does not hold for meaningless lip movements (Campbell et al., 2001), indicating that visual speech in particular has access to auditory and multisensory cortex.
Modulation of Performance Enhancement

Previous investigators have indicated that at low SNRs performance enhancement is maximal and that it decreases with increasing SNRs; meaning, the greatest ‘boost’ in performance will occur at the lowest SNRs and the ‘boost’ will decrease as the SNR is increased. This phenomenon is generally referred to as “inverse effectiveness” and it has been observed in behavioural (Bolognini, Frassinetti, Serino, & Ladavas, 2005; Grant & Seitz, 2000; Stein et al., 1988) and non-invasive population response studies (Callan et al., 2003; Calvert et al., 2000; Calvert et al., 2001; Stevenson et al., 2012; Stevenson & James, 2009). In AV speech-in-noise tasks, this modulation of performance gain has also been observed (Binnie et al., 1974; Erber, 1969, 1975; McCormick, 1979; Neely, 1956; Sumby & Pollack, 1954).

The applicability of inverse effectiveness to bimodal speech perception has been challenged by functional imaging studies (Beauchamp et al., 2004; Laurienti et al., 2005) and behavioural data (Holmes, 2007; Ross, Saint-Amour, Leavitt, Javitt, et al., 2007; Ross, Saint-Amour, Leavitt, Molholm, et al., 2007).

Evidence has been found that performance enhancement for identification tasks should instead be associated with a performance boost that is maximal at an intermediate noise level (reported as -12 dB by Ross, Saint-Amour, Leavitt, Javitt, et al. (2007) rather than the lowest noise levels (see also Ma et al. (2009). The zone of maximal integration or "integration sweet-spot" that was first described by (Ross, Saint-Amour, Leavitt, Javitt, et al., 2007) proposed a radically different model of integration. This model asserted that the integration process is a varying function of SNR that is maximally tuned
to a particular noise masking level - analogous to a neural tuning curve. Neural tuning curves associated with outer hair cells, for example, show increased sensitivity at certain sound frequencies. That is to say, the threshold for firing is lower at a particular frequency band. For this analogy to hold, the integration process must be preferentially tuned to a particular SNR level. Ross et al. (2007) determined this value to be -12 dB for their task, which used pink noise. These controversies notwithstanding, a consistent finding is that the amount of performance gain in AV speech is a function of the strength of the input signal(s).

Floor and ceiling effects are also a critical limiting factor in behavioural studies for the purposes of investigating AV speech integration; the performance “floor” is at 0% and the “ceiling” is at 100%. Because performance cannot go outside of these values, the floor and ceiling will exert effects on average performance as soon as some of the subjects start to hit the saturations points. The result is that the performance curve is de-linearized and the variance is reduced near the asymptotes. As a result of the suppression of variance, correlational analyses cannot be performed for these data points.

**Methodological considerations - SNR effects:**

SNR measures are widely accepted as the parametric scale used to manipulate the intelligibility of auditory and audiovisual speech. This measure is determined by two factors: the speech signal and the noise mask. The SNR value, measured in decibels is ten multiplied by the log of the power of the speech signal divided by the power of the noise mask. The lower the SNR, the less intelligible the speech will be. In AV speech,
the signal and noise have a complicated relationship that could be contributing to large
between-study variability in intelligibility.

Speech intelligibility is a complex quantity. Efforts to identify correlates indicate
contributions from the fundamental frequency (Bradlow et al., 1996; Laures & Bunton,
2003; Lu & Cooke, 2009), duration cues (Bond & Moore, 1994), lexical stress (Field,
2005) and vowel space (Bradlow et al., 1996). Further complicating the issue is
individual variability. Individual speaker characteristics will exert strong effects on the
intelligibility of the speech signal (Barker & Cooke, 2007; Grant & Braida, 1991;
O'Neill, 1954), as will the characteristics of the masking noise. The efficacy of the noise
mask depends on the number of parameters—such as the frequency distribution of the mask,
whether or not the noise is stationary, glimpsing effects and temporal correlates with the
target signals. In AV speech-in-noise studies, one of six noise types is typically used:
white noise (Blamey et al., 1989; Dodd, 1977; Erber, 1969; Ewertsen & Nielsen, 1971;
MacLeod & Summerfield, 1987, 1990; O'Neill, 1954; Robert-Ribes et al., 1998; Sumby
& Pollack, 1954; Watson et al., 1996), pink noise (Ma et al., 2009; Ross, Saint-Amour,
Leavitt, Javitt, et al., 2007; Ross, Saint-Amour, Leavitt, Molholm, et al., 2007),
broadband noise (Binnie et al., 1974; Rosenblum et al., 1996), low-frequency noise
(Erber, 1971a), speech-shaped noise (Barker & Cooke, 2007; Grant & Braida, 1991;
Grant & Seitz, 1998; Grant et al., 1998; Smith & Bennetto, 2007) or multi-talker babble
noise (Callan et al., 2003; Munhall et al., 2004; Sommers et al., 2005; Tye-Murray et al.,
2007a, 2007b; Tye-Murray et al., 2010). The interaction between talker and noise-type is
complex and time-varying that leads to widely varying measures of intelligibility. As
such, the reported effects of SNR on intelligibility are not easily interpreted. These degrees of freedom complicate between-study comparisons and have likely contributed to the wide variety of results reported from AV speech-in-noise performance studies.

The overarching goal of this study was to examine variability in AV speech-in-noise performance; we approached this in two ways. First, we examined between-subject variability, specifically its relationship to SNR level and to see if it could be explained by an interaction with word-frequency effects. And second, we examined between-study variability and whether this variability could be accounted for by normalizing the SNRs.

**Methods:**

**Participants:**

Seventy subjects (11 men; mean age of 20.4 years, std = 1.47 years) were recruited from the Queen's University Psych 100 subject pool (compensated with a course credit) as well as a paid-participant list (compensated with $10). All subjects were native English speakers between 18-24 years of age with self-reported normal hearing and vision (including corrected-to-normal). All subjects provided informed consent prior to participating.

**Stimuli:**

The MRC linguistics database (Coltheart, 1981) was used to generate a word list of 350 monosyllabic English nouns. The list was broken into 14 groups of 25 words. The list was then reviewed by the authors to remove homophones or words that were not
considered to be colloquially common. SUBTLEXus word-frequency (word count per million words) measures (Brysbaert & New, 2009) were collected for all words and used for subsequent word-frequency analysis.

The design was counterbalanced across subjects in a repeated-measures 2x6 factorial design. The within-subject conditions were AO and AV presentations that were collected across 7 SNR levels (-10, -6, -2, 2, 6, 10 and 14 dB). Pink noise was used to mask the audio files. Noise levels were adjusted to the peak RMS value of the clip audio track to achieve the seven SNR levels (-10, -6, -2, 2, 6, 10 and 14 dB). The SNR levels were chosen to span the spectrum of intelligibility in the AO domain. Each of the 14 lists of 25 words was assigned to a single condition and SNR level for a given subject. There were 14 counterbalancing conditions in total. Counterbalancing of the stimuli was done to minimize word effects - effects due to the different susceptibility/robustness to noise or the presentation condition. Over the course of the experiment each subject was presented all words across the SNR levels and presentation conditions. Each word was presented only once and assignment to SNR and presentation conditions was counterbalanced across subjects. During the experiment the presentations were fully randomized online to minimize order effects.

A 20-year-old female speaker was recorded uttering the 350 monosyllabic nouns that comprised the test set plus an additional 30 words that were used during practice trials. The clips were framed to include the speaker’s full face and tops of her shoulders. Recording were done using a high-definition camera (SONY full HD camcorder) and a small condenser shot gun microphone connected to a R-05 Roland wave/MP3 recorder.
The clips were edited in Final Cut Pro (v. 7.0) on a Mac (OS X v. 10.6.3). The clips were separated into movie (AVI) and audio components (WAV). The AVI files were trimmed to a 1024 x 768 px resolution with a frame rate of 29.97 frames/sec. All audio files were normalized to a target RMS value of 0.5 using custom written MatLab software.

*Presentation:*

Subjects viewed stimuli in an audiometric sound booth (ECKEL noise control technologies, C-Series model C-17 Mod) on a computer monitor (ViewSonic) with a resolution of 1920 x 1080 px. Audio was presented in stereo via a speaker system (Paradigm Electronics); volume was controlled via a Midimate 602 audiometer.

Subjects were seated comfortably approximately 0.75 m from the presentation monitor. They were permitted to adjust their position to remain comfortable. Stimuli were presented via DMDX software (version 4.1.1.0). Subjects controlled the presentation rate via a button press after they delivered their responses; there was no time limit for the length of each trial. They delivered their responses by repeating the word they heard out loud into a microphone. The experimenter sat outside the sound booth listening and watching the subject through the sound booth window recording all responses. The experimenter did not have access to the correct answer to avoid biasing responses. Pluralized and conjugated responses were recorded as such and subsequently marked as incorrect trials.
Analysis:

The analysis was completed using custom written software using Microsoft Excel (2007, SP3 MSO) and MatLab (v. 7.5.0.342 - R2007b) and additional statistical analysis was performed using SPSS Statistics version 21.

Results:

The results for all subjects plus the averaged results across subjects are illustrated in the top panel of Figure 1. Analysis was conducted with a two-way repeated-measures ANOVA (two within-subject effects: modality by six SNR levels), there was a main effect due to SNR \[F(4.223, 291.404) = 1205.483, p<0.001\] – Greenhouse Geisser correction for violation of sphericity] in that increasing SNR led to increased performance in the AO and AV conditions. This is consistent with SNR being a satisfactory manipulation of intelligibility. There was also a significant effect due to modality (AO vs. AV) \[F(1, 69) = 4087.609, p<0.001\] in that the performances in the AO and AV conditions were significantly different. There was also a significant interaction between modality and SNR \[F(5.379, 371.147) = 355.922, p < 0.001\] this interaction could have been due to ceiling effects in the AV modality that capped performance at higher SNRs, decreasing the slope of the AV curve.

The gain in performance from the unimodal to bimodal condition was calculated in two ways: ‘Difference Score’ (DS) (Ross, Saint-Amour, Leavitt, Javitt, et al., 2007) and the ‘Percent Gain’ (%Gain) (Meredith & Stein, 1986). The bottom panel of Figure 1 illustrates the DS, where DS = (AV-AO) as well as the PG where PG = [(AV-AO)/AV].
The error bars are the associated confidence intervals (CIs) as measured by bootstrapping (N=70 with repeats, 10,000 repeats, inner 95% range). The peak in the curve of the difference score at -6 dB corresponds to where the AO curve begins to exhibit floor effects due to the performance asymptote.
Figure 1: All collected data. Error bars represent the 95% confidence interval for the mean as measured via bootstrapping. **Top panel:** Average performance curves are in bold and individual subjects are illustrated in fine lines. Both functions are consistent with smooth sigmoid functions with performance asymptotes at 0 and 1. The AV function hits the ceiling at +10 dB. **Bottom panel:** the performance difference (AV-AO) is illustrated in fuchsia. The shape of the difference exhibits a peak at the -6 dB due to floor effects on the AO function. The percent gain (AV-AO)/AV is illustrated in purple; the shape of the curve is monotonically decreasing.
**Word-frequency**

The effect of word-frequency (WF) on performance was examined to determine if there was a differential effect on the AV vs. the AO modality. If word-frequency exerted a stronger effect on AV stimuli it might be a source of the high variability in P(AV). This was accomplished by running a between-stimulus effects two-way repeated-measures ANCOVA (analysis of covariance) with the AO and AV performances (for each word at each SNR averaged across all subjects) as the two repeated-measures and WF as the covariate. This analysis showed a significant effect of WF \([F(1,348) = 9.691; p = 0.002]\), a non-significant interaction between WF and Modality \([F(1,348) = 0.113; p=0.737]\) and a non-significant interaction between WF and SNR \([F(4.189, 1457.663) = 1.832; p=0.117]\). These results indicate that the WF covariate significantly modulated performance but it exerted the same influence on AO as it did on AV across SNRs. Therefore, the variability in AV was not due to WF interactions with the integration process.

Two follow up regression analyses were then performed to determine the relationships between WF and the average AO performance (averaged across SNRs), as well as between WF and the average AV performance. Both regressions resulted in positive beta values indicating that the main effect exerted by WF on AO and AV performances was a positive correlation.

These results indicated that the WF covariate significantly modulated performance, but it exerted the same influence on AO as it did on AV across SNRs.
Although WF exerted a modulatory effect on performance, this effect was not different between the AV and the AO condition.

**Correlations**

Within-subject correlations between the performances in each modality across SNRs were examined using the R\(^2\) value for each subject. The average correlation between the performances in each modality across SNRs was R\(^2\) = 0.902 (p = 0.001) indicating a strong link between performances in AO and AV for each subject (Figure 2).

Between-subject correlations were examined by comparing performance on AO trials to performance on AV trials using a full factorial multivariate linear regression. The AO values were entered as the predictor variables and the AV values were the dependent variables. The overall fit of the model was non-significant according to Pillai’s trace test [F(49,434) = 1.272; p =0.111]; indicating that the AO performance was not significantly correlated with AV performance.
Figure 2: The $P(AV)$ vs. $P(AO)$ averaged across subjects. Each point represents an SNR level. The correlation of the averaged data is $R^2 = 0.902$ ($p = 0.001$), representative of the high within-subject correlations between $P(AO)$ and $P(AV)$.

Variability in AV Performance

It is customary to report the standard error, however because the AV function hits a ceiling at higher SNRs and the AO function hits a floor at low SNRs, a non-parametric bootstrapping analysis was selected instead. By using this analysis method, the error bars were not bound to being symmetric about the mean. From the 70 subjects, a random subsample from the 70 subjects (with repeats) were chosen and averaged. This process was repeated 10,000 times and the 95% confidence interval (CI) of the mean was calculated (error bars in Figure 1).
The width of the CIs for AO and AV are illustrated in Figure 3 and the dashed lines connect points that are susceptible to floor/ceiling effects (where the saturation was >15%). The most dependable parts of the curves – those whose variance has not been depressed by floor/ceiling effects - are the points connected by solid lines. The width of the CI monotonically decreases with increasing SNR in the AV condition but not the AO condition. At low SNRs (-10 and -6 dB) the variability in the AV performance was greater than for any performance level in the AO condition.

Figure 3: The width of the 95% confidence interval (AO in blue, AV in red) is related to SNR in the AV condition, but not in the AO condition. Dashed lines between points indicate measures that contained more than 15% saturation (Performance at 0 or 1). Performance saturation will lead to depression of the variability and therefore must be interpreted cautiously. The low CI for AO at -10 dB is due to floor effects. The levelling of the CI for high AV values is due to ceiling effects.
Region of Maximal Integration

The results were examined to identify points of inflection that would be consistent with the zone of maximal integration or "integration sweet-spot" that was first proposed by (Ross, Saint-Amour, Leavitt, Javitt, et al., 2007). A point of inflection on a curve indicates a change in the rate of change of performance with respect to SNR the same way changes in acceleration indicate changes in the rate of change of position (i.e. velocity). These inflection points can be identified through local maxima/minima in the first derivative of the performance function that can be estimated using Equation 1. The results for the current study are illustrated in the left panel of Figure 4. The resulting function is monotonically decreasing and therefore has no maxima/minima. Equation 1 was also applied to the results reported by Ross, Saint-Amour, Leavitt, Javitt, et al. (2007) and are presented for comparison in the right panel of Figure 4.

\[
\text{Performance Differential} = \frac{AV_{n+1} - AV_n}{SNR_{n+1} - SNR_n}
\]

Equation 1: To estimate the first derivative of the performance curve - this equation calculates the slope of the line between each point on the performance curve. \(AV_n\) is the performance at the \(n^{th}\) SNR level. This equation measures the rate of change in performance at each two-point interval.
Figure 4: Estimates of the first derivatives of AV performance (calculated with Equation 1) for average AV data in the current study (left panel) and from Ross, Saint-Amour, Leavitt, Javitt, et al. (2007) (right panel). The presence of a local maximum (indicated with arrows) indicates points of inflection where the rate of change of performance peaks. The largest maximum is associated with the reported 'sweet spot' of integration. The smaller maximum is likely due to ceiling effects that depressed the performance at +14 dB. No such local maxima were observed in the current study.

**Variability in the Literature**

The results from 8 studies that measured performance in an AV speech-in-noise task were collected and are summarized in the top panel of Figure 5. When these results are superimposed, they illustrate a significant amount of cross-study variability. It is evident that the SNR measures are not super-imposable as the spectrums of intelligibility vary greatly across the abscissa. We normalized the SNR scale to performance in the AO condition (bottom panel of Figure 5) to determine if this would eliminate the variability. The normalized SNR scale locked AO performances of 0.26 and 0.7 to the same abscissa locations. The AV curves were shifted accordingly. This treatment did not eliminate the
variability in the AV condition. This observation indicates a source of variance that cannot be accounted for with respect to AO performance. Consistent results are dependent on a reliable mapping from the AO to the AV percept. Assuming the groups of subjects tested for each of these studies were approximately equivalent, these results could indicate large speaker effects, such as differentials in visual stimuli. Or alternatively, they could reflect a process that is differentially effective/recruited based on different study presentations and methodologies.

The SNR normalization process was further developed in Figure 6 where AV performance is plotted against AO performance. This plotting method removes SNR entirely and highlights the non-linear relationship between AO and AV. Figure 6 confirms a high degree of variability of the relationship between AO and AV performance.
Figure 5: Illustration of the variability in reported performances on speech-in-noise tasks from 8 studies of single words presented in noise – thick lines represent the results from the current study. AV performance is in red; AO performance is in blue. The results from the current study are bolded. **Top Panel:** The reported variability due to different SNR scales is great. **Bottom Panel:** Data from the same studies on a normalized SNR scale such that the AO performance at two points (0.26 and 0.70) is locked. Variability in the AV performances is reduced, but still remains large. The variability across studies is due to factors outside of AO performance and differing SNR measurements.
Figure 6: Combined analysis for 8 studies measuring performance in a speech-in-noise task for single words. The results from the current study are in bold (red). Due to a high degree of variability in SNR measures, it is more meaningful to present AV data as a function of AO performance. However, variability is not fully explained by SNR effects. Inter-study variability due to methodological differences and presentation effects remain substantial.
Discussion

The results from our study were consistent with most previous studies that measured the performance functions for AV and AO perception of words in noise (Erber, 1969, 1971b; Ewertsen & Nielsen, 1971; Ma et al., 2009; O'Neill, 1954; Sumby & Pollack, 1954). The performance curves were monotonically increasing with SNR, exhibiting smooth sigmoidal behaviour. High within-subject correlations indicated a tight relationship between AO and AV for each subject (Figure 2). This suggests that the integration process that produces enhanced performance in AV is recruited in a consistent way for a given subject across SNRs.

