Use Authorization

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at Idaho State University, I agree that the Library shall make it freely available for inspection. I further state that permission to download and/or print my thesis for scholarly purposes may be granted by the Dean of the Graduate School, Dean of my academic division, or by the University Librarian. It is understood that any copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Signature ___________________________________

Date ______________________________________

Some Normative Data for Audiovisual Speech Integration Skills

by

Daniel Carnley, B.S.

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in the Department of Communication Sciences and Disorders

Idaho State University

(Summer 2014)

Committee Approval

To the Graduate Faculty:

The members of the committee appointed to examine the thesis of Daniel Carnley find it satisfactory and recommend that it be accepted.

Dr. Nicholas Altieri,
Major Advisor

Dr. Anthony Seikel,
Committee Member

Dr. David Mercaldo,
Graduate Faculty Representative

Human Subjects Committee Approval Page

IDAHO STATE UNIVERSITY
HUMAN SUBJECTS COMMITTEE
NOTICE OF ACTION

RESEARCH PROPOSAL TITLE: Multisensory Integration

INVESTIGATORS: Dr. Nicholas Altieri, Daniel Carnley B.S.

SPONSORING AGENCY: ISU

PROPOSAL NO.: 3801

HUMAN SUBJECTS COMMITTEE ACTION:

_XX_ PROPOSAL APPROVED AS IS

____ PROPOSAL APPROVED PENDING MINOR MODIFICATIONS. SUBMIT MODIFICATIONS FOR FINAL APPROVAL. SEE PAGE 22-25 IN THE MANUAL OF POLICIES AND PROCEDURES OF THE HUMAN SUBJECTS COMMITTEE FOR AN EXPLANATION OF THE PROCEDURES TO BE FOLLOWED. **PLEASE BOLD ALL MODIFICATIONS MADE TO THE MODIFIED PROPOSAL!!**

____ PROPOSAL REQUIRES MAJOR MODIFICATIONS. SUBMIT 12 COPIES OF THE REVISED PROPOSAL FOR COMMITTEE REVIEW. SEE PAGE 22-25 IN THE MANUAL OF POLICIES AND PROCEDURES OF THE HUMAN SUBJECTS COMMITTEE FOR AN EXPLANATION OF THE PROCEDURES TO BE FOLLOWED. **PLEASE BOLD ALL MODIFICATIONS MADE TO THE MODIFIED PROPOSAL!!**

____ PROPOSAL WAS DISAPPROVED

Ralph Baergen____________ September 12, 2013
Human Subjects Committee Chair Date
(Signature Letter Available Upon Request)

Note: Approval is for a maximum period of one year. Projects extending beyond that time period should be renewed.

The researcher must notify Human Subjects Committee immediately in cases where the subject is harmed. Information (e.g. adverse reactions, unexpected events/outcomes) that may impact on the risk/benefit ratio should also be reported to, and reviewed by the HSC to ensure adequate protection of the welfare of the subjects.

_x_ Investigator
_x_ Dean of Research
_x_ Office of Sponsored Programs
_x_ Human Subjects Committee

Table of Contents

List of Figures……………………………………………………………………vi

List of Tables……………………………………………………………………vii

Abstract …………………………………………………………………………viii

Chapter 1: Introduction……………………………………………………………1

Processing Models…………………………………………………………………2

Measures of Audiovisual Integration………………………………………………7

Obtaining Normative Measures……………………………………………………16

Chapter 2: Experimental Methods…………………………………………………19

Participants………………………………………………………………………19

Audiometric Testing………………………………………………………………19

Single Word Recognition Task……………………………………………………20

Sentence Processing Task…………………………………………………………21

Chapter 3: Results…………………………………………………………………22

Sentence Processing Task…………………………………………………………22

Single Word Recognition Task……………………………………………………24

Chapter 4: Discussion and Conclusion……………………………………………28

Future Applications………………………………………………………………29

References………………………………………………………………………31

Appendix A………………………………………………………………………36

List of Figures

Figure 1: Visual representation of proposed integration models……………………………5

Figure 2: Mathematical representation of traditional capacity (C(t))……………………..9

Figure 3: A visual representation of unlimited, limited and super capacity systems……10

Figure 4: (C_I(t)) capacity measure including both accuracy and response times (RTs)..12

Figure 5: One individual's accuracy scores compared to response times under separate S/N ratios…………………………………………………………………………...13