Conversely, the between-subject correlation between average AO and AV was non-significant. This disconnect indicates that the integration process could be considerably different across people. There are two candidate explanations for this. First, the variability in AV is greater than the range of AO at low SNRs, washing out the correlation. Or second, that pre-processing of the auditory signal prior to integration introduces another source of variance that is not explained by unimodal performance (Calvert et al., 1997; Calvert & Campbell, 2003; Mottonen et al., 2004; Paulesu et al., 2003; Pekkola et al., 2005; Sams et al., 1991). If this were the case, the auditory signal that was used to evaluate the AV percept would be different from the auditory signal in the AO percept. Meaning that the underlying assumption that AO maps to AV simply through the addition of an integration process would be incomplete. Either way, the high degree of variability has substantial implications for the resolution of conclusions that can be drawn from AV studies. That is to say, we cannot assume that all subjects have

51
approximately the same integration mechanism and because of this, experimental manipulations that modify performance enhancement will be prone to sampling errors.

High subject variability at low signal-to-noise-ratios has been reported in numerous studies, but none have been able to fully account for the observed variance (Blamey et al., 1989; Erber, 1969; Grant et al., 1998; MacLeod & Summerfield, 1987; Middelweerd & Plomp, 1987).

The high degree of between-subject variability is also reflected in the CI widths that decrease with increasing SNR for the AV condition but not the AO condition. If we make the assumption that the underlying integration mechanism is structurally the same across people, then we assume that the differences in performance enhancement are due to differential efficacy with which the mechanism is recruited. At lower SNRs the increased variance reflects the different degrees to which the integration mechanism may be operating. Therefore, the variable AV performance at lower SNRs is consistent with an integration mechanism that is preferentially activated or recruited at lower SNRs – though this is not always reflected in all of the subjects, it is observed in more subjects than when the integration mechanism doesn’t need to work as hard (higher SNRs). Comparatively the variability in the AO condition is consistent with static additive noise for a process that is not SNR dependent.

The between-subject variability in AV decreases to below the AO variability baseline at higher SNRs (-2dB). There are two implications to this finding. First, that the quality of the AV percept is variable at low SNRs – that is to say, some people are good at the task, while some people are not. And second, at higher SNRs the AV percepts are more
consistent (lower variability) than the unimodal AO percept. This could indicate that a process akin to a Maximum-Likelihood-Estimation procedure is combining the auditory and visual unimodal cues such that the variance is being reduced in the bimodal percept.

Word-frequency had a significant effect on AO and AV performance in that higher frequency words were associated with greater average performance in both the AO and AV conditions. Since the interaction between the WF and the modality condition was non-significant, we can conclude that WF does not differentially impact performance in the two presentation modalities. Therefore, the WF effects did not cause the high degree of variability in performance in the AV condition. It must be noted however, that when the study was designed, beyond stabilizing the average and variance of the frequency values of the word lists, WF was not tightly controlled for. Nevertheless, a histogram of the WFs of the word list produced an exponential distribution that is similarly observed in the SUBTLEXus corpus - this distribution is the reason the metric log(WF) is sometimes used (Brysbaert & New, 2009). This finding supports the word list as being a representative sub-sample of the corpus of monosyllabic nouns.

The cross-study comparison indicated that there are sources of variance outside of the integration mechanism, such as speaker effects and subject differences, which are associated with the experimental methodology. Different studies report substantially different results for similar tasks - indicating problems with the reproducibility (Erber, 1969, 1971b; Ma et al., 2009; O'Neill, 1954; Ross, Saint-Amour, Leavitt, Javitt, et al., 2007; Sumby & Pollack, 1954). If the groups of subjects tested are assumed to be approximately equivalent (a very optimistic assumption), then the variability could be
due to speaker effects, different noise maskers, word effects or whether the task is open or closed-set. It has already been shown that performance on a speech-in-noise task will vary with different speakers (Grant & Braida, 1991). Additionally, different studies have used different word-types such as monosyllabic (Ma et al., 2009; Ross, Saint-Amour, Leavitt, Javitt, et al., 2007), generally bisyllabic (Sumby & Pollack, 1954), specifically spondees (Erber, 1969, 1971a), trochees (Erber, 1971a) or unreported (Ewertsen & Nielsen, 1971; O'Neill, 1954). Performance might be modulated by word type, making the comparison of the AV performances tenuous. Finally, the size of the test set has been shown to exert a significant effect on performance (Sumby & Pollack, 1954). Therefore - as previously discussed by Ross, Saint-Amour, Leavitt, Javitt, et al. (2007) - it might be problematic to compare results from open-set tasks (Ewertsen & Nielsen, 1971; Ma et al., 2009; Ross, Saint-Amour, Leavitt, Javitt, et al., 2007) to results from closed-set tasks (Erber, 1969, 1971b; O'Neill, 1954; Sumby & Pollack, 1954).

Differences in methodologies and presentation effects, such as speakers and stimuli, could influence performance by impacting either the efficacy or recruitment of the integration mechanism. This could be affecting the process by either modulating how information is combined, or the sensory interactions prior to integration. This review illustrates the importance of developing a more standardized model to examine speech-in-noise.

No intermediate zone of maximal integration was observed in this study and our results are therefore inconsistent with the findings of Ross et al. (2007). This was verified with an analysis of points of inflections in the AV performance data. The absence of any
points of inflection in the plot of the first derivative of the performance function (left panel of Figure 4) indicates that there is no integration 'sweet-spot' in our data. Conversely, there are two clear local maxima in the data from Ross et al. (2007) (indicated with arrows in the right panel of Figure 4). The large first maximum indicates the region of maximal integration that was their main finding. The second maximum is likely due to ceiling effects in the AV condition. It is possible that our sampling width between SNR levels was too wide and we missed a peak, though our SNR range spans the same AO intelligibility range as Ross, Saint-Amour, Leavitt, Javitt, et al. (2007) and has the same number of equally spaced SNR sampling levels. A possible difference between our methodologies is that the stimuli in the current study were counterbalanced across all SNRs and the two main conditions such that all words were presented at all SNRs and in both conditions. This was done to minimize word effects (some words are easier to speechread than others while some words are more robust to noise than others) and could have influenced the results. Ross, Saint-Amour, Leavitt, Javitt, et al. (2007) and Ross, Saint-Amour, Leavitt, Molholm, et al. (2007) did not report counterbalancing their presentations. Word effects could have caused the 'integration sweet spot', in that more words in the list for the SNR level of -12 dB in the AV condition, by chance, were easier to speechread, leading to a disproportionately high AV score. The impact of counterbalancing is illustrated in Figure 7 where results are shown for 3 counterbalancing conditions only (15 subjects). The results suggest an enhancement in performance at – 6 dB, but when performance is averaged across all counterbalancing conditions, this peak
diminishes. This result highlights the importance of counterbalancing and recruiting enough subjects to get an accurate representation of the main performance.

![Performance vs SNR graph]

Figure 7: Performance on 3 of the 14 counterbalancing conditions. In this subset the AV performance appears to be enhanced at – 6 dB, however this effect is not present when all the counterbalancing conditions are averaged together.

Our results indicate that the integration mechanism is consistently recruited within-subjects as indicated by the high within-subject correlation, but variably recruited across subjects as indicated by the high cross-subject variability. Further, our results indicate that cross study variability is substantial. These results speak to the inherent volatility of the AV integration process and put into perspective the importance of strict controlling of experimental design when examining this process.
Chapter 3: Measuring Performance Gain in an Audiovisual Speech-in-Noise task

Keywords: Audiovisual, Speech, words, Super-additivity, Inverse Effectiveness, Gain metric

Abstract:

Performance for audiovisual speech-in-noise tasks can be analyzed in several ways using different metrics designed to measure performance enhancement relative to the unimodal conditions. Recent investigations into the validity of the Principal of Inverse Effectiveness (PoIE) have seen these metrics used interchangeably. However, as was previously reported by Ross, Saint-Amour, Leavitt, Javitt, et al. (2007), the different metrics produce distinct results, some of which are consistent with the PoIE and some of which are not. Therefore, these gain measures are describing different things. In this experiment we measured performance on an identification task for monosyllable nouns in noise, presented aurally, visually and audiovisually. Using these data we measured the broader characteristics of six ways to measure performance enhancement in audiovisual speech-in-noise tasks. We propose modifications to traditionally used metrics to ensure that they explicitly describe performance boosts associated with integration. The previously described “Integration Enhancement” metric (Blamey et al., 1989) is developed to describe the magnitude of the audiovisual performance that is unaccounted-for by a linear sum of the auditory-only and visual-only performances in an open-set word identification task.
Introduction:

Performance on an audiovisual (AV) speech-in-noise task depends on three sources of information: information from the auditory stream, information from the visual stream and intermodal information. Intermodal information comprises integration-specific data that is unavailable during unimodal presentations. It is the focus of much research in AV speech perception because it leads to super-additive performance - performance that is greater than the sum of the two unimodal inputs. Being able to measure the degree of super-additivity in AV speech is of critical importance because it will allow investigators to examine in greater detail the integration process. Specifically, finding correlates of super-additivity would provide deeper insight into how the brain chooses to bind the auditory and visual signals and how efficiently these channels are combined. One parameter that affects this process is the manipulation of the signal-to-noise ratio (SNR), which degrades the auditory information. This manipulation also impacts the performance enhancement that is generally exhibited more strongly at lower SNRs – this phenomenon is known as “inverse effectiveness”.

There is currently no consensus on a standardized gain metric that describes the degree of performance enhancement associated with integration. For this reason, this paper outlines the meanings of the metrics described in the literature and strives to identify a metric that is most representative of the super-additive performance on bimodal speech perception tasks.
**Gain Metrics:**

AV performance can be estimated as a linear composite of the unimodal performances: auditory-only (AO), visual-only (VO) and an integration specific term that describes the degree of super-additivity (Equation 2).

\[ AV = AO + VO + Int \]

Equation 2: A general equation of audiovisual performance. The multisensory performance (AV) is the sum of the unimodal inputs (AO and VO) plus the performance gain associated with integration specific processes (Int).

The term in Equation 2 that we are most interested in is the integration term (Int) – because we would like to use a gain metric that measures performance gain associated with the integration process specifically.

There are several metrics that have been used to quantify the boost in performance from the unimodal speech condition to the AV condition. These metrics can be grouped into three main categories, with each category being described by one equation. In the first category are metrics that measure performance enhancement associated with the addition of concurrent visual information. These metrics measure the gain estimated by (VO + Int). The second category consists of the metric that estimates the performance enhancement due to the addition of auditory information. These metrics measure gain estimated with (AO + Int). The third and last category of gain metric estimates the performance enhancement associated with the bimodal integration process that is outside of the unimodal auditory and visual processing streams. These metrics
measure gain only associated with integration specific processing – described by the term ‘Int’.

The equations used to describe each of the three metrics are analogous. That is to say that there are three general mathematical ways to calculate a gain metric. The first equation is the difference score (Equation 4), which is the magnitude of the difference between the AV condition and the starting condition (referred to as “X” in Equation 4). The starting condition “X” can be the VO, AO or AVpredicted (AV_pred) condition, where the AV_pred is a modeled estimate of AV performance (Equation 3). In this paper, AV_pred will always be the linear estimate of the bimodal performance (Equation 3). The AV_pred is not expected to be equal to the observed audio-visual performance. The question is rather whether AV can be traced back to a consistent transformation from AV_pred. The second equation is the Percent Gain (%Gain) (Equation 5); this describes the percent of the observed performance due to a single or two conditions. For example, the standard %Gain definition is (AV-AO)/AV (Meredith & Stein, 1986). This substitutes AO into Equation 5 and measures the proportion of the AV performance that is not due to AO. That is, the performance gain that is due to VO plus integration-specific processes (VO + Int). The third equation is the Normalized Enhancement (NE in Equation 6), which describes the amount the performance increased from the initial condition to the bimodal condition relative to the maximum it could have increased. Because the starting condition X can be one of VO, AO and AV_pred, there are therefore nine total equations describing the gain metric. Six of these nine possible equations have been previously described in the literature.
The difference score for visual gain has been previously described as the difference between performance in the AV and AO conditions (Ross, Saint-Amour, Leavitt, Javitt, et al., 2007) (Equation 4 with $X = AO$). This description gives an estimate of the performance boost associated with $(VO + Int)$. The Integration Enhancement (IE) score described by Blamey et al. (1989) is another version of the difference score, and it describes the magnitude of the difference between the observed AV performance and the linear estimate for a closed-set of stimuli; it is similar to $AV_{pred}$ in Equation 3. Therefore, IE, which is Equation 4 with $X = AV_{pred}$, is an estimate of the magnitude of the performance gain associated specifically with the bimodal integration process. Though it must be emphasized that their calculation of $AV_{pred}$ was different than the one used in this paper.

A previously described definition of the $\%$Gain (Meredith & Stein, 1986; Ross, Saint-Amour, Leavitt, Javitt, et al., 2007), as applied to behavioural data for speech-in-noise perception, refers to the percentage of the performance that cannot be accounted for by the AO performance (Equation 6 with $X = AO$). It therefore describes the proportion of the observed performance attributable to the $(VO + Int)$.

The visual-enhancement score has been previously described (Grant & Walden, 1996; Sumby & Pollack, 1954) (Equation 6 with $X = AO$) and refers to the proportion of the performance gain from the auditory channel to the AV condition relative to the maximum possible performance gain – this is the gain associated with $(VO + Int)$ normalized by maximum possible gain. Similarly, the auditory enhancement (Rabinowitz, Eddington, Delhorne, & Cuneo, 1992) (Equation 6 with $X = VO$) describes
the proportion of the performance gain between VO and AV (AO + Int) relative to the maximum possible performance gain. Finally, the normalized integration enhancement (Tye-Murray et al., 2007a) (Equation 6 with X = AV_{pred}) describes the percentage of the performance gain from the linear estimate for a closed-set task (AV_{pred}) to the actual AV performance relative to the maximum possible performance gain (Int alone).

The metrics associated with performance boosts attributed to elements other than those in the VO condition will be very similar to those in the AV condition. This is because the VO measure has been found to be constant across SNRs (Ross, Saint-Amour, Leavitt, Javitt, et al., 2007). As a result, the equations for which X is equal to VO add very little to the analysis of the AV gain. For this reason, we limit our analysis to the remaining six metrics, which exclude the performance gain due to the auditory channel plus integration (where X =VO).

\[ AV_{pred} = AO + VO - R \]

Equation 3: A linear estimate of AV performance. The R is the Redundancy, a corrective term that is measured experimentally which describes the amount of shared information between AO and VO.

\[ DS = AV - X \]

Equation 4: Difference Score, where X could be VO, AO or AV_{predicted}.

\[ \% \text{Gain} = \frac{AV - X}{AV} \]

Equation 5: Percent Gain, where X could be VO, AO or AV_{predicted}.
\[ NE = \frac{AV - X}{1 - X} \]

Equation 6: Normalized Enhancement, where X could be VO (auditory enhancement), AO (visual enhancement) or AV\(_{predicted}\) (normalized integration enhancement).

**Redundancy:**

The Redundancy (R) term on the right-hand-side of Equation 3 is a corrective term that compensates for overlap between auditory and visual information. Previous work has defined a similar corrective term as AO*VO (Blamey et al., 1989; Ronan et al., 2004; Tye-Murray et al., 2007a). AO*VO is a good approximation for closed-set tasks where the auditory and visual channels are estimated to be independent and there is a good chance that a correct response will be given based on the chance of the co-occurrence of a correct guess in the auditory and visual channels (false positive). To compensate for such false positives, the probability of a guess co-occurring in the auditory and visual channels that gives a correct response (AO*VO) is subtracted from the linear estimate (AO+VO) of the performance. This means that the leftover performance will truly be due to the correct perception of the stimuli. However, when the task is open-set this likelihood becomes negligible. This is because it is highly unlikely that a subject will correctly guess a word or phrase in both the auditory and visual channels. Rather, what is more likely is that the two channels contain redundant information which, if counted twice, could lead to estimated bimodal performances that are greater than 1. This redundancy is likely beneficial to AV performance (Barlow, 2001) and therefore should be eliminated to compensate for performance boosts.
associated with redundancy. Additionally, the R term prevents the $AV_{\text{pred}}$ estimate from exceeding 1, making it more easily comparable to the AV performance. This magnitude of the redundancy is greatest at higher SNRs where words that are correctly speechread in the VO condition are also likely to be correctly understood in the AO condition. At low SNRs the R will be very small because the auditory and visual streams are largely independent. In this paper we outline an experimental procedure to estimate the R. Conceptually, the need for such a term is illustrated in Figure 8.

Figure 8: If the two sets (AO and VO) have an area of overlap (R) then the total area is calculated as $(AO + VO - R)$. This is because the simple addition $(AO + VO)$ will lead to the overlap region (R) being counted twice as it occurs in both the AO and VO conditions.
Modulation of Performance Gain:

The super-additive enhancement in performance in AV speech perception, as compared to AO or VO perceptions of speech, has traditionally been described as being “inversely effective” on the basis that the greatest performance enhancements occur when unimodal performance is most degraded. That is to say, the response enhancement associated with the bimodal signal will be greatest when the input signals are weakest. This concept was adapted from neurophysiological studies which found that super-additive spiking rates associated with bimodal as compared to unimodal stimuli were greatest when the unimodal spiking rates were lowest (Meredith & Stein, 1986). Subsequent research has revealed that the modulatory effect of the input signals is substantial enough to bring the enhancement down to the additive or even sub-additive level as the input signals get stronger (Stanford & Stein, 2007). This effect has also been described for behavioural tasks (Bolognini et al., 2005; Grant & Seitz, 2000; Stein et al., 1988).

The strict definition of the PoIE has been challenged by several recent studies that have instead proposed that it is either not applicable to AV speech-in-noise (Ross, Saint-Amour, Leavitt, Javitt, et al., 2007; Tye-Murray et al., 2010) or is only sometimes applicable to behavioural data (Holmes, 2007). The interpretation of the PoIE for many studies is complicated by a disconnect between the phenomenon on which the principle was based and the nature of the tasks to which it has been applied. Neural spiking rates are most easily comparable to either an AV detection task (Bolognini et al., 2005; Grant & Seitz, 2000) or localization task (Stein et al., 1988) – wherein a non-speech stimulus
has been found to easier to localize in space than a purely auditory or purely visual stimulus. This kind of task requires a binary decision that can be related to a neural accumulator model (Usher & McClelland, 2001; Vickers, 1970). However an identification task will have a more nuanced processing scheme that involves detection, processing and classification stages. As a result this can no longer be directly compared to a detection or localization task and, as such, the link to the PoIE becomes more tenuous. Performance enhancement may therefore vary with SNR in an identification task, but not necessarily follow a monotonically decreasing function.

In addition to the confusion arising regarding the modulation of AV performance enhancement across SNR levels, the misapplication of gain metrics has also lead to confusion on this issue. This is because the parameters used to define PoIE are not consistently defined across the literature. For example, the normalization terms in the denominator of Equation 6 will compensate for the shrinking space between the performance functions and the asymptote at 1 and, as such, will either minimize or eliminate any inverse effectiveness in the data. Equation 6 is therefore an interesting analysis tool, however it may be inappropriate to judge adherence to the PoIE (Tye-Murray et al., 2010). Meredith and Stein (1986) used the Percent Gain metric (Equation 5) in their research, and it will be shown that this gain metric will always follow the PoIE for behavioural data even when others do not.
Study Expectations:

The difference score is the magnitude of the performance difference between AV and the starting condition (either AO or AV\textsubscript{pred}). The lack of a normalization term means that at high SNRs ceiling effects in the AV curve will depress this value. As well, at low SNRs, floor effects in the AO curve will also depress this value. The difference score curve is the difference between two sigmoids (the psychometric functions for AV and AO) and therefore is expected to be consistent with a Gaussian function across SNRs.

The Percent Gain is the difference score normalized to the AV score. It is a measure of the proportional contribution of a particular condition to the overall observed performance. It has previously been described as Equation 5 with X = AO (Ross, Saint-Amour, Leavitt, Javitt, et al., 2007). This represents the proportion of the performance gain that cannot be attributed to the AO performance. Therefore it is the proportion of the performance gain associated with complementary information in the visual channel as well as integration-specific performance gain. At low SNRs the AO value is at floor, therefore this value will be 1 (where it is equal to AV/AV). At high SNRs, AV will be at ceiling and this value will go to zero. Therefore, for the case where X=AO, this value is always expected to follow the PoIE. To calculate the proportion of the performance gain due to only integration specific information, Equation 5 should be used with X = AV\textsubscript{pred}.

The most popular metric reported in the literature is that of normalized enhancement (Equation 6), which is instantiated as visual enhancement, auditory enhancement or the normalized integration enhancement. The visual enhancement is the proportional performance improvement associated with the presentation of the
concurrent VO channel (Grant & Walden, 1996; Sumby & Pollack, 1954) (Equation 6 with X = AO). This performance improvement includes auditory information as well as bimodal-specific performance gain. This is the proportional amount the performance increased relative to how much it could have gone up from the AO condition. The normalized integration enhancement is the proportional performance improvement associated only with the bimodal-specific performance gain (Equation 6 with X = AV_{pred}).

The purpose of the normalization term (the denominator) in the normalized enhancement equation is to counteract ceiling effects in the AV condition and floor effects in the AO condition. However, this gain metric is not immune to floor and ceiling effects. Essentially, the normalized enhancement metrics attempt to linearize non-linear data, emphasizing the proportional relationship between the starting condition X, and AV, rather than the magnitude of the difference between the two. For example, in the case of visual enhancement (Equation 6 with X = AO), the relationship between AO=0.3 and AV=0.8 is equal to the relationship between AO=0.7 and AV=0.915 (both of these pairs have visual enhancement metrics of 0.71). However the absolute difference between these performance values are 0.5 and 0.215 respectively. Another drawback of this metric is that the denominator value is not meaningful when X=1 as it becomes infinity or undefined.

In the case of visual enhancement, at low SNRs AO ceases to be SNR dependent and is instead bound at approximately 0. Conversely, the AV performance will continue to be SNR dependent, this is likely because AV speech perception is able to use the
auditory information that is insufficient for whole word identification in the AO condition. The gain metric is therefore not representing a difference in integration capacity at this SNR level, because the AO performance is not an accurate measure of the amount auditory information available in the AV condition. Therefore the visual enhancement metric therefore is not immune to floor and ceiling effects and is not expected to follow the PoIE.

In this study we will demonstrate that different gain metrics cannot always be directly compared because they are measuring different things. We will not attempt a linearization procedure because normalization does not eliminate floor and ceiling effects. We will instead develop a gain metrics that describes the proportion or magnitude of the AV performance that is associated specifically with the integration process.

**Methods:**

Participants:

Seventy subjects were recruited from the Queen's University Psych 100 subject pool as well as a paid-participant list (10 men; mean age = 19.8 yrs, std = 0.29). All subjects were native English speakers between 18-24 years of age with self-reported normal hearing and normal or corrected-to-normal vision. All subjects provided informed consent prior to participating.
Stimuli:

The MRC linguistics database was used to generate a word list of 350 monosyllabic English nouns. The list was broken into 14 groups of 25 words. The list was then reviewed by the authors to remove homophones or words that were not considered to be colloquially common. SUBTLEXus word-frequency (word count per million words) measures (Brysbaert & New, 2009) were collected for all words and used for subsequent word-frequency analysis.

The design was counterbalanced across subjects in a repeated-measures 2x6 factorial design to minimize word effects - effects due to the different susceptibility/robustness to noise or the presentation condition. The within-subject conditions were AO and AV presentations that were collected across 6 SNR levels (-10, -6, -2, 2, 6, and 10 dB). The SNR levels were chosen to span the spectrum of intelligibility in the AO domain. Twelve of the lists were assigned to the AO or AV condition at each SNR level. The final two lists were assigned to the VO condition. There were 14 counterbalancing conditions in total. Over the course of the experiment each subject was presented all words across the SNR levels and presentation conditions. Each word was presented only once and assignment to SNR and presentation conditions was counterbalanced across subjects. During the experiment the presentations were fully randomized online to minimize order effects.

A 20-year-old female speaker was recorded uttering the 350 monosyllabic nouns that comprised the test set plus an additional 30 words that were used during practice trials. The clips were framed to include the speaker’s full face and tops of her shoulders.
Recordings were done using a high-definition camera (SONY full HD camcorder) and a small condenser shot gun microphone connected to an MP3 recorder. The clips were edited in Final Cut Pro. The video files were trimmed to a 1024 x 768 px resolution with a frame rate of 29.97 frames/sec. All audio files were normalized to a target RMS value of 0.5 using custom written MatLab software.

*Presentation:*

Subjects viewed stimuli in an audiometric sound booth (ECKEL noise control technologies, C-Series model C-17 Mod) on a computer monitor (ViewSonic) with a resolution of 1920 x 1080 px. Audio was presented in stereo via a speaker system (Paradigm Electronics) whose volume was controlled via an audiometer. Pink noise was used to mask the audio files. Noise levels were adjusted to achieve the six SNR levels (-10, -6, -2, 2, 6 and 10 dB) for AO and AV trials. VO trials were presented in silence.

Subjects were seated comfortably approximately 0.75 m from the presentation monitor. They were permitted to adjust their position to remain comfortable. Stimuli were presented via DMDX software (version 4.1.1.0). Subjects controlled the presentation rate via a button press after they delivered their responses; there was no time limit for the length of each trial. They performed an open-set single word identification task, delivering their responses by repeating the word they heard out loud into the microphone. Subjects were instructed to guess when they were unsure of a word, but were also permitted to respond, “I don’t know”, when they were unable to guess. Presentations of AO, VO and AV trials were randomized on-line. The experimenter sat
outside the sound booth listening and watching the subject through the sound booth window recording all responses. They did not have access to the correct answer to avoid biasing responses. Pluralized and conjugated responses were marked as such and subsequently marked as incorrect.

Analysis:

Analysis was done using custom written software using Microsoft Excel (2007, SP3 MSO) and MatLab (v. 7.5.0.342 - R2007b). Additional statistical analysis was performed using SPSS Statistics version 21.

Results:

Mean performance on the VO, AO and AV speech-in-noise task are illustrated in Figure 9. Performance in AO and AV are SNR dependent and are consistent with a smooth sigmoid function that increases with SNR. Performance in the AV condition was greater than performance in the AO condition at all SNR levels. Group analysis was conducted with a 2x6 repeated-measures ANOVA with the AO vs. AV modality being the first repeated measure, and the 6 SNR levels being the second. There was a significant effect due to modality [F(1,69) = 1966.7; p <0.001], as well as a significant effect due to SNR [F(5, 345) = 1270.7; p<0.001]. There was also a significant interaction between the modality and SNR [F(5, 345) = 70.0; p<0.001]. VO was not included as a covariate in the analysis because a two-way repeated-measures ANCOVA revealed that the VO*modality interaction was significant [F(1,68) = 26.085; p<0.001] meaning that
VO co-varied differently with AO than it did with AV. This is consistent with the VO performance being related to the AV performance, but not the AO performance. Therefore, multivariate regression analyses, using VO as a predictor, were conducted separately for AO and AV.

A repeated-measures ANCOVA analysis was performed and the AV condition was found to be significantly correlated with VO performance \( [F(1,68) = 33.548; \ p<0.001] \), and SNR \( [F(5,64) = 114.306; \ p<0.001] \). The interaction between SNR*VO
was also significant \( [F(5,64) = 3.262; p = 0.011] \). Conversely, AV was not correlated with AO performance according to a Pillai’s trace test from a full-factorial multivariate regression analysis \( [F(36, 378) = 1.279; p=0.136] \). Performance between the two unimodal conditions was compared using a repeated-measures ANCOVA – using VO as a covariate for the repeated-measures of AO performance – and it was determined that the AO and VO performances were not significantly correlated \( (F(1,68) = 1703.085, p =0.841) \) and the interaction between SNR*VO was also non-significant \( [F(3.699, 251.523) = 2.103; p=0.086; \text{Greenhouse Geisser correction for violation of sphericity}] \).

Analysis:

*Estimating Redundancy:*

The R term from Equation 3 was estimated based on the observed score for items (words) by comparing the rate of correct responses to each word at each SNR in the VO condition and the AO condition. Words that were identified correctly in both conditions would qualify as having redundant information in both information streams and therefore this value was considered the redundancy score. It should be clarified that the use of the term “redundant” in this situation does not necessarily mean *identical* information, but rather information that leads to the same response. For example, the rate of correct responses to the word “jaw” was 0.30 in the VO condition (averaged across all participants presented with this word in the VO condition, \( N = 10 \)). Comparatively, the performance on the word “jaw” in the AO condition varied across SNRs as \([0, 0.25, 0.4, 0.8, 1, 1]\). The R was described as being the lower of the VO and AO performance score.
for a particular word across each SNR. For instance, based on the above scores, the R for the word “jaw” was [0, 0.25, 0.3, 0.3, 0.3, 0.3]. The average R across all words at each SNR is illustrated in Figure 10.

![Figure 10: Estimated Redundancy between the AO and VO channels at each SNR. This value will never exceed VO performance (0.17). Error bars represent the 95% confidence intervals measured with bootstrapping (N=70 with repeats; 10,000 trials)](image)

**Validity of AV predicted:**

The corrected linear estimate of the bimodal performance ($AV_{pred}$) was examined by correlating the observed AV with $AV_{pred}$ and comparing it to the simple linear sum of AO+VO. The correlation between AV and (AO+VO) was found to be significant ($R^2 = 0.295$, $p < 0.001$). Similarly, the correlation between AV and $AV_{pred}$ was also significant ($R^2 = 0.290$, $p < 0.001$). Therefore, AV performance correlates just as well with $AV_{pred}$ as it did with (AO+VO). Because $AV_{pred}$ is a corrected term that is controlled to not go above 1 it can be used as a performance estimate that can be directly compared with AV.
Gain Metrics:

The six gain metrics discussed in the introduction section were applied to the subject performance data from this study and the results are displayed in Figure 11. The gain metric curves are noticeably different from one another. They have different relationships with the SNR and therefore cannot be used interchangeably.

The difference score and IE metrics exhibited peaks at -2 and -6 dB respectively (though the peak of the difference score was not significantly different at points -2 and -6 dB). The visual enhancement and normalized integration enhancement metrics are analogous and are both monotonically increasing. This was expected given the effects of the normalization terms. Finally, the %Gains for (VO + Int) and Integration alone are similarly shaped functions which largely decrease with increasing SNR. The %Gain for (VO + Int) (left panel of Figure 11) is completely consistent with the PoIE because at low SNRs, where the AO performance approaches zero, the value approaches 1. This is not the case for the %Gain of Integration (right panel of Figure 11) that subtracts out the VO performance. As a result, at lower SNRs where AO performance approaches zero, the bimodal-specific processes boost (which include integration) plateaus and will presumably approach VO as the bimodal performance will eventually become equivalent to speechreading (Erber, 1971a).
Figure 11: Different Gain metrics applied to the data from this study. When a metric is normalized to the total possible performance increase (visual enhancement (VE) and normalized integration enhancement (IEnh)) it will not follow the PoIE. Metrics that are not normalized (difference score (DS) and integration enhancement (IE)) as well as %Gain (which is normalized to the total performance) are consistent with the PoIE. **Left Panel**: Gain Metrics that estimate contributions from VO plus Integration. These are calculated using AV and AO values. **Right Panel**: Gain Metrics measuring the deviation of AV from a linear estimate. These metrics estimate the contribution due to the integration mechanism that is outside of the unimodal streams; this includes integration gain modulation, intermodal information and lexical constraints.

**Metrics that Describe the Gain Modulation Associated with Bimodal Speech Perception:**

The most meaningful interpretation of gain distinguishes the amount of the observed bimodal performance associated with each of the three mechanisms of bimodal speech perception: First the visual channel, second the auditory channel and third, the integration process that is only recruited during AV presentations (Grant & Seitz, 1998). In Figure 12 the AV performance is broken down into these three components. The
individual components were calculated as follows: \( \% \text{Gain}_{\text{AO}} = (\text{AO}/\text{AV}) \), \( \% \text{Gain}_{\text{VO}} = (\text{VO-R})/\text{AV} \), and \( \% \text{Gain}_{\text{Integration}} = (\text{AV} - \text{AV}_p)/\text{AV} \). These are all adaptations of Equation 5. The plot in Figure 12 corrects the contribution of visual information (estimated with VO performance) as being SNR dependent because of the R. The performance enhancement due to the bimodal integration process is consistent with a Gaussian curve which peaks at SNR = -6 dB. The approach is couched in the assumption of Equation 2 that assumes that the AO and VO performances are an accurate reflection of the auditory and visual signals in AV perception. That is to say, we are assuming that integration employs additional processes that supplement, but don’t modify, the auditory and visual channels. Given mounting evidence for cortical interactions, this is likely an oversimplification.
Figure 12: Proportion of the performance due to each element in bimodal speech perception. The sum of these elements will always equal 1. The contributions are from auditory information (blue), visual information (green) and bimodal-specific performance gain (pink). The %Gain Integration curve (pink) is the same as the one illustrated in the right panel of Figure 11.

Discussion:

The SNR-dependent performance on the AV speech-in-noise task was consistent with a smooth monotonically increasing sigmoid function (Figure 9). This result was consistent with previous results that measured performance on word stimuli as a function
of the SNR (Erber, 1969, 1971a, 1971b; Ma et al., 2009; O'Neill, 1954; Sumby & Pollack, 1954). The VO performance was found to be 0.17 ± 0.09. This is consistent with previous studies that have found speechreading performance for words to be around 25% in hearing subjects with no training (Dodd, 1977; Ewertsen & Nielsen, 1971; Tye-Murray et al., 2007a). Performance in the AV condition was super-additive, wherein AV was greater than the corrected sum of AO and VO at all SNRs except for +10 dB.

Consistent with previous findings, AV was significantly correlated with VO (Blamey et al., 1989; Ewertsen et al., 1970; MacLeod & Summerfield, 1987, 1990). Between-subject variance in the VO condition accounted for 34% of the variability in the averaged AV condition. The correlation between AO and AV was much lower, with the variance in the averaged AO condition accounting for only 6% of the variance observed in the AV condition. No significant correlation was found between AO and VO performance. The low correlation between AO and AV could have been due to the high inter-subject variability in the AV performance that has been previously reported (Blamey et al., 1989; Erber, 1969; MacLeod & Summerfield, 1990; Middelweerd & Plomp, 1987). The higher AV variability means that the range in AV performance is greater than the range in AO performance that could have weakened the correlation. This could be due to differential item effects in the AO as compared to the VO conditions. The correlation between the AV performance and the linear estimate (AO+VO) and the corrected linear estimate (AO+VO-R) were significant, with both models accounting for 29% of the variance observed in the averaged AV condition. This value was less than the correlation between AV and VO. Therefore, the AO performance did not add to the predictive power of the
estimate of bimodal performance. This decoupling between AO and AV performances could be due to signal modification wherein the visual channel is able to modify the processing of the auditory information prior to integration. There is physiological evidence for this in the form of cortical connectivity that would permit cortical (Falchier et al., 2002; Romanski, Bates, & Goldman-Rakic, 1999; Smiley et al., 2007) and subcortical interactions (Hornickel, Skoe, Nicol, Zecker, & Kraus, 2009; Musacchia, Sams, Skoe, & Kraus, 2007). This is also supported by some functional imaging studies indicating cross-modal activation (Calvert et al., 1997; Calvert & Campbell, 2003; Mottonen et al., 2004; Paulesu et al., 2003; Pekkola et al., 2005; Sams et al., 1991) – though not all studies have observed this (Bernstein et al., 2002).

The term R was defined to describe redundancy in the AO and VO responses at different SNRs (Figure 8 and Figure 10). This metric is a corrective term developed to measure the deviation of the AV performance from the predicted linear sum of AO and VO in an open-set task (Figure 12). Hence, the %Gain was successfully calculated for the three processes involved in integration: visual processing (estimated with VO), auditory processing (estimated with AO) and the bimodal-specific integration process. The integration process is likely a multi-stage process that involves the aforementioned early cortical interactions, lexical constraints (Tye-Murray et al., 2007b), and a neural gain modulating process.

Six different gain metrics are plotted in Figure 11. These results showed that the different treatments of performance data lead to substantially different results. These
metrics are not interchangeable as each metric provides a different interpretation of the data.

Previously, the difference score between the AV and AO has been used as a means of illustrating performance gain from the AO condition (Ross, Saint-Amour, Leavitt, Javitt, et al., 2007). However, because this measure only subtracts out the AO performance, it, in fact, is the magnitude of the performance boost associated with both the VO channel and the integration process. We therefore recommend using the IE metric (Equation 4 where $X = AV_p$) instead of the difference between AV and AO, as the IE measure removes the unimodal contributions from both AO and VO. The IE is therefore a measure of the magnitude of the performance boost associated only with the bimodal-specific integration process (assuming that the auditory and visual signals in AV are processed the same as they are in AO and VO respectively – a common though likely incorrect assumption).

Previous studies have also used the visual enhancement gain metric to measure integration gain (Grant & Walden, 1996; Sumby & Pollack, 1954; Tye-Murray et al., 2007a). For reasons similar to those described above regarding the difference score between AV and AO, the visual enhancement descriptor is also incomplete. Because visual enhancement accounts for performance gain outside of the performance in the AO channel, it describes the performance gain associated with the VO channel as well as integration. We therefore recommend using the metric normalized integration enhancement instead of visual enhancement. The normalized integration enhancement is the normalized performance gain associated with the bimodal-specific integration
process. It must be emphasized that normalized integration enhancement cannot be used
to study the relationship between the degree of AV performance enhancement and the
auditory signal strength because the normalization term will largely remove the SNR
dependence.

Previous studies have used the %Gain_{AO} (Meredith & Stein, 1986) to describe the
amount of the bimodal performance that can be attributed to integration. However, for
the same reasons as outlined above, the %Gain_{AO} is also an incomplete description. To
actually account for performance boosts associated with integration rather than the VO
channel plus integration, the corrected %Gain_{AVp} should instead be used (Equation 5 with
X = AV_p – calculated with Equation 3). This corrected measure of the %Gain gives the
proportion of the performance that is unaccounted for by performance in the AO and VO
conditions. This is therefore the best estimate of the proportion of performance gain
associated with integration.