Figure 6: Mean accuracy scores for sentence recognition………………………………23

Figure 7: Three graphs discussing values associated with the word recognition task…...26

List of Tables

Table 1: Truth Table……………………………………………………………………..11

Table 2: Accuracy Scores and Race Model Predictions…………………………………15

Table 3: Mean Accuracy Scores Across Modalities…………………………………......24

Table 4: Average C(t) Scores and Standard Deviations (SDs)…………………………..27

Table 5: Mean Scores for C_I(t) and C(t)………………………………………………..28

Thesis Abstract—Idaho State University (2014)

Most normal hearing people primarily rely on auditory information during conversational speech. Research has demonstrated that information from other modalities contributes to spoken language comprehension (Summerfield, 1987). Traditional measures for quantifying the contribution of visual speech cues during conversational speech have utilized accuracy. In a real-world conversational experience this may not be the most effective measure alone. Including accuracy with response time (RT) measures allows for a more representative real-world integration skill snapshot (Altieri & Townsend, 2012). Altieri, Townsend and Wenger (2014) recently developed a capacity measure including accuracy along with RTs to more accurately measure the gains achieved during audiovisual speech perception versus unimodal situations. The purpose of this thesis was to obtain a normative sample of audiovisual speech integration skills using this new measure of capacity alongside traditional audiovisual gain measures using accuracy (e.g., Grant, Walden, & Seitz, 1998; Sumby & Pollack, 1954).

Chapter 1: Introduction

Audiovisual speech perception involves responses to auditory and visual signal input, and interaction among these signals (Gelfand, 2009). The McGurk Effect is one classical illustration of how visual information can affect speech perception. The McGurk effect consists of a perceptual incongruency that occurs when auditory and visual speech signals are semantically mismatched or incongruent. For example, the McGurk effect is known to occur when a fusion of the auditory signal of /ba/ is combined with a visual signal [ga] yielding a perceptual conglomeration of the two signals which is perceived as an illusory "da" or "tha" (McGurk & MacDonald, 1976).

More germane to this study, Stein, Stanford, Ramachandran, Perrault, and Rowland (2009) discussed audiovisual speech integration in regards to the principal of *inverse effectiveness*. The principal of "inverse effectiveness" states that responses to multisensory signals *increase* relative to unisensory (i.e., auditory or visual-only) responses, as the saliency of the unisensory signals *decrease*. Significantly, this general principle applies to audiovisual speech perception. For speech perception, Sumby and Pollack (1954) observed in their seminal paper that visual speech cues can significantly enhance the identification of auditory speech. Their research demonstrated that being able to see a talker's face, especially in difficult listening conditions, can yield the equivalent of up to a 15 dB gain in the auditory domain (see also Erber, 1969). Akin to the principle of inverse effectiveness, the greatest visual benefit tends to occur when the auditory signal is substantially degraded by noise.

Processing Models

In a review of several major accounts of multisensory integration during speech processing, Summerfield (1987) emphasized that information about place of articulation during phonation is obtained via the visual modality (e.g. bilabial). Place of articulation is critical for understanding consonants or the higher frequency sounds, especially when the auditory signal is compromised by noise or hearing loss. Second, Summerfield (1987) discussed the measurement and the values of independent auditory and visual parameters, noting that; third, he discussed the filter function of the vocal tract. The vocal tract filters the raw sound made by one's vocal folds, molding the signal based on the space available through which the sound may travel before it leaves the body. This is primarily how the audio signal heard during speech is created. These roughly-hewn open sounds are mostly vowels and account for the majority of the acoustic energy and much of the information gleaned by the listener during speech perception. Lastly, Summerfield (1987) discussed articulatory dynamics of the vocal tract structures. These structures include the oral and nasal cavities, tongue, lips, etc. all of which assist in filtering the vocal fold's raw sound into the speech sounds a listener would hear, and in the case of many consonants.

Various neuro-cognitive models of how multisensory information interacts in the brain have been described in recent years. For example, one emerging proposal is that the auditory and visual information is translated into a common code prior to the integration of the two inputs (Rosenblum, 2005). Based on a review of audiovisual speech literature, Rosenblum (2005) argued that the information sharing involved in audiovisual speech perception is integrated in the earliest stages of perception; before word recognition even happens. The evidence gleaned by Rosenblum's research supported his hypothesis that audiovisual integration during speech perception occurred during the early stages; this included neurophysiological evidence –including brain mapping techniques, previous studies by Bernstien and Summerfield, and even infant studies (Rosenblum, 2005, p. 54-55).