The results of %Gain_{AVp} and IE are consistent with an integration process that
recruits the visual channel to access auditory information that is insufficient for
identification during AO presentations (Calvert et al., 1997; Mottonen et al., 2004;
Paulesu et al., 2003; Sams et al., 1991; Stevenson et al., 2012; van Wassenhove et al.,
2005). This would explain the persisting SNR dependence of the AV function at SNRs
below 0% performance rate in the AO condition (Erber, 1969, 1971a; Ma et al., 2009).
The shape of this integration curve could be consistent with the results of Ross, Saint-
Amour, Leavitt, Javitt, et al. (2007), who reported a boost in performance at an
intermediate SNR level; they attributed this boost to a zone of maximal integration or
“sweet spot”. That being said, our results are slightly different from those of Ross et al. because we observed no sharp discontinuity of the slope of the AV curve. Rather, the boost was related to floor effects in the AO curve that predicted zero contribution from the auditory stream. However, the low performance in AO does not reflect the actual informational content of the auditory signal. It appears as though the AV performance is building something out of nothing, but in fact there was information in the auditory signal, just not enough for AO identification. It is possible that if the AO curve were instead a measure of auditory information the IE curve would take a very different shape. Another possible reason for the difference in the findings is that the visual information was different in both studies. Ross, Saint-Amour, Leavitt, Javitt, et al. (2007) reported a VO score of 9%, while ours was 17%. The AV performance enhancement, which we previously explained is SNR dependent, could also depend on the visual information. Modeling results indicate that impoverished visual signals will affect the global characteristics of the difference score between AV and AO (Ma et al., 2009).
Figure 13: The magnitude of the three components associated with AV performance. These curves result from taking each point in Figure 12 and multiplying the point by the AV performance at a given SNR. As a result, the sum of the three curves at any SNR will equal the AV. The Contribution of VO decreases as the R between AO and VO increases. The contribution from the bimodal-specific integration process (pink curve) is lowest at higher SNRs, increasing as the SNR is decreased. At -6 dB the contribution begins to fall off as information necessary for whole word identification in the AO channel is critically degraded by noise.

The %Gain curves for AO, VO and AV\textsubscript{p} can be applied to the performance data to give estimates of the magnitude of AV performance at each SNR due to each of the three components: auditory processing, visual processing and Integration (Figure 13). From this, the magnitude of the performance that is unaccounted for by the unimodal streams can be calculated. The shape of the %Gain\textsubscript{Integration} curve (Figure 12) and the magnitude
of the performance attributed to Integration (Figure 13) are different, with peaks at -6 and -2 dB, respectively. This difference is because of the shape of the AV curve and it highlights the importance of making the distinction between the magnitudes of the contributions, as opposed to the proportion of the performance attributable to each condition.
Chapter 4: Lexical Congruency between Unimodal Confusions Accounts for Some (But not All) of the Performance Enhancement in Audiovisual Speech Perception

**Keywords**: Audiovisual, Speech, Integration, Words, Super-additivity, Class-conditional Independence, Lexical Constraints, Neighbourhood Activation Model, PRE labeling model

**Abstract:**

Performance in an audiovisual word-identification task can be greater than the sum of the unimodal performances. We developed a categorical lexical constraints model using response confusions in the auditory and visual-only conditions to predict audiovisual speech perception in noise in an open-set word identification task. For words that were correctly identified in both unimodal confusions, the model was able to account for experimentally observed performance enhancement. However, because the model could not be applied to all words, the averaged predicted audiovisual performance was still less than the observed performance. These results indicate that the unimodal streams may be processed independently when speech signals are strong in both channels. However, when one or both channels contain weak information there remains a source of performance enhancement that is unaccounted for. This is consistent with the existence of a dependent processing mechanism, wherein signal enhancement occurs due to either the interaction between or the convergence of the unimodal streams prior to lexical token assignment, at either the phonemic or sub-phonemic levels. These results highlight a limitation of models of integration that assume class-condition independence, wherein whole unimodal estimates of the solution are combined into the final response.
**Introduction:**

The goal of this study was to build a predictive model for audiovisual speech perception, based on experimentally measured unimodal performance data from the auditory-only (AO) and visual-only (VO) domains, that could be compared to experimentally measured audiovisual (AV) speech perception performance data. Using a word identification task we modeled AV performance based on AO and VO confusion errors, selectively enhancing predicted AV performance when the solution appeared in both unimodal confusions. We used this model to investigate whether such selective enhancement, which reflects a simple lexical constraints rule, could account for super-additive performance in AV speech perception data. This approach allowed us to also address the limitations of the independent model of speech processing, wherein the auditory and visual channels are processed separately, each providing independent estimates of the solution (Braida, 1991; Massaro & Cohen, 2000). If integration occurs at a late stage of processing then AV performance should be associated with the unimodal performances.
The Neighbourhood Activation Model:

The development of our model was based on previous work regarding the Neighbourhood Activation Model (NAM) (Luce & Pisoni, 1998; Mattys et al., 2002; Tye-Murray et al., 2007b). NAM asserts that auditory and visual speech signals each activate sub-sets of the lexicon that are confusable with the target word; these sub-sets are termed the “lexical neighbourhoods”. The auditory neighbourhoods are partially distinct from the visual neighbourhoods because the auditory and visual channels encode complementary as well as redundant, information. For example, the auditory signal is known to carry strong information on voicing and nasality, while signals of place of articulation are weak. Conversely, the visual signal carries robust information on place of articulation for consonants produced in the front of the vocal tract, but weak information regarding voicing and nasality. This distinctiveness leads to small intersections between the neighbourhoods from which the solution is selected. The auditory neighbourhood can be estimated by adding, deleting or substituting phonemes from the target word. This methodology is based on the work of Nusbaum, Pisoni, and Davis (1984) and is available through the Lexical Neighbourhood Database maintained at the Psychology Department at Washington University (http://neighborhoodsearch.wustl.edu). Mattys et al. (2002) developed the measure of visual lexical equivalence based on visual confusability and lexical constraints (Auer & Bernstein, 1997). This principle has since been further developed to include relational word recognition – that is to say, the facilitative effect of the relationship of a word to other words in the lexicon (Feld & Sommers, 2011). Tye-Murray et al. (2007b) combined the approaches for AV words by collecting AO
neighbourhood data, and then generating VO neighbourhoods (analogous to the lexical equivalence class defined by Auer and Bernstein (1997)) from homophenes (the visual equivalent of auditory homophones) that excluded vowels, which usually have poor visibility. They then classified the words into two groups according to the degree of overlap between the two neighbourhoods. When the neighbourhoods overlapped by 6 or more words – they were considered to have a “high intersection density”. When the AO and VO neighbourhoods overlapped by 5 words or fewer they were considered to have a “low intersection density”. They found that AV performance in the high intersection density and low intersection density conditions was significantly different. They went on to posit that a lexical constraints model such as NAM could account for AV performance enhancement and thereby eliminate the theoretical need for a separate mechanism of integration.

We used a similar approach to NAM, but replaced the neighbourhoods based on large lexicon statistics with measured confusions. We also expanded the approach to operate on a baseline estimate of AV performance, based on a summation model. Using the model we selectively enhanced performance based on coincidence in the unimodal confusions, and thereby provided a quantifiable measure that could be compared to experimentally measured AV values. This approach was chosen for two main reasons. First, using the actual confusion responses would control for the individual speaker effects that could deviate the results from the lexical neighbourhood estimates. That is to say, different speaker characteristics can elicit different types of confusions that are unaccounted-for when the neighbourhoods are estimated beforehand. Second, if
integration occurs at the sub-lexical level, the confusability will not necessarily be only equivalent to the manipulation of phonemic or visemic elements of the target word – which have previously been used to estimate neighbourhoods. Fundamental speech features, such as prosody, stop closure durations and formant transitions, may also contribute to the classification process in a way that is not anticipated by certain instantiations of the NAM model, which may not account for confusions outside of the phonemic/visemic variations of the target word. By using the actual responses, this model avoids making any assumptions about the encoding features. There is expected to be a lot of overlap in the results from this approach and that of NAM because they both will be heavily influenced by the lexical similarity. However, this approach has the advantage of also considering speaker idiosyncrasies that may increase the variability of the confusions from a lexical estimate.

For instance, studies examining patients suffering from auditory verbal agnosia suggested that speech processing is mediated through bilateral amodal representations before interfacing with the lexicon (Poeppel, 2001). Additionally, work in the field of language acquisition in infants indicates that the discovery of statistical and prosodic patterns leads to phoneme and word recognition (Jusczyk, 2000; Kuhl, 2004). This could reflect the development of the hierarchical language-processing scheme based on sub-phonemic encoding which precedes phonemic and word encoding. More generally, there is evidence that speech signals, regardless of modality, are encoded via an amodal intermediate speech-element representation that maps to the lexicon. Green (1998) observed that cross-modal speech influences occur at the featural level, such as voice-
onset time, which is more fundamental than phoneme representations. As well, Skipper et al. (2007) reported fMRI evidence consistent with the existence of gestural speech code elements in frontal motor areas.

Because of the converging evidence that speech encoding occurs at the sub-phonemic level, it is likely inaccurate to assume that lexical neighbourhood can be fully estimated by single phonemic/visemic differences alone. The NAM model requires that the experimenter choose the features that make words confusable – though these estimate have been shown to be accurate (Luce & Pisoni, 1998; Mattys et al., 2002) using the current approach will also take into account effects from talker particularities. Although one can expect confusability to occur between pairs of words that differ by only one phoneme, we cannot be certain that this is the only feature that leads two words to be confused. If the incoming sensory information is first assembled into sub-phonemic representations, then confusability could arise from overlap in this sub-phonemic encoding. Because the speech code is unknown, we cannot be certain that single phonemic differences capture all the places where these speech code elements overlap. By using error confusions we sidestep this issue because we make no assumptions about what makes words confusable with each other. The trade-off in using our approach is that our estimates are limited by the data. Insufficient sampling would lead to underestimation of the confusions and therefore inaccuracies in the model.
Methods:

Participants:

Seventy subjects were recruited from the Queen's University Psych 100 subject pool as well as a paid-participant list (10 men; mean age = 19.8 yrs, std = 0.29). No participants had served in Experiment 1. These are the same data presented in Chapter 3. All subjects were native English speakers between 18-24 years of age with self-reported normal hearing and vision (including corrected-to-normal). All subjects provided informed consent prior to participating.

Stimuli:

The MRC linguistics database was used to generate a word list of 350 monosyllabic English nouns. The list was broken into 14 groups of 25 words. The list was then reviewed by the authors to remove homophones or words that were not considered to be colloquially common. SUBTLEXus word-frequency (word count per million words) measures (Brysbaert & New, 2009) were collected for all words and used for subsequent word-frequency analysis.

The design was counterbalanced across subjects in a repeated-measures 2x6 factorial design to minimize word effects - effects due to the different susceptibility/robustness to noise or the presentation condition. The within-subject conditions were AO and AV presentations that were collected across 6 SNR levels (-10, -6, -2, 2, 6, and 10 dB). The SNR levels were chosen to span the spectrum of intelligibility in the AO domain. Twelve of the lists were assigned to the AO or AV condition at each
SNR level. The final two lists were assigned to the VO condition. There were 14 counterbalancing conditions in total. Over the course of the experiment each subject was presented all words across the SNR levels and presentation conditions. Each word was presented once and assignment to SNR and presentation conditions was counterbalanced across subjects. During the experiment the presentations were fully randomized online to minimize order effects.

A 20-year-old female speaker was recorded uttering the 350 monosyllabic nouns that comprised the test set plus an additional 30 words that were used during practice trials. The clips were framed to include the speaker’s full face and tops of her shoulders. Recordings were done using a high-definition camera (SONY full HD camcorder) and a small condenser shot gun microphone connected to an MP3 recorder. The clips were edited in Final Cut Pro. The video files were trimmed to a 1024 x 768 px resolution with a frame rate of 29.97 frames/sec. All audio files were normalized to a target RMS value of 0.5 using custom written MatLab software.

**Presentation:**

Subjects viewed stimuli in an audiometric sound booth (ECKEL noise control technologies, C-Series model C-17 Mod) on a computer monitor (ViewSonic) with a resolution of 1920 x 1080 px. Audio was presented in stereo via a speaker system (Paradigm Electronics) whose volume was controlled via an audiometer. Pink noise was used to mask the audio files. Noise levels were adjusted to achieve the six SNR levels (-10, -6, -2, 2, 6 and 10 dB) for AO and AV trials. VO trials were presented in silence.
Subjects were seated comfortably approximately 0.75 m from the presentation monitor. They were permitted to adjust their position to remain comfortable. Stimuli were presented via DMDX software (version 4.1.1.0). Subjects controlled the presentation rate via a button press after they delivered their responses; there was no time limit for the length of each trial. They performed an open-set single word identification task, delivering their responses by repeating the word they heard out loud into the microphone. Subjects were instructed to guess when they were unsure of a word, but were also permitted to respond, “I don’t know”, when they were unable to guess. Presentations of AO, VO and AV trials were randomized on-line. The experimenter sat outside the sound booth listening and watching the subject through the sound booth window recording all responses. They did not have access to the correct answer to avoid biasing responses. Pluralized and conjugated responses were marked as such and subsequently marked as incorrect.

Analysis:

Analysis was done using custom written software using Microsoft Excel (2007, SP3 MSO) and MatLab (v. 7.5.0.342 - R2007b). Additional statistical analysis was performed using SPSS Statistics version 21.

Confusions Constraints Model:

We developed a model based on the bimodal interpretation of the Neighbourhood Activation Model (NAM) (Mattys et al., 2002; Tye-Murray et al., 2007b). In this study,
the term “confusions” describes the set of responses for a given target word, including
the correct response. This provides an estimate of the set of words that are most strongly
associated with the presented target, as well as how strongly the target is associated with
itself under varying noise conditions. The overlap in responses between the AO and VO
confusions is used to predict performance enhancement or depression in the AV
condition. Performance enhancement was hypothesized to occur when the target word
(the correct response) occurred in both unimodal confusions. If multiple overlapping
solutions occur across the unimodal confusions then these predictions become the
solution set from which the final response is chosen. This process could be compared to a
conjunction search task (Treisman & Sato, 1990), wherein the visual “features” (VO
responses) that are in conjunction with the auditory “features” (AO responses) narrow the
set of candidate solutions to only those which overlap in the AO and VO condition. It
should be emphasized that the use of the word “features” is for the purpose of the
analogy with a conjunction search task, we do not base our model on auditory or speech
features per-se rather they are based on whole word identification responses. The
presented word is the “target”, and incorrect words that are also co-activated in the VO
and AO condition are “distractors”. Distractors are thus words that satisfy the same
unimodal criteria as the target and are most easily confused with the target. Because of
the use of error confusions to impose constraints onto the solution selection process, we
have entitled this model the “Confusion Constraints Model” (CC). This model is
different from the NAM model developed by Tye-Murray et al. (2007b) (which was
based on the works of Luce and Pisoni (1998) and Mattys et al. (2002)) who proposed a
priori lexical neighbourhoods based on a standardized lexicon that were defined by the experimenter. This approach therefore has the advantage of automatically considering the effects of the speaker in generating confusions, as such particulars will not be estimated by a large lexicon based model.

Using the linear sum of the unimodal inputs as a baseline, performance on individual words was selectively enhanced in the model based on the characteristics of the unimodal error confusions. The linear estimate was calculated according to the equation $AV_{\text{LIN}} = (AO + VO - R)$, where $R$ is the experimentally measured “redundancy”, which estimates the redundant signal information in AO and VO at a given SNR; this value prevents $AV_{\text{LIN}}$ from exceeding 1 and is calculated by taking the minimum between the two unimodal performances at a given SNR. For example, if a word is successfully speechread 25% of the time and correctly heard 35% of the time, the $R$ would be 25% and the linear estimate for that single word would be 35%. This means that the linear estimate will always be less than or equal to $(AO+VO)$. For this task the average $R$, across all words, at each SNR level was [0.01, 0.05, 0.11, 0.13, 0.15, 0.16] for the noise levels -10, -6, -2, +2, +6 and, +10 dB. Previous research has used analogous corrective terms that apply to closed-set tasks (Blamey et al., 1989; Tye-Murray et al., 2007a). $R$ in the current study is a generalization of the corrective terms from the closed-set studies to an open-set task. The single measure of VO performance was compared with all AO performances at each SNR level.

The presented words can be grouped into three broad cases (Figure 14). In Case 1 the presented word activates unimodal confusions that overlap over the correct solution
(CC Prediction = Target). This constraint boosts performance, and in this model AV\textsubscript{CC} performance was maximally boosted to 1. In Case 2 the unimodal confusions do not overlap and the predicted performance is estimated to be the sum of AO and VO (no CC enhancement). In Case 3 the unimodal confusions overlap over an incorrect solution (Prediction = Distractor). Therefore the predicted performance is depressed below the linear estimate, and in this model AV\textsubscript{CC} performance was maximally depressed to 0. In this way, the CC model was constructed as a categorical model. In some cases there were multiple solution overlaps that required the system to choose a winner. In these cases the probabilities-of-selection for each word in the AO and VO conditions were multiplied. A winner-take-all operation was performed and the word with the highest product was chosen. Therefore, the predicted outcome of these words therefore subsequently fell into Case 1 or 3. Figure 15 illustrates an example of the competition between the words “push”, “bush” and “porch”. The winner in this case is “bush”, which was indeed the target word and therefore the model predicts performance enhancement (Case 1). Figure 16 illustrates an example of the competition between the words “gross”, “roast” and “rose”. The winner, “gross”, is a distractor solution and the model therefore predicts performance depression (Case 3).
Figure 14: Overlap in the unimodal error confusions can lead to a prediction that is used to constrain the bimodal solution. In Case 1 the AO and VO error confusions overlap over the correct solution (Prediction = Target word) and performance is boosted. In Case 2 there is no overlap between the unimodal confusions and therefore there is no constraint and the system will perform at the rate of the sum of AO and VO. In Case 3 the AO and VO confusions overlap over an incorrect solution (Prediction = Distractor) and the system is biased away from the target word, depressing performance.

Figure 15: The initial AO and VO inputs activate a set of possible solutions during the co-activation stage. Competition between multiple overlapping solutions leads to a correct prediction for the word “bush”. High co-activation in both the AO and VO conditions means that “bush” out-competes “push” and “porch”.

99
Figure 16: Competition between multiple overlapping solutions leads to an incorrect prediction of the word “gross” for the presented word “roast”. High probability of selection in the AO condition and co-activation with the probability of selection in the VO condition leads to maximal activation of a distractor.

An overview of the CC model is provided in Figure 17. A word is presented audio-visually. The auditory and visual channels process this word independently. Each channel operates as it would if the stimuli were presented unimodally, and projects estimates of the word onto the lexicon where the word representations are stored. Co-activation in the two unimodal channels leads to the restriction of the solution set from which the final response is chosen. If there are multiple solutions in the co-activated solution set the solution with the greatest probability (AO*VO) is chosen. The word confusions were measured experimentally (Figure 17 depicts the responses to the presentation of the word “bush”). Performance enhancement was set to the maximum (AV_{CC} = 1) and performance depression was set to the minimum (AV_{CC} = 0). When there was no prediction the linear estimate was used (AV_{CC} = AV_{LIN}).
Figure 17: Overview of the Confusion Constraints (CC) Model. Co-activation of the AO and VO channels is used to constrain the solution set from which the final bimodal response is chosen. When the model predicts the target word performance is maximally enhanced. When the model predicts a distractor the performance is depressed. When the model is not able to make a prediction the performance is estimated to be the sum of the unimodal performances.

Results:

Performance in the unimodal conditions was recorded by stimulus (word performance) rather than by subject. Average performance in the VO condition was 0.168 ± 0.040. Performance in the AO condition ranged from 0.026 (± 0.008) to 0.832 (± 0.024) (lowest to highest SNR condition). Confidence intervals (95%) were calculating using non-parametric bootstrapping (N=350 with repeats; 10,000 trails). The unimodal performances were combined into a linear estimate of the AV performance ($AV_{LIN}$) that is illustrated in Figure 18 (yellow line of the top panel).

Performance in the AV condition was significantly greater than $AV_{LIN}$ [$F(1,347) = 541.423; p<0.001$] and there was a significant interaction between the SNR and AV measures [$F(4.324, 1500.374) = 41.800; p < 0.001$; Greenhouse Geisser correction for violation of sphericity] . Statistical analysis was conducted using a 2x6 mixed ANOVA with Estimated vs. Observed AV performance across the 6 SNRs. The Integration
Enhancement (IE) was calculated using $AV_{\text{LIN}}$. The IE value measures the amount of AV performance that is unaccounted-for by the model and is calculated by subtracting the observed performance from the model prediction: $IE = AV - AV_{\text{predicted}}$. In Figure 18 the model is the linear estimate $AV_{\text{LIN}}$ and therefore the IE value is $IE_{\text{LIN}}$. The $IE_{\text{LIN}}$ is illustrated in the lower panel of Figure 18. Floor effects in the AO condition carry over to $AV_{\text{LIN}}$ and suppressed $IE_{\text{LIN}}$ at -10 dB. Otherwise, the $IE_{\text{LIN}}$ curve is generally decreasing with increasing SNR, indicating that the largest amount of unaccounted performance is at lower SNRs and therefore the integration process is contributing the most to the enhancement at these noisier SNRs.

As long as the IE value is greater than zero there is super-additive performance in AV that is unaccounted for by the model. We can see from the bottom panel of Figure 18 that AV is super-additive at all SNRs and therefore $AV_{\text{LIN}}$ is a poor model of integration that under-predicts performance. As described above, the goal in this study is to determine if a lexical constraints model will have an IE value closer to, or equal to zero.
Figure 18: **Top Panel:** Group results for the observed AV performance (red line) and the linear estimate $AV_{LIN}$ (yellow line). Error bars are the 95% confidence interval for the mean as measured by bootstrapping (N=350 with repeats; 10,000 trials). **Bottom Panel:** Integration Enhancement (IE) is the difference between the observed performance ($AV_{obs}$) and the predicted performance. This figure depicts the linear model of AV performance ($AV_{LIN}$).

The predictions from the Confusion Constraints (CC) model were compared to those of the linear model by examining the IE values (Figure 19). Statistical analysis was conducted using a 2x6 mixed ANOVA with CC vs. LIN estimated AV performance across the 6 SNRs. The $IE_{CC}$ value was significantly different than that for the linear
estimate (IE\textsubscript{LIN}) \[ F(1,349) = 100.536; \text{p} < 0.001 \] and there was a significant interaction between SNR and the model-type \[ F(4.063, 1417.903)=29.456; \text{p}<0.001; \] Greenhouse Geisser correction for violation of sphericity]. However, the CC model did not account for all of the super-additive performance, as the IE\textsubscript{CC} value was greater than zero at all SNRs.

Figure 19: The IE for the Confusion Constraints model (IE\textsubscript{CC}) is lower than for that of the linear model (IE\textsubscript{LIN}). This means that the CC model accounts for some, but not all, of the super-additivity.

To measure the differential effect of performance enhancement for the three word classifications (described in Figure 14) the IE\textsubscript{CC} was plotted for each case and compared across the word classifications (Figure 20). The CC model predicted performance for
words belonging to Case 1 (Prediction = Target) and was not significantly different from the actual AV performance on those same words \[F(1,213) = 2.587; p = 0.109\]. Statistical analysis was conducted using a 2x6 mixed ANOVA with Estimated vs. Observed AV performance across 6 SNRs for the restricted set of words (Case 1 in Figure 14, 60% of the wordlist). A within-stimuli mixed ANOVA was performed (3 Cases X 6 SNRs) which confirmed that significant differences existed between the 3 cases \[F(2,347) = 43.259; p<0.001\]. Post-hoc multiple comparisons, with Bonferroni corrections, indicated that Case 1 was significantly different from Cases 2 and 3 (both p<0.001). While Cases 2 and 3 were not significantly different from each other (p = 0.125). The results described in Figure 20 indicate that for words that are correctly predicted by the model (Case 1), the model under-predicts performance at low SNRs, where the IE\(_{CC}\) is greater than zero, and over-predicts performance at higher SNRs, where the IE\(_{CC}\) is less than zero. The model does not account for the performance enhancement observed for the words that have either no prediction (Case 2, 18% of words) or whose prediction is a distractor (Case 3, 20% of words).
Figure 20: The IE\textsubscript{CC} broken down into the three CC cases (see Figure 14). Showing that super-additivity is accounted-for in Case 1 (60% of the wordlist), where the presented word and the predicted word coincide (Prediction = Target). The IE remains high for words with no Prediction (18% of the wordlist) and for words whose prediction was incorrect (Prediction = Distractor) (22% of the wordlist).

**Discussion:**

When the prediction was the target word, the CC model accounted for super-additive AV performance. However, because the CC model could only apply to 60% of the words, it was not able to account for all of the AV performance enhancement. This is likely because CC is limited by the unimodal signal quality and has no access to intermodal information. In Case 1 (where Prediction = Target), the unimodal signals are
strong, and it is therefore not entirely surprising that the model largely predicts the performance enhancement. It is possible that strong unimodal signals allow for independent channel processing, possibly through a lexical constraints mechanism. But it appears as though either the CC model incompletely describes the process responsible for the performance enhancement, or there is an additional process being employed when the unimodal signals are weak. There are two reasons a lexical constraints model would not capture this. First, because the whole-word lexical tokens could be too removed from the pre-phonemic speech-code elements, they are therefore unable to capture lower level features used to discriminate words during bimodal presentations. That is to say, the scoring method of whole word identification was not sensitive enough to capture the actual informational content in the auditory and visual channels. Or second, because intermodal information is not encoded in a lexical constraints model there is not access to relevant temporally coincident information that might aid in discrimination.

There is substantial behavioural evidence of pre-phonemic level integration in behavioural studies (Fowler, 1986; Green, 1998; Green & Miller, 1985), as well as evidence from functional imaging studies (Skipper et al., 2007) and lexical modeling (Iverson, Bernstein, & Auer, 1998). Physiologically, there is evidence that visual cortex has direct access to auditory cortex at the earliest levels or sensory processing (Calvert et al., 1997; Calvert & Campbell, 2003; Mottonen et al., 2004; Paulesu et al., 2003; Pekkola et al., 2005; Sams et al., 1991). Additionally, visual cortex could have access to auditory cortex via long-range oscillations (Schroeder et al., 2008). These results indicate that interactions between the auditory and visual processing streams can occur well before
lexical word tokens are activated. If integration occurs via the combination of lower level pre-phonemic elements then the true AV solution set – hypothesized in this study to be the intersection between the AO and VO confusions – could be determined not by whole word coincidence, but rather by the intersection of lower level features; features such as formant transitions, place of articulation, voicing and nasality. And even if integration occurs at the phonemic level, a model that only accounts for single phoneme differences might not adequately represent the true AV lexical neighbourhood. The CC model and NAM model are predicated on the assumption that the intersection of the unimodal confusions is equivalent to the intersection of the confusions between the sub-phonemic encodings of the speech signals. Indeed, in our results we observed unexpected AO and AV confusions that did not follow the single phoneme difference rule. In the AO condition, the word “bell” was commonly confused with the word “valve” (12.3% rate of confusion), the word “king” was commonly confused with “pain” (14.3% rate of confusion), and “milk” was confused with “note” (13.6% rate of confusion). These confusions would not have been predicted by a single phoneme difference because they, in fact, have no phonemes in common. Additionally, in the AV condition many words were confused with solutions not observed in either unimodal case. For example, the word “green” was repeatedly confused with the word “wing”, the word “hay” with the word “egg” and the word “ink” with “leak”. These confusions are inconsistent with the hypothesis that AV solution selection is a result of straightforward overlap between AO and VO confusions. Though it should be mentioned that many confusions were visemically close to the target words though they were not measured in the VO
confusions. This could reflect differential processing in the AV condition, or a
shortcoming in the limited sampling of the VO confusion. Another fundamental issue is
that performance in VO is an incomplete measure of the usefulness of the VO
information. Partial VO information has been shown to be enough to boost performance
(Fort et al., 2013). Salient features in the VO channel, which do not transfer over directly
to high VO performance on a whole word identification task, could well be assisting in
bimodal word identification. And on the whole, there is significant evidence that the
NAM estimation of each unimodal stream directly accessing the lexicon is likely an
oversimplification.

Previous studies have found that temporal coincidence of the acoustic envelopes
and amplitudes of visible articulations enhance sensitivity in a speech detection task
(Grant & Seitz, 2000; Vatakis, Maragos, Rodomagoulakis, & Spence, 2012). Statistical
evaluation of speech signals has also indicated that temporal correlations between
auditory and visual signals are rich sources of information (Chandrasekaran, Trubanova,
Stillittano, Caplier, & Ghazanfar, 2009; Yehia, Kuratate, & Vatikiotis-Bateson, 2002),
making them likely sources of information during speech perception. Visual speaking
rate has also been shown to alter the perception of voicing in speech (Green & Miller,
1985), and the perception of the McGurk illusion, which is presumably a measure of
auditory and visual fusion, is dependent on temporal asynchronies (Munhall et al., 1996;
van Wassenhove et al., 2002; van Wassenhove, Grant, & Poeppel, 2007). Additionally,
speech intelligibility in noise has been found to decrease when auditory and visual inputs
are asynchronous (Grant & Greenberg, 2001; Tanaka, Sakamoto, Tsumura, & Suzuki,
Taken together, these results indicate that the intermodal information carried by the temporal correlations between auditory and visual speech signals is a source of information during speech perception. The CC model does not account for such elements because the lexical constraint is applied after independent unimodal processing.

The assertion that super-additivity in bimodal integration is due to lexical constraints only, rather than a separate integration mechanism (Tye-Murray et al., 2007b), could be the case for words for which the unimodal signals are strong. However, the lexical constraints model fails to account for more general performance enhancement. Additionally, there remains the substantial body of evidence that supports the existence of a separate integration mechanism. For example, evidence that variance in AV is unaccounted-for by unimodal performance (Blamey et al., 1989; Erber, 1969; Grant & Seitz, 1998; MacLeod & Summerfield, 1987; Middelweerd & Plomp, 1987) and functional imaging studies that have described brain activity specifically associated with bimodal speech presentations (Calvert et al., 2000; Calvert et al., 2001; Mottonen et al., 2004). This shortcoming in the CC and NAM models indicates that the bimodal speech perception system could be either exploiting intermodal information, integrating sub-phonemic speech elements, or performing a gain operation on the speech signal to boost the signal clarity.

The model correctly anticipated a performance depression when the predicted word was a distractor (Case 3 in Figure 20). This can be inferred because we know that the model depresses performance for Case 3 (Prediction = Distractor) and it does not perform additional operations when there is no prediction (Case 2). The fact that IE_{CC} for
Case 2 and Case 3 are not significantly different from each other implies that the CC model prediction is as different from the true AV performance in both cases. Since we know that the model performance of Case 3 is depressed, we know that the actual AV performance was also depressed. Therefore, lexical confusions seem to be an adequate measure of the magnitude of performance depression, if not the performance enhancement. This shows that performance suppression due to a confusion effect is separable from the integration enhancement.

The CC model could potentially be improved by implementing the constraints at the phonemic or sub-phonemic level. This modification would challenge the fundamental assumptions of the NAM, which keeps the processing of the auditory and visual streams independent up until projection onto the lexicon. The assigning of a lexical token to the signals prior to the integration makes NAM and CC late-stage integration models, which conflicts with research indicating that integration occurs at a pre-phonemic stage (Buchwald, Winters, & Pisoni, 2009; Sams, Manninen, Surakka, Helin, & Kättö, 1998) and that prior to the integration of the auditory and visual information, the two signals are translated into amodal “common currency” that can be directly combined (Kim, Davis, & Krins, 2004; Rosenblum, Miller, & Sanchez, 2007; Walton & Bower, 1993).

Additionally, previous theoretical models of AV speech perception have proposed that the auditory and visual channels are projected onto an amodal space where the signals can be combined into an estimate of the solution. The Fuzzy-Logic-Model-of-Speech (FLMP) (Massaro & Friedman, 1990; Massaro & Stork, 1998) asserts that feature vectors of sub-lexical speech elements are optimally combined to provide likelihood
estimates for a closed-set of candidate solutions. The PRE-labeling model (Braida, 1991) asserts that sub-lexical features constitute the axes of a multi-dimensional space in which lexical elements are coded. Lexical tokens are therefore associated with a particular coordinate value in the feature space. The Motor Theory of Speech Perception (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985) and Analysis-by-Synthesis (Halle & Stevens, 1962; van Wassenhove, 2007) models assert that sub-lexical elements are encoded as intended speech articulations in motor cortex. All of the foregoing models assume a common-currency approach to the speech signal in which auditory and visual information can be combined into a single bimodal speech signal. These models provide options as to how sub-phonemic constraints could be enforced, however the CC approach is most compatible with the PRE model. Because of the high degree of sub-phonemic feature redundancies between confusable words, multi-dimensional scaling would become a necessity. Essentially, the confusion approach illustrated in Figure 16 and Figure 17 would be expanded into multi-dimensional space to include more fundamental features of speech rather than just a single dimension of word candidates. Though it is not definitively known what these features are, they could be speech elements such as voicing, nasality and place of articulation. While the auditory signal would provide information on the first two of these features, visual would provide information on the third. The introduction of concurrent visual information introduces additional dimensions along which solution discrimination can occur. Therefore solutions, which are easily confused in the two auditory dimensions, might be easily discernable in the three AV dimensions. A
substantial complication in the development of such a model is that the fundamental speech elements are not definitively described in the literature.

An appropriate place to start, in light of the speech code being unknown, could be by biting off a manageable problem, so to speak, and constructing such a model for phonemic confusions. This would provide an indication of whether the lexical constraints model supplemented with phonemic constraints would provide an adequate estimate of super-additive AV performance. Indeed this was the approach of Braida (1991). The difference in this case would be to build the predictions to the word level rather than restrict it to syllables. Theoretically, this approach could be expanded to longer utterances through the stimulus set would need to be large to average out differential segmentation effects, lexical facilitation effects and semantic facilitation effects.

In summary, we constructed a lexical constraints model based on unimodal error confusions that was able to account for super-additive bimodal performance in 60% of the stimuli. These results indicate that models of AV integration that rely on independent treatments of the auditory and visual streams are incomplete. The integration enhancement in the remaining 40% of our stimulus set could rely on bimodal integration at the sub-phonemic/phonemic level, a gain modulatory mechanism or intermodal information garnered from temporal correlational analysis between the auditory and visual streams.
Chapter 5: Susceptibility to the McGurk Illusion is Negatively Correlated with Performance in an Audiovisual Speech-in-Noise Task

**Keywords:** McGurk Effect, Audiovisual, Speech-in-noise, speechreading

**Abstract:**

We present results from a behavioural study that critically examines the mechanisms underlying the perception of the McGurk illusion. McGurk susceptibility was measured using five stimuli consisting of VCV nonsense syllable combinations. Its relation to performance on a single word audiovisual (AV) speech-in-noise task was measured using multiple regression analysis. We found that susceptibility to different McGurk stimuli varied significantly between the stimuli. Only one of these McGurk stimuli significantly correlated with AV performance. But perhaps most surprising, was that the correlation between the McGurk susceptibility and the AV performance was negative, meaning that higher susceptibility to McGurk was associated with lower performance in an AV speech-in-noise task. Furthermore, a significant negative correlation was also observed between McGurk susceptibility and speechreading performance, a unimodal performance measure that enlists no bimodal integration. There was no significant relationship between the degree of bimodal integration enhancement and the susceptibility to the McGurk Effect - indicating that the relationship between susceptibility and AV was not driven by the integration process, but rather by speechreading ability. These results are inconsistent with the McGurk illusion as an
appropriate proxy measure for the integration mechanism that underlies AV speech-in-noise perception.

**Introduction:**

Perception of audiovisual (AV) speech-in-noise has been observed to be a super-additive process (Dodd, 1977; Ewertsen & Nielsen, 1971; Ma et al., 2009; Risberg & Lubker, 1978; Ross, Saint-Amour, Leavitt, Javitt, et al., 2007; Tye-Murray et al., 2007a; Wright et al., 2003). When the auditory signal is degraded with noise, performance in the bimodal AV condition is greater than the sum of the auditory-only (AO) and visual-only (VO) performances. The convergence of concurrent auditory and visual speech signals therefore leads to improved performance. The mechanism underlying this integration process is the subject of ongoing research. Here we investigate the relationship between the performance enhancement of AV speech and a frequently used proxy for bimodal integration, the McGurk Effect.

AV performance enhancement can be quantified by calculating how much the observed AV performance deviates from the linear sum of the unimodal performances. The linear estimate, or predicted AV (AV_{pred}), is calculated according to the equation $AV_{pred} = (AO + VO - R)$. Where R is a corrective term that compensates for redundant information in the auditory and visual channels. When AV performance is super-additive it means that AV > AV_{pred}. The amount of super-additivity can therefore be calculated through a measure called the Integration Enhancement (IE) which is the difference between the two: $IE = (AV - AV_{pred})$ (Blamey et al., 1989). This is the gain in AV
performance that cannot be accounted for by the unimodal performances and is therefore explicitly associated with the integration process.

The integration process enlisted during AV speech perception results in improved behavioural performance as compared to the performance in unimodal speech perception tasks. It has been proposed that the performance gain could be due in part to the modification of the auditory information by the visual signal in auditory cortex (Calvert et al., 1997; Calvert & Campbell, 2003; Mottonen et al., 2004), either via direct projections (Falchier et al., 2002; Rockland & Ojima, 2003), or long-range oscillatory effects (Fingelkurts, Fingelkurts, Krause, Möttönen, & Sams, 2003; Lakatos, Chen, O'Connell, Mills, & Schroeder, 2007). It has also been proposed that lexical constraints imposed on incoming sensory information could account for super-additive performance (Tye-Murray et al., 2007b). The most pervasive theory however, is that which describes an integration mechanism receiving projections from the auditory and visual channels and then outputting the bimodal signal (Watson et al., 1996). Evidence from functional imaging studies has found that multisensory convergence associated with speech may be occurring in the superior-temporal-sulcus (STS) (Callan et al., 2003; Murase et al., 2008; Reale et al., 2007; Wright et al., 2003), making the STS the front-running candidate for the site of this integration process.