Another approach described by Arnold, Tear, Schindel, and Rosebloom (2010) combined the existing models: a probability summation model and a model that assumed that auditory and visual cues are encoded as a unitary psychological process (linear summation model), and tried to fit them to audiovisual processing in speech perception. Probability summation can occur when redundant information is encoded by separate sensory systems. A decision is made when either system (e.g. auditory or visual) exceeds the requisite sensory threshold leading to identification of a precept. This model does not necessitate sensory integration; rather, it requires two independent sensory systems. The difference of the linear summation model is that the sensory estimate is subject to only one source of neural stimulation, as opposed to at least two utilizing the probability summation model. The authors found that the latter (linear summation) model was a more accurate fit by utilizing audiovisual signals and comparing results to auditory only and visual only accuracy scores. Findings such as these indicate that a decision process incorporates both modalities of information (audio and visual) and integrates them early into a unitary sensory "code" before a decision is reached.

Contrary to Rosenblum (2005), Bernstein (2005) also cited several studies that indicate that integration may occur at the later processing stages. Bernstein proposed that auditory and visual information are processed simultaneously and in separate pathways citing that, "a single phonetic processing area that is independent of sensory modality appears not to have been implemented in the speech perceiving brain" (Bernstien, 2005, p. 79). This proposition, along with supporting evidence from his research indicate that integration does occur but at the later processing stages and not necessarily in one "processing area," but perhaps in various areas. For example when studying the McGurk effect, large onset asynchronies between the audio and visual modalities failed to abolish the effect (Massaro, Cohen, & Smeele, 1996). Another study showed that the McGurk effect varies in strength depending on the culture of the listeners (Sekiyama & Tohkura, 1993). Finally the McGurk effect can be reduced in strength as a result of training (Massaro, 1987). Studies such as these suggest that the auditory and visual inputs can be attended to separately—at least to some degree—during recognition.

Empirical evidence. To address these issues, Altieri and Townsend (2011) examined questions related to audiovisual processing architecture described by Rosenblum, Bernstein, and Massaro. As described above, previous literature outlined two competing models of integration (e.g., Bernstein; 2005; Rosenblum, 2005; Massaro, 2004). A visual representation of these models can be seen in Figure 1. One proposed model is a "co-active model" which is similar to the "early integration" model proposed by Rosenblum (2005) and two, the convergent model discussed by Massaro (2004). This co-activation framework assumes that all inputs are translated into a common code while undergoing simultaneous processing. Second, the parallel model, or "late integration" model, is similar to the non-convergent model described by Massaro (2004) and suggests that both auditory and visual signals are processed independently and subsequently integrated prior to an "and/or" decision being reached (Townsend & Wenger, 2004). An "and/or" decision assumes two processing channels and either both finish processing together (and) or one finishes first (or) before a decision about what is seen, heard, or both can be reached. For example, suppose a listener is presented with an auditory /b/ and a visual "b" and needs to make a decision of what is perceived. Using a parallel processing model, the information would be processed by separate channels and integrated just before a decision is reached (or). The convergent model would conversely "integrate" the information into a common code, or singular pathway, and be processed as a singular information strand (and).

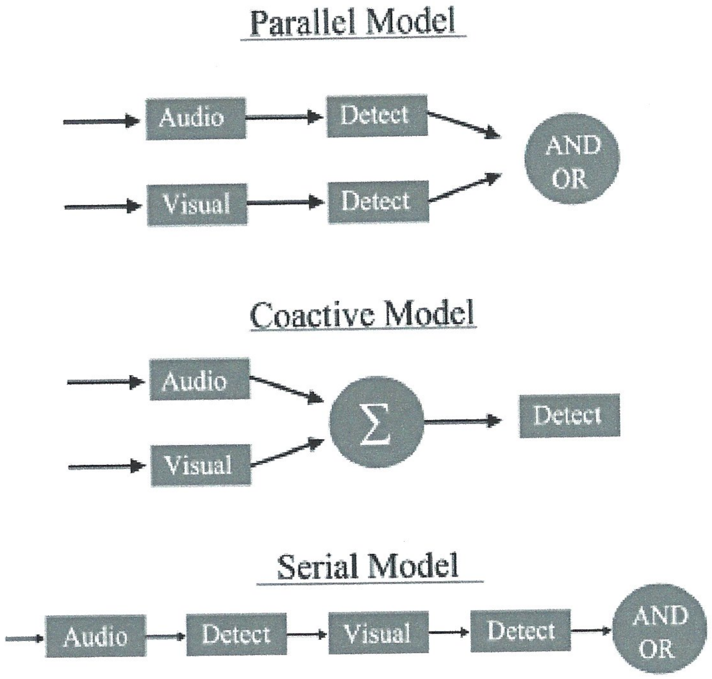


Figure 1. A visual representation of the proposed integration models: parallel "late integration" model, coactive "early integration" model and serial model (This figure appeared in Altieri & Townsend 2011; and Townsend & Nozawa, 1995).