There is a large body of research devoted to elucidating the underpinnings of the integration process - be it an emergent property of cortical dynamics or a discrete gain modulatory process. A great deal of this research has used the McGurk Effect as a proxy for bimodal integration. The McGurk Effect, discovered by McGurk and MacDonald
(1976), is a perceptual illusion often thought to measure the degree of AV speech integration. In this illusion, concurrent but incongruent AV stimuli are presented to a subject. For example, an auditory /aba/ utterance is played with a visual /aga/ utterance. The subject will often hear an illusory utterance, such as /ada/ or /atha/. Because the illusory response is clearly heard, it was originally described as an example of how the visual signal is able to modify the auditory signal (MacDonald & McGurk, 1978), but the definition has evolved and it is frequently used as a proxy measure of the AV speech integration mechanism (Massaro & Cohen, 1993). In McGurk studies, the frequency of illusory responses – called McGurk susceptibility - is used as a measure of the subject’s binding efficiency (Alsius et al., 2005; Sekiyama et al., 2003; Tomaskovic et al., 2008; van Wassenhove et al., 2002). Experiments that vary parameters that modulate the susceptibility - such as stimulus-onset-asynchrony (Munhall et al., 1996; van Wassenhove et al., 2002) - are therefore assumed to vary parameters that modulate the integration process.

McGurk susceptibility can be measured using different stimuli. The most commonly used stimuli are nonsense syllables, such as an auditory /ga/ combined with a visual /ba/ (McGurk & MacDonald, 1976). But McGurk susceptibility has also been measured for different syllable combinations, such as auditory /ka/ and visual /pa/ (McGurk & MacDonald, 1976; Sams et al., 1998), auditory /ba/ and visual /da/ (Nath, Fava, & Beauchamp, 2011; Soto-Faraco & Alsius, 2009), and auditory /ta/ with visual /pa/ (Kaiser, Hertrich, Ackermann, Mathiak, & Lutzenberger, 2005; MacDonald & McGurk, 1978). It can also be measured in real words (Ali & Ingleby, 2005; Alsius et al.,
There is little discussion regarding whether these different types of McGurk stimuli can be used interchangeably to measure the “integration” mechanism. In the current study, we collected data for five nonsense syllable combinations to measure if their susceptibilities are significantly different from each other and whether they correlated differentially with AV performance. The five combinations could conceivably cover a range of speech features and thus might provide a more complete estimate of AV speech binding.

Because super-additive performance on an AV speech-in-noise task results from the integration of the unimodal inputs, and the McGurk susceptibility is considered a measure of the strength of AV integration, it seems logical that susceptibility to the McGurk will be positively correlated with AV performance across subjects. However, this relationship has not been explicitly tested. In the current study, we present the results of a simple experiment from which we measure susceptibility to the McGurk illusion for five VCV nonsense syllables and correlate this susceptibility with performance on an AV speech-in-noise task for high-frequency monosyllabic words.

**Methods:**

**Participants:**

Thirty subjects were recruited from the Queen's University Psych 100 subject pool as well as a paid-participant list (4 men; mean age = 19.1 yrs, std = 1.3). All subjects were native English speakers between 18-22 years of age with self-reported
normal hearing and vision (including corrected-to-normal). All subjects provided informed consent prior to participating.

**Stimuli:**

The MRC linguistics database was used to generate a word list of 300 monosyllabic English nouns. The list was broken into three groups of 100 words. Word-frequency was controlled across lists using SUBTLEXus word-frequency (word count per million words) measures (Brysbaert & New, 2009). The lists were reviewed by the authors to remove homophones and words that were considered to be colloquially uncommon.

The three sets of word lists were randomly assigned to the AO, VO and AV conditions - to create three counterbalancing conditions. Presentation of the groups of words was counterbalanced across subjects. Each subject was therefore presented the whole word-list, viewing each word once in one of the three conditions. This was done to minimize stimulus effects: effects due to different susceptibility/robustness to noise or the presentation condition. During the experiment the presentation of words and conditions were fully randomized online to minimize order effects.

A 20-year-old female speaker was recorded uttering the 300 monosyllabic nouns that comprised the test set plus an additional 30 words that were used during practice trials. The clips were framed to include the speaker’s full face and tops of her shoulders. Recordings were done using a high-definition camera (SONY full HD camcorder) and a small condenser shot gun microphone connected to an MP3 recorder. The clips were
edited in Final Cut Pro. The video files were trimmed to a 1024 x 768 px resolution with a frame rate of 29.97 frames/sec. All audio files were normalized to a target RMS value of 0.5 using custom written MatLab software.

During a separate filming session, the same speaker was recorded uttering the following 6 VCV nonsense syllables: /aba/, /ada/, /aga/, /apa/, /ata/, and /aka/. Four recordings of each syllable were made. The audio and video tracks were separated to create the following five incongruent AV presentations (audio is listed first, then video): [aba, aga], [aga, aba], [aka, apa], [aba, ada], and [ata, apa]. In subsequent referencing these McGurk stimuli will be referred to by their respective intervocalic consonants (with audio listed first, and then video), for example [aba, aga] = [B, G]. Using the 4 exemplars of each consonant recording, 16 combinations of each incongruent stimulus pairs were made. Multiple stimuli were used to achieve a broader characterization of McGurk susceptibilities for the subjects. The audio tracks for the incongruent trials were modified such that the onset of the auditory burst associated with the middle consonant coincided with the burst from the audio file associated with the video file to which it was being aligned. Once the consonant burst onsets were aligned, the onsets of the first vowels were also aligned. This was accomplished by modifying the length of the stop closure duration (by adding or deleting silence) between the offset of the first vowel and the acoustic burst of the intervocalic consonant. In addition to the incongruent trials, congruent control trials were created for all 6 syllables and included in the final presentation. Each of the 5 incongruent McGurk stimuli was presented twice, so that each stimulus consonant type
was presented 32 times. Additionally, 36 congruent control trials were included (6 of each recorded syllable) in the presentations. There were therefore 192 trials in total.

Presentation:

Subjects viewed stimuli in an audiometric sound booth (ECKEL noise control technologies, C-Series model C-17 Mod) on a computer monitor (ViewSonic) with a resolution of 1920 x 1080 px. Audio was presented in stereo via an HD speaker system (Paradigm Electronics), the volume of which was controlled via an audiometer. Pink noise was used to mask the audio files. A single noise level of -5 dB was used to mask the AO and AV trials while the VO trials were presented in silence.

Subjects were seated comfortably approximately 0.75 m from the presentation monitor. They were permitted to adjust their position to remain comfortable. Responses were delivered via keyboard entry after each trial – the responses were open-set. The subjects were instructed to write whatever they heard. Stimuli were presented via DMDX software (version 4.1.1.0). Subjects controlled the presentation rate via a button press after they delivered their responses; there was no time limit for the length of each trial.

The experiment was conducted in two consecutive sessions; first, the subjects performed the McGurk task. They were told that they would be watching a series of clips of a woman uttering syllables starting and ending with the letter “a” and that it was important that they keep their eyes on the screen. They were explicitly instructed to write down what they heard. They were also told that some of the videos might seem “odd” and that the audio and video tracks might seem mismatched. They were provided with a
A list of 18 possible responses to serve as a guide if they were unsure of what they heard, but they were also explicitly instructed that they could write down whatever they heard, even if it was not on the list. This response list included common fusion, combination responses as well as all the actual identities of the AO and VO stimuli. The list was [ABA, AGA, AKA, APA, ATA, ADA, ATHA, AVA, AHA, ASA, ABGA, AGBA, APKA, AKPA, APTA, ATPA, ABDA, ADBA]. A response was recorded as a “McGurk response” when it was an illusory percept that corresponded to neither unimodal stimuli. That is to say, any response that corresponded to either the auditory or visual track was not counted as an illusory response. VO-driven responses were not counted as McGurk responses to ensure the trials were bimodally processed. Many studies report VO-driven responses as McGurk responses therefore the analysis was run with and without the VO-driven responses and both analyses produced consistent results.

In the second part of the experiment the subjects performed an open-set single word identification task, delivering their responses by typing-in the word. Subjects were instructed to guess when they were unsure of words, but were also permitted to respond, “IDK” (I Don’t Know), when they were unable to guess. Presentations of AO, VO and AV trials were randomized on-line. Pluralized and conjugated responses were marked as such and subsequently marked as incorrect.

**Analysis:**

Analysis was done using custom written software using Microsoft Excel (2007, SP3 MSO) and MatLab (v. 7.5.0.342 - R2007b). Additional statistical analysis was performed using SPSS Statistics version 21.

122
Results:

McGurk Susceptibilities:

The average McGurk susceptibilities, defined as the percent of illusory responses (corresponding to neither the AO or VO stimulus), for the five stimulus categories are illustrated in Panel A of Figure 21. Three subjects never experienced the McGurk effect for any of the 5 McGurk stimuli (neither illusory nor VO-driven responses), these subjects were not dropped from subsequent analysis.

Speech Perception:

The average performance on the three speech perception tasks – consisting of single words in noise for AO and AV, and in silence for VO (noise was set at -5 dB) – is illustrated in Panel B of Figure 21. A repeated-measures ANOVA was performed using the three tasks as the within-subject effects. Performance on the three tasks was found to be significantly different \( [F(1.569, 45.514) = 500.792; p<0.001] \). A post-hoc pairwise comparison with Bonferroni correction indicated that the AO and VO performances were not significantly different from each other \( (p=1.000) \). The AV performance was significantly different from AO \( (p<0.001) \) and VO \( (p<0.001) \). Performance in the AV condition was substantially greater than the sum of the unimodal inputs, illustrating that super-additive performance enhancement took place.

A multiple linear regression analysis was performed to model AV using AO and VO as predictors. The resulting regression was significant \( [F(2,27) = 19.994; p<0.001] \), and \( R^2 = 0.597 \). The correlational coefficients indicated that VO was a significant...
predictor of AV \( \beta_1 = 0.810; t(27) = 6.242; p<0.001 \) but AO was not \( \beta_1 = 0.340; t(27) = 1.582; p=0.125 \). A stepwise regression was then performed with VO as the initial input; this determined that VO accounted for 56% of the variance observed in AV, while the composite model accounted for 59.7% of the variance.

Figure 21: Behavioural results. Panel A: Average susceptibilities to the McGurk illusion for nonsense VCV syllables presentations, the flanking vowel was /a/. Stimuli are labeled as [auditory, visual] consonant combinations. Results were measured as the average responses that were not the presented audio or video. The four significantly different token pairs are labeled. Error bars indicate the variance. Panel B: Average performance on three speech perception tasks – AO at -5 dB, VO in silence and AV at -5 dB. Stimuli were single monosyllabic high frequency words. Performance in AO and VO were not significantly different from each other. Performance in AV was significantly different from AO and VO and was substantially greater than the sum of the unimodal performances. Error bars indicate the standard deviation.
Relationship between McGurk Susceptibility and Speech Performance:

Multiple linear regressions were performed to predict the performance on the speech perception tasks using the 5 McGurk stimuli as predictors. The multicollinearity of the 5 McGurk stimuli was tested by calculating the tolerance between each stimulus and the remaining 4 stimuli using multiple linear regression analysis. All tolerance values were greater than 0.10 (between 0.258 and 0.934), indicating that each predictor contained a significant degree of variance that was unaccounted-for by the other predictors. Therefore, we were able to proceed with the multiple linear regressions using the 5 McGurk susceptibilities as independent predictors of performance.

The McGurk stimuli did not significantly predict performance in the AO condition [F(5,24) = 0.909; p = 0.491]. The McGurk stimuli did however significantly predict VO performance [F(5,24) = 2.675; p =0.047] with $R^2= 0.358$. The correlation coefficients indicated that the [K, P] McGurk stimulus was the only significant contributor to the model [$\beta_1 = -0.189; t(24) = -2.431; p =0.023$]. A step-wise multiple regression was then performed with susceptibility to the [K, P] stimulus entered first, and the remaining stimuli entered second. The susceptibility to [K, P] stimulus accounted for 29.8% of the variance in the VO performance. The negative correlation is illustrated in Figure 22.
A multiple regression analysis was performed to predict AV performance using the 5 McGurk stimuli as predictors. The resulting model significantly fit the data \([F(5, 24) = 3.693; p = 0.013]\) with \(R^2 = 0.435\). The stimulus [K, P] was the only significant contributing correlational coefficient \([\beta_1 = -0.166; t(24) = -2.156; p = 0.041]\). A subsequent step-wise multiple regression indicated the correlation with the [K,P] stimulus accounted for 24.9\% of the variance observed in the AV performance.

Another multiple regression analysis was performed to predict the difference in the AV and VO performance (AV-VO). This gave an estimate of a correlation between McGurk susceptibility and with the AV performance that was attributable to the auditory
channel and the integration process. This correlation was non significant \( F(5,21) = 1.103; p = 0.388 \). This indicated that McGurk susceptibility was not correlated with the AV performance that was not associated with VO performance.

The aforementioned regressions were repeated using a less stringent classification of McGurk stimuli - the susceptibility for McGurk stimuli that also included visually-driven responses. Similar to the previous results, the fit for the multiple regression to AV performance was significant \( F(5,24) = 2.994; p = 0.031; R^2 = 0.384 \). However, none of the individual correlation coefficients (the five McGurk stimuli) were significant predictors. This indicated that overall the model prediction was still significant, but the inclusions of visually-driven responses reduced the overall predictive power of the model.

**Relationship between McGurk susceptibility and Integration Enhancement:**

The McGurk susceptibilities were correlated with integration enhancement (IE) to determine the relationship between susceptibility and the AV performance enhancement associated with the integration process. The IE value was calculated according to the equation \( IE = (AV - AV_{pred}) \), where \( AV_{pred} = (AO + VO - R) \) - where R is the estimate of the redundant information in the auditory and visual channels. This equation is a generalized version of one proposed for closed-set tasks (Blamey et al., 1989; Tye-Murray et al., 2007a). R is measured experimentally as the rate of correct responses that occur in both the AO and VO conditions. This value is what the AV performance would
be if it were not super-additive, but rather if it were the sum of AO and VO performances. Next the IE was calculated as the difference between the linear sum $AV_{pred}$ and the true value for AV. This gives a measure of the amount of non-linear performance gain that is associated with the integration process.

A multiple regression analysis was conducted to determine if the McGurk stimuli could be used as predictors for the linear estimate of IE. That is to say, if the McGurk stimuli could predict the degree of performance enhancement associated with the integration process. This was a critical comparison because McGurk susceptibility is used as a proxy measure for the efficiency of the integration process. The correlation was found to be non-significant $[F(5,24) = 0.789; p=0.568]$, and none of the individual McGurk stimuli were significantly correlated with IE.

**Discussion:**

Susceptibility to five versions of the McGurk effect for VCV nonsense syllables was measured and compared to performance on three speech perception tasks for monosyllabic words: VO in silence, AO in noise and AV in noise. There were two main hypotheses. First, that McGurk susceptibility would be positively correlated with the IE metric which measures the degree of performance enhancement. And second, that McGurk susceptibility would not be correlated with the unimodal performances. Our results did not confirm these hypotheses.

In the current study, performance in the speech perception task agreed with previous findings for AO and AV performance (Ewertsen & Nielsen, 1971; O'Neill,
1954; Sommers et al., 2005; Sumby & Pollack, 1954), as well as VO performance (Dodd, 1977; Ewertsen & Nielsen, 1971; Tye-Murray et al., 2007a). The AV performance was also found to significantly correlate to the VO performance, which is consistent with the results of some previous work (Blamey et al., 1989; Ewertsen et al., 1970; MacLeod & Summerfield, 1990). Performance in the AO condition was not significantly correlated to performance in AV.

Previously reported rates of susceptibilities have been widely variable and have proven receptive to modulation. For example, it has been found that the [B,G] stimulus has susceptibilities between 0.45 (Paré, Richler, ten Hove, & Munhall, 2003) and 0.68 (Grant & Seitz, 1998). In the current study, our measured susceptibility was 0.31. Previously reported susceptibilities to the [K,P] stimulus range from: 0.18 (MacDonald & McGurk, 1978) to 0.90 (Sams et al., 1998) – though the latter result was for McGurk stimuli presented in noise. In the current study, the susceptibility was only 0.49. The low-reproducibility of the McGurk effect can be attributed to speaker effects, quality of the track alignments, large inter-subject variabilities (Schwartz, 2010) and different scoring methods. These factors make cross study comparisons for the magnitude of susceptibility problematic. For this reason, most research has concentrated on identifying modulatory effects on McGurk susceptibility. Indeed, the susceptibility has been found to be influenced by age (McGurk & MacDonald, 1976; Rosenblum et al., 1997), cochlear implant usage (Skipper et al., 2007), rate of the visual presentation (Green & Miller, 1985), stimulus onset asynchrony (Jones & Callan, 2003; Massaro & Cohen, 1993; Munhall et al., 1996; Tomaskovic et al., 2008; van Wassenhove et al., 2002), different
speakers (Cienkowski & Carney, 2002), noise masking (Sekiyama et al., 2003), attention effects (Alsius et al., 2005; Alsius et al., 2007; Soto-Faraco et al., 2004), lexical effects (Braccazio, 2004), context effects (Windmann, 2004) and priming (Kilian-Hutten et al., 2011). It is has also been reported that McGurk susceptibility requires that the subject be processing the visual stimulus as a face (Munhall et al., 2009) and the audio as speech (Vroomen & Stekelenburg, 2011). All of these studies are investigations into audiovisual presentations of speech, and they rely on the fundamental assumption that susceptibility to the McGurk is a proxy measure of speech integration. Modulation of the rate of susceptibility is assumed to be equivalent to the modulation of speech integration.

The relationship between McGurk susceptibility and speech-in-noise performance was investigated using multiple regression analysis. This method built a linear estimate of the speech performance using the McGurk susceptibilities as coefficients. It was found that only one of the McGurk stimuli, [K, P], significantly predicted performance. This stimulus predicted performance in the VO and the AV conditions and the correlation was negative – the opposite of the relationship that was expected. Further, the susceptibility was not expected to correlate with unimodal performance since this performance involves no multisensory integration. However a significant negative correlation was found between susceptibility for the [K, P] token and VO performance. Therefore, in our study greater McGurk susceptibility appears to be associated with a slightly decreased ability to identify the visual information as incongruent with the auditory information for this particular stimulus pair. This
relationship was relatively weak, with a low $\beta_1$-value, and most of the McGurk stimuli did not significantly correlated with AV performance.