To assess these hypotheses using reaction time modeling approaches, Altieri and Townsend (2011) carried out two experiments to assess whether parallel or coactive architecture best described audiovisual integration phenomena. Altieri and Townsend outline the various processing models and utilized a decision stopping rule, which determines whether all items or channels must finish processing before system termination and a decision is made. Their first experiment was designed to investigate processing architecture and decision rule with a discrimination task using eight monosyllabic words (see methodology section) along with five different signal-to-noise ratios (-30, -24, -18, -12, and -6 dB). Both reaction time (RT) and accuracy scores were recorded for audiovisual, auditory only and visual only presentations.

Altieri and Townsend (2011) argued, based on the speed at which their participants were able to maintain accuracy in their responses, that the parallel processing model was the most accurate in describing how the cognitive system processes audiovisual speech information. This conclusion was reached utilizing the Double Factorial Paradigm (DFP) approach (for more details about this approach see Altieri & Townsend, 2011). This approach essentially allowed them to empirically distinguish between a coactive model and a parallel model as theories of integration. They used mathematical formulation along with behavioral data to test these two models. The same methodology was also used to test architecture (parallel vs. coactive), decision stopping rule, as well as to assess workload capacity or integration efficiency. The basic design of the DFP involves identification of targets, usually presented in one or more channels (e.g. a visual signal and an auditory signal). These stimuli would then be presented by themselves in separate trials, redundant trials and absent trials, where only a blank screen would be presented. An "OR" design response mapping would require a "yes" response when a single target trial is presented; be it auditory or visual, or when a redundant trial with both signals is presented, and a "no" response on target-absent trials. The "OR" decision is reached when one stimulus *or* the other is processed so as to reach a decision on what is being presented. While response times (RT) and accuracy are utilized with the DFP to assess capacity, RT's are the crucial measure with this paradigm. As we shall see in the following section, measurement of capacity, including response times and accuracy, is essential to be able to compare individuals.

Measures of Audiovisual Integration

Development and application of a dynamic measure of audiovisual speech perception is critical to be able to use this information regarding visual benefit in clinical practice. Therefore, understanding the processing and integration of auditory, visual, and audiovisual signals is the first step in development of this measure. Once a measure of audiovisual integration is successful, a normative sample utilizing said measure would be needed against which comparisons for individuals could be made. Understanding how human sensory systems and the brain process audiovisual information is crucial to the development of a method to measure these processes.

Capacity analyses: reaction time measure. RTs have been the traditional measure of capacity throughout the years; however, inclusion of accuracy scores can give researchers and therapists a more accurate measure of conversational integration skills. Investigating communication between the auditory and visual pathways using model theoretic tools is crucial for understanding integration processes. One model theoretical tool for calculating integration efficiency involves comparing audiovisual RTs to parallel independent model predictions (e.g., Altieri & Townsend, 2011) derived from auditory and visual-only experimental conditions. This measure is known as capacity, and it was introduced by Townsend and Nozawa (1995). Specifically, capacity measures how the number of working channels affects processing efficiency in the processing time domain. It measures whether there is a cost, benefit, or no change in processing efficiency when multiple input channels are present relative to the conditions when only one channel is present. This instantiation of capacity implemented by Altieri and Townsend (2011) is shown in Figure 2 assumes an "OR" stopping rule. In other words, when information is presented via separate modalities (auditory and visual) a decision is reached when one *or* the other modality is recognized. This is opposed to an "AND" stopping rule which denotes that *both* modalities must be recognized before processing is halted and a decision is made (Townsend & Wenger, 2004). Information obtained from response times (RT) and accuracy gives researchers an idea of the efficiency associated with the processing task; specifically RTs measured in the presence of a unimodal signal as compared with bimodal signals (Altieri & Townsend, 2011; Townsend & Altieri, 2012). With a parallel independent model one would expect singular versus dual to be at least as fast as one another (unlimited capacity) with perhaps dual being faster (super capacity). Super capacity may be due to mutually beneficial interactions between input channels or from coactive processing (Townsend & Nozawa, 1995).