Only one previous study has correlated McGurk susceptibility, amongst a battery of other tests, to speech-in-noise performance (Grant & Seitz, 1998). In that study, a significant positive correlation was found between the McGurk susceptibility and both the performance gain for AV consonants, and AV sentences. Unfortunately, our results cannot be directly compared to these because Grant and Seitz (1998) used a different measure of gain: $(AV-VO)/(1-AO)$. As discussed in Chapter 3 the choice of gain measurement can greatly affect the results. Our study employed a non-normalized gain score that did not treat the auditory and visual contributions to AV performance as independent. In an fMRI study of the McGurk effect, Jones and Callan (2003) reported that greater activity in V5 was associated with fewer perceptions of the McGurk effect. They attributed this observation to modulation of visual cortex by the auditory input. In the context of our results, if V5 activity is correlated with VO performance then this agrees with the observed negative correlation. In a study manipulating voicing boundaries, Brancazio (2004) asserted that decreased perception of the McGurk effect must be associated with decreased usage of the visual information. However, he observed that VO information continues to be used even when the McGurk effect was not observed. He concluded that McGurk susceptibility might be underestimating AV integration. These results also fit well with the observations from the current study. If the “VO modulation” of Brancazio’s study is comparable to the VO performance in our study, then increased VO performance is correlated with fewer McGurk illusions, as is
illustrated in Figure 22. However, this assertion is tempered by the findings of Cienkowski and Carney (2002) who observed no significant correlation between speechreading ability and the rate of McGurk responses.

Further work needs to be done to elucidate the complete relationship between susceptibility to the McGurk illusion and real speech stimuli. However, our results indicate that susceptibility to this illusion might not be associated with the same integration mechanism that enhances AV performance. Further, the correlation between susceptibility and VO is troubling because it suggests that modulation of the McGurk effect is not a proxy for integration modulation, but rather for the modulation of the VO input into the integration mechanism. However, this was a weak relationship that was only observed for only one of the five McGurk stimuli.
Chapter 6: Discussion

This thesis described four studies that examined characteristics of AV performance enhancement on a speech-in-noise task for open-set word identification. Our overarching goal was to develop foundational tools to examine the super-additive nature of AV performance enhancement. In the first study, it was determined that AV performance is highly variable across subjects as well as across studies, underscoring the volatility of the integration process. In the second study, a metric called the Integration Enhancement (IE) was developed to quantify the AV performance enhancement due to bimodal specific processes, such as integration, cortical interaction and cross-modal lexical congruency. This metric was specifically developed to be applicable to an open-set task. In the third study, a predictive model was developed to determine whether congruency in the AO and VO responses could account for the AV performance enhancement. This model was partially successful, accounting for the performance enhancement in 60% of the stimuli (words). The discrepancy indicates that full word confusions for AO and VO responses underrepresent the informational content of the auditory and visual signals in the AV percept. Therefore integration is occurring at the sub-lexical level. In the final study, the susceptibility to the McGurk illusion – a phenomenon that is often employed to investigate AV speech - was examined to determine its efficacy as a proxy for AV performance enhancement and by extension, integration. It was found that McGurk susceptibility was not correlated with the magnitude of the integration enhancement (as measured by the IE metric). Therefore a process other than the one that results in AV performance enhancement for words-in-
noise governs the susceptibility to the McGurk illusion.

In the first study, the relationship between AV performance and SNR was found to be consistent with a smooth, monotonically increasing, sigmoid function. In the second study AV performance was consistent with the phenomenon of super-additivity, wherein AV performance enhancement was greater than the corrected sum of AO and VO at all SNRs except for +10 dB. The results from these studies were consistent with most previous studies that measured the performance functions for AV and AO perception of words in noise (Erber, 1969, 1971b; Ewertsen & Nielsen, 1971; Ma et al., 2009; O'Neill, 1954; Sumby & Pollack, 1954).

The difference between AO and AV was found to be SNR dependent, generally decreasing with increasing SNR. In the second study, the difference score between AV and AO exhibited a broad peak at -6 and -2 dB. The VO performances reported in the second and fourth study were consistent with previous studies (Dodd, 1977; Ewertsen & Nielsen, 1971; Tye-Murray et al., 2007a). High within-subject correlations indicated a tight relationship between AO and AV for each subject. This suggests that the integration process is recruited in a consistent way for a given subject across SNRs.

The cross-study comparison of the first study indicated that there is a source of variance outside of the integration mechanism that is associated with the experimental methodology. The review illustrated the importance of developing a more standardized model to examine speech-in-noise. Different studies report substantially different results for similar tasks, indicating problems with reproducibility (Erber, 1969, 1971b; Ma et al.,
The variability could be due to subject features, speaker effects, different noise masks, word effects or whether the task uses open- or closed-sets. It has already been shown that performance on a speech-in-noise task will vary with different speakers (Grant & Braida, 1991). Additionally, different studies have used different word-types such as monosyllabic (Ma et al., 2009; Ross, Saint-Amour, Leavitt, Javitt, et al., 2007), generally bisyllabic (Sumby & Pollack, 1954), specifically spondees (Erber, 1969, 1971a), trochees (Erber, 1971a) or unreported (Ewertsen & Nielsen, 1971; O'Neill, 1954). Performance can also be modulated by word type, making the comparison of the AV performances tenuous. Finally, the set size of the task has been shown to exert a significant effect on performance (Sumby & Pollack, 1954). Therefore, as previously discussed by Ross, Saint-Amour, Leavitt, Javitt, et al. (2007), it may be problematic to compare results from open-set tasks (Ewertsen & Nielsen, 1971; Ma et al., 2009; Ross, Saint-Amour, Leavitt, Javitt, et al., 2007) to those from closed-set tasks (Erber, 1969, 1971b; O'Neill, 1954; Sumby & Pollack, 1954).

Differences in methodologies and presentation effects could also have influenced performance by impacting either the efficacy or recruitment of the integration mechanism. As previously reported Ma et al. (2009), the quality of the auditory and visual information can influence the magnitude of the AV performance enhancement. Therefore any methodological effects on the unimodal signals can also modulate AV performance enhancement.
In the second study, the IE metric was developed to quantify the amount of AV performance enhancement associated specifically with bimodal presentations. This measure is the absolute difference between the observed AV performance and a linear estimate of \( \text{AV}_{\text{predicted}} \). The \( \text{AV}_{\text{predicted}} \) value treats the AO and VO channels as completely independent sources of information; therefore, there are no super-additive properties in this estimate. The difference \( \text{AV} - \text{AV}_{\text{predicted}} \) is therefore the performance enhancement due to the bimodal-specific processes. These processes could be either gain modulatory processes that enhance the speech signal or processing constraints that reduce noise. For example, van Wassenhove (2007) proposed that during an AV presentation the visual signal, which arrives slightly before the auditory signal, primes cortex for optimal processing of the incoming auditory information. The system is maximally tuned to auditory signals that are congruent with the visual information. This means that the visual signal is applying constraints to the processing of the auditory signal, which effectively reduce noise in the auditory channel. This proposal is supported by EEG research indicating that visual information could put the auditory cortex into a high excitability state that increases intelligibility (Lakatos et al., 2007; Schroeder et al., 2008).

The term R (Redundancy) was defined to describe redundant information in the auditory and visual streams that lead to correct identification of the presented word at different SNRs. This metric is a corrective term developed to measure the deviation of the AV performance from the predicted linear sum of AO and VO in an open-set task.
Redundancy in the auditory and visual signal is a source of performance enhancement (Barlow, 2001). The IE was therefore a measure of the magnitude of the performance that is unaccounted-for by the unimodal streams.

We did not observe the ‘sweet-spot’ phenomenon – an intermediate zone of maximal integration (Ross, Saint-Amour, Leavitt, Javitt, et al., 2007) – in the first study. In the second however, the IE curve – a measure of AV performance enhancement – peaked at -2 dB. This could have been due to the sweet spot phenomenon. But it might also be explained by a limitation of the data. Because the task was whole word identification the performance in the AO condition underestimated the actual amount of auditory information that was available in the AV condition. The AO performance reflects the availability of information sufficient for whole word identification. This information may be insufficient for performance gains in the AO condition, but it might be sufficient when coupled with visual information. Therefore the difference score (AV-AO) and IE metric represent performance differences, but these are not equal to the differences in informational content. Information that is unavailable to boost performance in the AO condition, such as auditory timing, is easily exploited in the AV condition.

The under-estimation of auditory information by the AO performance measure could partially explain the super-additive enhancement in AV performance. To illustrate this point Figure 23 shows a different scoring method applied to the data from Chapter 3 (see Figure 9 in Chapter 3). Performance reflects correct responses to the nucleus of the
syllable in each word stimulus, for our word list this was always the vowel (underlined phonemes listed in Appendix A). Therefore, if the response word provided was incorrect, but that word had the same vowel as the presented word, then in this scoring scheme, the response was marked as correct. For example, if the word “map” was presented, the vowel for this word is “a” - corresponding to the phoneme /AE/, if the response given were “cash” this would be marked as correct because the vowel of the word “cash” is also /AE/. Comparatively, if the response were “milk”, this would be marked as incorrect because the vowel of “milk” is “/IH/. The performance in this case, as compared to whole word scoring, is more robust to noise across all SNRs for the AO, VO and AV conditions, illustrating how there is a substantial amount of information in the auditory signal that is underestimated by whole word AO performance. At -10 dB for example, whole word identification is 2%, whereas the performance on vowels is 23%. The IE metric has been calculated for these performances (bottom panel of Figure 23). This curve exhibits different behaviour than the one calculated from the performance in whole word identification (Figure 13 in Chapter 3). The IE curve is only super-additive at SNRs below -2 dB; above this noise level the performance enhancement is either additive or sub-additive. In the identification task the AV performance enhancement (measured with the IE metric) was super-additive at all SNR levels. This high IE value was due, at least in part, to the underestimation of the auditory information in AV, by the AO performance – particularly at low SNRs. The more modest range of IE values for the vowel performance is consistent with findings for neural spiking and behavioural detection tasks which indicate that super-additive response enhancement for bimodal stimuli only
occurs when the unimodal signals are very degraded, and when the signal quality is improved performance enhancement quickly drops to additive or sub-additive levels (Stanford & Stein, 2007).

Figure 23: Performance on vowels. A different scoring method was applied to the data from Chapter 3. Performance was measured according to whether the response word had the same vowel as the presented word. The high performance values show how whole word identification performance underestimates the actual informational content of the signals.
The variability in the AV performance at low SNRs in the first study is consistent with an integration process that uses complementary visual information to extract meaning from auditory information that would, on its own, be insufficient for word identification (Calvert et al., 1997; Mottonen et al., 2004; Paulesu et al., 2003; Sams et al., 1991; Stevenson et al., 2012; van Wassenhove et al., 2005). Alternatively, the visual channel could act by depressing noise in the auditory channel (van Wassenhove, 2007) at lower SNRs.

The high degree of between-subject variability is also reflected in the analysis of 95% confidence interval widths that decrease with SNR for the AV condition but not the AO condition (Figure 3 of Chapter 2). The variability in the AO condition is consistent with static additive noise and is not SNR dependent. The variable AV performance at lower SNRs is consistent with a source of variance in the AV speech perception mechanism that is a function of the SNR. Further, because the differences in performance enhancement at lower SNRs are due to differential efficacy with which the mechanisms associated with performance enhancement are recruited. The increased variance reflects the different degrees to which the integration mechanism may be operating. This is consistent with a mechanism that is preferentially activated or recruited at lower SNRs.

High subject variability at low signal-to-noise-ratios has been reported in numerous studies, but none of these studies have been able to fully account for the observed variance (Blamey et al., 1989; Erber, 1969; Grant et al., 1998; MacLeod & Summerfield, 1987; Middelweerd & Plomp, 1987). The between-subject variability in AV decreases to below the AO variability baseline at higher SNRs (-2dB). There are two implications for
this finding. First, the quality of the AV percept is variable at low SNRs – that is to say, some people are good at the task, while some people are not. And second, at higher SNRs the AV percepts are more consistent (lower variability) than the unimodal AO percept. A combinatorial process that leads to an overall decrease in variance could indicate that AV integration is a Maximum-Likelihood-Estimation procedure – as previously suggested by Massaro and Cohen (2000).

\[ AV_{\text{predicted}} \text{ removed the covariance of } VO \text{ from } AV \text{ Performance} \]

Consistent with previous findings, the average AV scores in the second study were significantly correlated with VO (Blamey et al., 1989; Ewertsen et al., 1970; MacLeod & Summerfield, 1987, 1990) across subjects. Variance in the VO condition accounted for 33.7% of the variance in the averaged AV condition (averaged across the SNRs -10 to +6 dB). In the fourth study, where AV performance was only measured at -5 dB, correlation with VO accounted for 56% of the variance. No significant correlation was found between AO and VO performance for either the second or fourth study. Additionally, the correlation between AO and AV was also non-significant for both the second and third data collection.

The correlation between the average AV performance and the linear estimate \( AV_{\text{predicted}} \) was significant, accounting for 29% of the variance observed in the average AV condition in the second study and 55.4% of the variance observed in the fourth study. The higher correlation in the latter study is likely because the AV performance was collected at only one SNR (-5 dB). As was illustrated in the second study (Figure 13 in
Chapter 3) the contributions of auditory information (estimated as AO performance), visual information (estimated as VO performance) and Integration are SNR dependent. The contribution from visual information falls quickly from nearly 50% of the AV performance at -10 dB to near zero a +10 dB. Therefore it might be expected that AV performance will be more correlated with VO performance at low SNRs where the contribution is greatest. This could explain why the correlation between AO and VO in the fourth study is greater than that of the second study.

The correlations between AV and AV_{predicted} were approximately the same for the respective correlations between AV and VO for the second and fourth study. The AO therefore did not add to the predictive power of the estimate of bimodal performance. This indicated that individual variance in the AO performance did not predict individual variance in the AV performance, while individual variance in the VO performance did. This was also reflected in the aforementioned low correlation between AO and AV performance. This low correlation can at least be partially attributed to the large inter-subject variability in the AV performance that has been previously reported (Blamey et al., 1989; Erber, 1969; MacLeod & Summerfield, 1990; Middelweerd & Plomp, 1987). As a result, the variability in AV exceeds the range of AO performances, diminishing the correlation. The covariance between the two measures could also be due to pre-integration signal modification by the VO channel (van Wassenhove, 2007). There is physiological evidence for such an interaction in the form of early cortical interactions (Falchier et al., 2002; Romanski et al., 1999; Smiley et al., 2007), which is also supported by functional imaging studies (Calvert et al., 1997; Calvert & Campbell, 2003; Mottonen
et al., 2004; Paulesu et al., 2003; Pekkola et al., 2005; Sams et al., 1991). A third possible explanation for the low correlation between AO and AV performance is a discrepancy in the statistical analyses. Firstly, we assumed that the VO performance would be constant across all SNR level and therefore VO performance measures were only collected in silence. We analyzed the covariance between AV and VO using a multivariate repeated-measures ANOVA or MANCOVA to assess the relationship. As mentioned earlier, this relationship was significant. Comparatively, AO and AV were both repeated-measures. The research in to statistical methods to correlate two dependent variables on repeated-measures is ongoing (Roy, 2006, 2009). Therefore the analysis was approached in a similar way as it was to correlate AV and VO. A MANCOVA was used with all six AO performances entered as covariates. Each AO performance was therefore treated as an independent covariate of the repeated-measures of AV performance. This left the covariance between the AO measures unaccounted for. This assumption could have led to an underestimation of the relationship between AV and AO performance. Although the individual variances in AO do not reflect those in AV, as reported in the first study, the two measures covary strongly across SNRs.

Predicting Super-Additive Performance:

In the third study a predictive model was developed to account for super-additive performance enhancement in AV word presentations. This model used congruency in the unimodal confusions to impose constraints on which word would be enhanced. Hence, it
was named the ‘Confusion Constraints’ model. The prediction accounted for super-additive performance in 60% of the stimuli (Figure 19 in Chapter 4).

The 40% discrepancy is consistent with a sub-lexical integration process that is not captured by whole word constraints. Other studies have found evidence that integration occurs at the phonemic or sub-phonemic level (Fort et al., 2013; Fowler, 1986; Green, 1998; Green & Miller, 1985; Skipper et al., 2007). The model was limited to cases where the AO signal quality is strong enough to provide enough information for whole word identification, and it has no access to intermodal information. AO performance underestimates the informational content of the actual auditory information available in the AV condition, which is why AV performance remains SNR dependent at noise levels where AO performance is at floor (Erber, 1969, 1971a; Ma et al., 2009). Additionally, intermodal information carried by the temporal correlations between auditory and visual speech signals is a source of information during speech perception (Grant & Greenberg, 2001; Grant & Seitz, 2000; Tanaka et al., 2009; Vatakis et al., 2012). The model does not account for such elements because the lexical constraint is applied based on independent unimodal responses. This was also reflected in the subgroup of words whose super-additive performance was predicted by the model. The IE analysis revealed that the model under-predicted performance at low SNRs (Figure 20 of Chapter 4); indicating the model was not able to capture cues that facilitated perception at low SNRs.

The results of the confusions constraints model were inconsistent with the late-stage independent lexical-level integration process proposed by Tye-Murray et al.
(2007b), but consistent with the dependent model of integration wherein the sensory signals are modified and do not provide estimates of the solution prior to integration. This model of integration is also consistent with the observed low correlation between AO and AV in studies 2 and 3 as well as other evidence that visual cortex has a modulatory effect on auditory cortex (Calvert et al., 1997; Calvert & Campbell, 2003; Mottonen et al., 2004; Paulesu et al., 2003; Pekkola et al., 2005; Sams et al., 1991).

**The McGurk Effect is not a Measure of AV Speech Integration:**

In the fourth study we examined the relationship between AV performance enhancement and performance on a task often used to measure AV integration – the McGurk illusion. Performance on the McGurk task is measured as susceptibility to an illusory percept resulting from the presentation of particular incongruent AV stimuli. It was found that McGurk susceptibility did not correlate with the degree of performance enhancement (measured with IE). Further, the susceptibility was negatively correlated with both the AV and VO performances, but not with the performance difference (AV-VO). Meaning that better performance in the VO condition indicates a greater ability to recognize incongruence in the McGurk stimulus. Indicating that lower McGurk susceptibility indicated an ability to detect incongruence and that these subjects were able to discard the VO signal. These results indicate that McGurk susceptibility is a measure of a subject’s propensity for erroneous VO influences, and not necessarily the integration mechanism associated enhanced AV performance. Further, the susceptibility for different types of McGurk stimuli was found to be variable within-subjects, indicating that the
process that produced the illusory percept could have been differentially recruited for each McGurk stimulus. Alternatively, differences in the informational content of the auditory and visual channels were not constant across these pairs and might also explain the observed differences. Inherent characteristics of the consonants, as articulated by our speaker, likely account for large differences in the susceptibilities. Prior studies have found substantial speaker effects (Cienkowski & Carney, 2002; Paré et al., 2003) and it could also be possible that within-speaker differences could have accounted for some of the variability.

There is some evidence that McGurk and AV speech perception are separate mechanisms. Functional activity associated with the perception of McGurk is different that that for more general incongruent stimuli (Szycik, Stadler, Tempelmann, & Munte, 2012). Further, activation for the illusory McGurk percept is different from the activity for the auditory-driven responses (Skipper et al., 2007). Finally, perception of the McGurk illusion leads to increased ERP signal latencies indicating longer processing times (van Wassenhove et al., 2005). Given that we observed a negative correlation with VO performance it is possible that this additional or differential processing reflects an evaluation of the visual data. Better visual information – as would be collected by people with greater speechreading skills – could allow the brain to better identify the visual signal as incongruent and irrelevant for the task – which is to report what was heard. Therefore, the susceptibility would not be a measure of AV integration, but rather the efficiency of the evaluation of VO.
Conclusions:

The results presented in this thesis provide insight into several aspects of AV speech integration. First I highlighted the importance of having more tightly controlled experimental designs in audio-visual speech perception research. Tighter controls would help the research community better investigate the nebulous integration process by making studies more readily comparable.

Second, I developed a gain metric that quantifies the AV performance enhancement for an open-set task. This measure isolates the AV performance gain that cannot be accounted for by the AO and VO performance data. It is therefore specifically associated with the bimodal condition.

Third, I developed a data-driven model to predict performance enhancement in an open-set task. This has not been attempted before as previous estimates have only been made for closed-set syllable stimuli. Although the results from this model were only able to account for super-additive performance in 60% of the words in our stimulus set, this methodology could be developed and expanded to model phonemic and sub-phonemic AV performance enhancement. This would allow future research to determine whether the integration processes is happening via an independent or dependent mechanism: An independent mechanism being one in which auditory and visual channels are independently processed in the presence of constraints, and a dependent mechanism being one in which the a bimodal presentation provides the system with intermodal information unavailable during bimodal presentations.
Finally, I addressed a long-standing assumption in AV speech perception literature: that the susceptibility to the McGurk effect is a measure of integration. I was unable to find evidence that this is the case. I instead found evidence that susceptibility to the McGurk is uncorrelated with super-additive AV performance enhancement and is instead negatively correlated with speechreading performance. Therefore, susceptibility to this illusion does not appear to be associated with the same integration mechanism that enhances AV speech performance. This result would be well supplemented with an additional study that examines the susceptibility to McGurk for non-sense syllables and compares it to performance on an AV syllable-in-noise task. Such a study would remove the lexical facilitation from the AV performance measure and hence might better correlate with the susceptibility measure.

Research regarding multi-sensory integration involves competing theories regarding how integration is achieved. In particular, there has been an evolution from simple theories of neural convergence (Stein et al., 1988) to more dynamic models of independent processing (Braida, 1991; Massaro & Friedman, 1990), and now growing support for dependent processing (van Wassenhove, 2007). My results contributed to this field by addressing assumptions inherent to AV perception research, and tackling problems regarding experimental design and metrics for data analysis described in the literature. These contributions will help guide future studies examining theories of AV speech integration.
References


Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and


enhancement of speech comprehension under noisy environmental conditions. 
Schizophr Res, 97(1-3), 173-183. doi: 10.1016/j.schres.2007.08.008


170


### Appendix A: Word List with Phonemic Transcription

<table>
<thead>
<tr>
<th>Word</th>
<th>Phonemic Transcription</th>
<th>Word</th>
<th>Phonemic Transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td>aim</td>
<td>EY M</td>
<td>bride</td>
<td>B R AY D</td>
</tr>
<tr>
<td>ant</td>
<td>AE N T</td>
<td>bridge</td>
<td>B R IH JH</td>
</tr>
<tr>
<td>ape</td>
<td>EY P</td>
<td>bronze</td>
<td>B R AA N Z</td>
</tr>
<tr>
<td>arm</td>
<td>AA R M</td>
<td>bump</td>
<td>B AH M P</td>
</tr>
<tr>
<td>axe</td>
<td>AE K S</td>
<td>burn</td>
<td>B ER M P</td>
</tr>
<tr>
<td>bag</td>
<td>B AE G</td>
<td>bus</td>
<td>B AH S</td>
</tr>
<tr>
<td>ball</td>
<td>B AO L</td>
<td>bush</td>
<td>B UH SH</td>
</tr>
<tr>
<td>band</td>
<td>B AE N D</td>
<td>cab</td>
<td>K AE B</td>
</tr>
<tr>
<td>bank</td>
<td>B AE NG K</td>
<td>calf</td>
<td>K AE F</td>
</tr>
<tr>
<td>bar</td>
<td>B AA R</td>
<td>camp</td>
<td>K AE M P</td>
</tr>
<tr>
<td>bark</td>
<td>B AA R K</td>
<td>cap</td>
<td>K AE P</td>
</tr>
<tr>
<td>bat</td>
<td>B AE T</td>
<td>carp</td>
<td>K AA R P</td>
</tr>
<tr>
<td>bath</td>
<td>B AE TH</td>
<td>cart</td>
<td>K AA R T</td>
</tr>
<tr>
<td>bay</td>
<td>B EY</td>
<td>cash</td>
<td>K AE SH</td>
</tr>
<tr>
<td>beach</td>
<td>B IY CH</td>
<td>cast</td>
<td>K AE S T</td>
</tr>
<tr>
<td>bear</td>
<td>B EH R</td>
<td>cat</td>
<td>K AE T</td>
</tr>
<tr>
<td>bed</td>
<td>B EH D</td>
<td>chain</td>
<td>CH EY N</td>
</tr>
<tr>
<td>bee</td>
<td>B IY</td>
<td>chair</td>
<td>CH EH R</td>
</tr>
<tr>
<td>bell</td>
<td>B EH L</td>
<td>cheek</td>
<td>CH IY K</td>
</tr>
<tr>
<td>belt</td>
<td>B EH L T</td>
<td>chest</td>
<td>CH EH S T</td>
</tr>
<tr>
<td>bench</td>
<td>B EH N CH</td>
<td>chief</td>
<td>CH IY F</td>
</tr>
<tr>
<td>bend</td>
<td>B EH N D</td>
<td>chin</td>
<td>CH IH N</td>
</tr>
<tr>
<td>bike</td>
<td>B AY K</td>
<td>clam</td>
<td>K L AE M</td>
</tr>
<tr>
<td>bill</td>
<td>B IH L</td>
<td>clamp</td>
<td>K L AE M P</td>
</tr>
<tr>
<td>bin</td>
<td>B IH N</td>
<td>claw</td>
<td>K L AO</td>
</tr>
<tr>
<td>bird</td>
<td>B ER D</td>
<td>clay</td>
<td>K L EY</td>
</tr>
<tr>
<td>bite</td>
<td>B AY T</td>
<td>cloak</td>
<td>K L OW K</td>
</tr>
<tr>
<td>blood</td>
<td>B L AH D</td>
<td>clock</td>
<td>K L AA K</td>
</tr>
<tr>
<td>bolt</td>
<td>B OW L T</td>
<td>cloth</td>
<td>K L AO TH</td>
</tr>
<tr>
<td>bond</td>
<td>B AA N D</td>
<td>cloud</td>
<td>K L AW D</td>
</tr>
<tr>
<td>bow</td>
<td>B AW</td>
<td>club</td>
<td>K L AH B</td>
</tr>
<tr>
<td>bowl</td>
<td>B OW L</td>
<td>coal</td>
<td>K OW L</td>
</tr>
<tr>
<td>box</td>
<td>B AA K S</td>
<td>coast</td>
<td>K OW S T</td>
</tr>
<tr>
<td>brain</td>
<td>B R EY N</td>
<td>cold</td>
<td>K OW L D</td>
</tr>
<tr>
<td>brass</td>
<td>B R AE S</td>
<td>cord</td>
<td>K AO R D</td>
</tr>
<tr>
<td>bread</td>
<td>B R EY D</td>
<td>cork</td>
<td>K AO R K</td>
</tr>
<tr>
<td>Word</td>
<td>phonemes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>----------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>corn</td>
<td>K AO R N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cow</td>
<td>K AW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>crab</td>
<td>K R AE B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cream</td>
<td>K R IY M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>crew</td>
<td>K R UW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>crow</td>
<td>K R OW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cube</td>
<td>K Y UW B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cult</td>
<td>K AH L T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cup</td>
<td>K AH P</td>
<td></td>
<td></td>
</tr>
<tr>
<td>curb</td>
<td>K ER B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dark</td>
<td>D AA R K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dart</td>
<td>D AA R T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dent</td>
<td>D EH N T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>desk</td>
<td>D EH S K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dime</td>
<td>D AY M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dirt</td>
<td>D ER T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>disc</td>
<td>D IH S K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td>D AO G</td>
<td></td>
<td></td>
</tr>
<tr>
<td>doll</td>
<td>D AA L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>drug</td>
<td>D R AH G</td>
<td></td>
<td></td>
</tr>
<tr>
<td>drum</td>
<td>D R AH M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>duck</td>
<td>D AH K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dusk</td>
<td>D AH S K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dust</td>
<td>D AH S T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ear</td>
<td>IY R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fall</td>
<td>F AO L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fan</td>
<td>F AE N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fang</td>
<td>F AE NG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>farm</td>
<td>F AA R M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fast</td>
<td>F AE S T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fault</td>
<td>F AO L T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>feast</td>
<td>F IY S T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fee</td>
<td>F IY</td>
<td></td>
<td></td>
</tr>
<tr>
<td>felt</td>
<td>F EH L T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fence</td>
<td>F EH N S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>film</td>
<td>F IH L M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>firm</td>
<td>F ER M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fish</td>
<td>F IH SH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fist</td>
<td>F IH S T</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word</th>
<th>phonemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>flag</td>
<td>F L AE G</td>
</tr>
<tr>
<td>flood</td>
<td>F L AH D</td>
</tr>
<tr>
<td>flow</td>
<td>F L OW</td>
</tr>
<tr>
<td>fog</td>
<td>F AA G</td>
</tr>
<tr>
<td>fork</td>
<td>F AO R K</td>
</tr>
<tr>
<td>fox</td>
<td>F AA K S</td>
</tr>
<tr>
<td>frame</td>
<td>F REY M</td>
</tr>
<tr>
<td>frog</td>
<td>F R AA G</td>
</tr>
<tr>
<td>fruit</td>
<td>F R UW T</td>
</tr>
<tr>
<td>fund</td>
<td>F AH N D</td>
</tr>
<tr>
<td>fur</td>
<td>F ER</td>
</tr>
<tr>
<td>fuse</td>
<td>F Y UW Z</td>
</tr>
<tr>
<td>gang</td>
<td>G AE NG</td>
</tr>
<tr>
<td>gas</td>
<td>G AE S</td>
</tr>
<tr>
<td>gem</td>
<td>JH EH M</td>
</tr>
<tr>
<td>germ</td>
<td>JH ER M</td>
</tr>
<tr>
<td>gift</td>
<td>G IH F T</td>
</tr>
<tr>
<td>gold</td>
<td>G OW L D</td>
</tr>
<tr>
<td>golf</td>
<td>G AA L F</td>
</tr>
<tr>
<td>gown</td>
<td>G AW N</td>
</tr>
<tr>
<td>green</td>
<td>G R IY N</td>
</tr>
<tr>
<td>guest</td>
<td>G EH S T</td>
</tr>
<tr>
<td>gum</td>
<td>G AH M</td>
</tr>
<tr>
<td>harm</td>
<td>HH AA R M</td>
</tr>
<tr>
<td>hat</td>
<td>HH AE T</td>
</tr>
<tr>
<td>hawk</td>
<td>HH AO K</td>
</tr>
<tr>
<td>hay</td>
<td>HH EY</td>
</tr>
<tr>
<td>hill</td>
<td>HH IH L</td>
</tr>
<tr>
<td>hint</td>
<td>HH IH N T</td>
</tr>
<tr>
<td>hole</td>
<td>HH OW L</td>
</tr>
<tr>
<td>hook</td>
<td>HH UH K</td>
</tr>
<tr>
<td>horn</td>
<td>HH AO R N</td>
</tr>
<tr>
<td>host</td>
<td>HH OW S T</td>
</tr>
<tr>
<td>hound</td>
<td>HH AW N D</td>
</tr>
<tr>
<td>hunt</td>
<td>HH AH N T</td>
</tr>
<tr>
<td>hut</td>
<td>HH AH T</td>
</tr>
<tr>
<td>ice</td>
<td>AY S</td>
</tr>
<tr>
<td>ink</td>
<td>IH NG K</td>
</tr>
<tr>
<td>jam</td>
<td>JH AE M</td>
</tr>
<tr>
<td>jar</td>
<td>JH AA R</td>
</tr>
<tr>
<td>jaw</td>
<td>JH AO</td>
</tr>
<tr>
<td>jet</td>
<td>JH EH T</td>
</tr>
<tr>
<td>jolt</td>
<td>JH OW L T</td>
</tr>
<tr>
<td>joy</td>
<td>JH OY</td>
</tr>
<tr>
<td>judge</td>
<td>JH AH JH</td>
</tr>
<tr>
<td>jug</td>
<td>JH AH G</td>
</tr>
<tr>
<td>keg</td>
<td>K EH G</td>
</tr>
<tr>
<td>kick</td>
<td>K IH K</td>
</tr>
<tr>
<td>king</td>
<td>K IH NG</td>
</tr>
<tr>
<td>kiss</td>
<td>K IH S</td>
</tr>
<tr>
<td>knee</td>
<td>N IY</td>
</tr>
<tr>
<td>lamb</td>
<td>L AE M</td>
</tr>
<tr>
<td>lamp</td>
<td>L AE M P</td>
</tr>
<tr>
<td>land</td>
<td>L AE N D</td>
</tr>
<tr>
<td>lap</td>
<td>L AE P</td>
</tr>
<tr>
<td>leg</td>
<td>L EH G</td>
</tr>
<tr>
<td>lie</td>
<td>L AY</td>
</tr>
<tr>
<td>lift</td>
<td>L IH F T</td>
</tr>
<tr>
<td>limb</td>
<td>L IH M</td>
</tr>
<tr>
<td>limp</td>
<td>L IH M P</td>
</tr>
<tr>
<td>link</td>
<td>L IH NG K</td>
</tr>
<tr>
<td>lip</td>
<td>L IH P</td>
</tr>
<tr>
<td>list</td>
<td>L IH S T</td>
</tr>
<tr>
<td>lock</td>
<td>L AA K</td>
</tr>
<tr>
<td>log</td>
<td>L AO G</td>
</tr>
<tr>
<td>lord</td>
<td>L AO R D</td>
</tr>
<tr>
<td>luck</td>
<td>L AH K</td>
</tr>
<tr>
<td>lung</td>
<td>L AH NG</td>
</tr>
<tr>
<td>lust</td>
<td>L AH S T</td>
</tr>
<tr>
<td>mall</td>
<td>M AO L</td>
</tr>
<tr>
<td>map</td>
<td>M AE P</td>
</tr>
<tr>
<td>mark</td>
<td>M AA R K</td>
</tr>
<tr>
<td>mat</td>
<td>M AE T</td>
</tr>
<tr>
<td>math</td>
<td>M AE TH</td>
</tr>
<tr>
<td>meal</td>
<td>M IY L</td>
</tr>
<tr>
<td>milk</td>
<td>M IH L K</td>
</tr>
<tr>
<td>mist</td>
<td>M IH S T</td>
</tr>
<tr>
<td>mop</td>
<td>M AA P</td>
</tr>
<tr>
<td>moss</td>
<td>M AO S</td>
</tr>
</tbody>
</table>
pub  PAHB
pulp  PAHLP
pulse  PAHLS
pump  PAHMP
queen  W IY N
rack  RAEK
raft  RAEFT
ram  RAE
ramp  RAE MP
rash  RAE SH
rat  RAE T
ray  EY
rent  EHNT
rib  IH
rim  IHM
ring  IHN G
risk  IH SK
roast  OWST
rock  AAK
rod  AAD
rug  AAG
rum  AAM
rust  AAST
sack  AEK
salt  AO LT
sand  AE ND
scar  K AR
screw  KR UW
scroll  KR OW L
self  EHLF
sheep  HYP
sheet  HIY T
shell  HEHL
shift  IHFT
ship  IHP
shoe  UW
sigh  AY
sign  AYN
silk  IHLK
sin  IHN
sink  IHN K
skin  K IY N
sky  KAY
smoke  M OWK
snail  N EYL
snow  NO W
sock  AA K
soil  OYL
song  AO NG
soul  OWL
soy  OY
spear  PIH R
spike  P AY K
spoon  P UW N
spray  P REY
spring  P R IH NG
sprout  P R AW T
staff  TAE F
stain  TEY N
stair  TEHR
star  T AR
steak  TEY K
steam  TIYM
steel  TIY L
string  TRIH NG
sun  AHN
sweat  W EH
sack  K R OW
tap  AE P
tar  AA R
task  AESK
taunt  AO NT
tea  IY
tent  EHN T
test  EHS T
thief  IYF
thigh  AY
thread  R EH D
throat  R OW T
<table>
<thead>
<tr>
<th>Word</th>
<th>Phoneme</th>
<th>Word</th>
<th>Phoneme</th>
</tr>
</thead>
<tbody>
<tr>
<td>tin</td>
<td>T IH N</td>
<td>wand</td>
<td>W AA N D</td>
</tr>
<tr>
<td>tint</td>
<td>T IH N T</td>
<td>wax</td>
<td>W AE K S</td>
</tr>
<tr>
<td>toll</td>
<td>T OW L</td>
<td>wealth</td>
<td>W EH L TH</td>
</tr>
<tr>
<td>tomb</td>
<td>T UW M</td>
<td>web</td>
<td>W EH B</td>
</tr>
<tr>
<td>tool</td>
<td>T UW L</td>
<td>wedge</td>
<td>W EH JH</td>
</tr>
<tr>
<td>toy</td>
<td>T OY</td>
<td>west</td>
<td>W EH S T</td>
</tr>
<tr>
<td>train</td>
<td>T R EY N</td>
<td>wheat</td>
<td>W IY T</td>
</tr>
<tr>
<td>tray</td>
<td>T R EY</td>
<td>wheel</td>
<td>W IY L</td>
</tr>
<tr>
<td>treat</td>
<td>T R IY T</td>
<td>whip</td>
<td>W IH P</td>
</tr>
<tr>
<td>trend</td>
<td>T R EH N D</td>
<td>wind</td>
<td>W AY N D</td>
</tr>
<tr>
<td>trust</td>
<td>T R AH S T</td>
<td>wing</td>
<td>W IH NG</td>
</tr>
<tr>
<td>twig</td>
<td>T W IH G</td>
<td>wink</td>
<td>W IH NG K</td>
</tr>
<tr>
<td>valve</td>
<td>V AE L V</td>
<td>wolf</td>
<td>W UH L F</td>
</tr>
<tr>
<td>van</td>
<td>V AE N</td>
<td>wood</td>
<td>W UH D</td>
</tr>
<tr>
<td>vault</td>
<td>V AO L T</td>
<td>wool</td>
<td>W UH L</td>
</tr>
<tr>
<td>vent</td>
<td>V EH N T</td>
<td>worm</td>
<td>W ER M</td>
</tr>
<tr>
<td>vest</td>
<td>V EH S T</td>
<td>wrench</td>
<td>R EH N CH</td>
</tr>
<tr>
<td>wage</td>
<td>W EY JH</td>
<td>yard</td>
<td>Y AA R D</td>
</tr>
<tr>
<td>wall</td>
<td>W AO L</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>