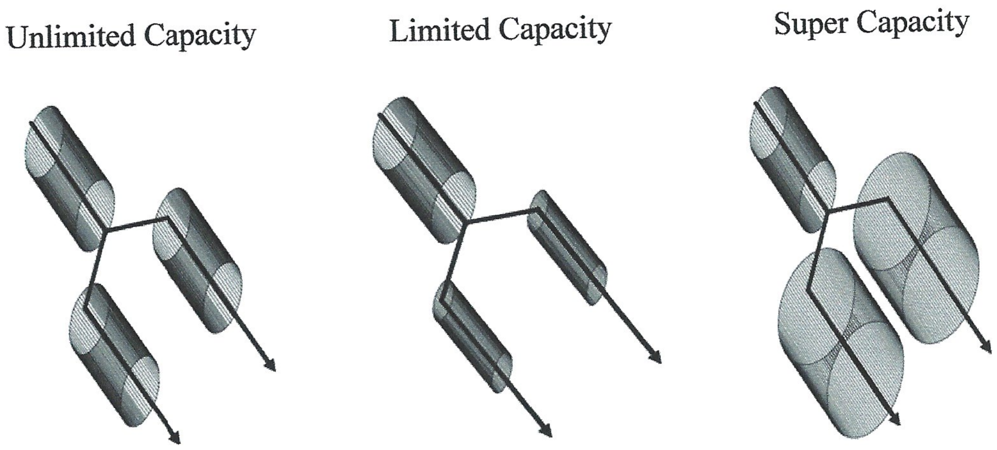
The methodology used in the present study to calculate RT capacity measures (during the single word identification task) are those described by Townsend and Altieri (2012). RT measures are used to assess the efficiency of having two channels (audiovisual) present as opposed to just one (auditory or visual) by comparing the audiovisual RT distributions to the sum of the auditory only and visual only distributions. A hazard function is the rate of probability of decision reached (H) is changing at time (t). Integrated hazard functions (H(t)) are calculated for the RTs from audiovisual, auditory only and visual only trials (Figure 2). The capacity coefficient (C(t)) allows comparison of processing times from trials to independent, parallel, race model predictions as a benchmark. This function can be interpreted as cumulative amount of work done, or expanded energy (C) by time (*t*) in each stimulus condition (auditory, visual, and audiovisual). This *capacity coefficient* is then used to compare total amount of work done (H) during the audiovisual trials (H_{AV}) (as compared to the auditory only (H_A) and visual only (H_V) trials): H_{AV} (t) = - Log{1 – F_{AV}(t)}, where F(t) denotes the cumulative distribution function (Altieri, Townsend & Wenger, 2014).

$$C(t) = \frac{H_{AV}(t)}{H_A(t) + H_V(t)}$$

Figure 2. Mathematical representation of traditional capacity (C(t)) measured by response time (*t*) for auditory (H_A), visual (H_V), and audiovisual (H_{AV}) signals.

C(t) is a ratio with three possible outcomes: first, C(t) < 1 indicates *limited capacity* consistent with parallel models with inhibitory cross talk between auditory and visual channels which is considered inefficient audiovisual integration. The second outcome would be C(t) = 1 which indicates *unlimited capacity* consistent with parallel unlimited independent models which assumes that integration is *not* occurring. The third outcome would be C(t) > 1 indicates *super capacity* consistent with parallel models with faciliatory cross-talk between the auditory and visual modalities which is indicative of efficient audiovisual integration.

Figure 3 is a visual representation of these capacity systems. The upper tube in each picture represents a single modality (e.g. visual or auditory). The lower tubes in each picture represent how the system will manage given a dual input. Theoretically the unlimited capacity processes at the same speed for bimodal input as unimodal (i.e. the lower tubes are the same diameter as the upper tube). The limited capacity system will process at a slower speed given bimodal input as opposed to unimodal (i.e. the lower tubes have a smaller diameter than the upper tube. The super capacity system operates faster given bimodal information (i.e. the lower tubes have a larger diameter than the upper tube).



Unlimited Capacity

Limited Capacity

Super Capacity

Figure 3. A visual representation of unlimited, limited and super capacity systems. The upper tube on all three figures denotes the presence of one input modality. The lower tubes indicate how the system will manage given two modalities. The more narrow the less processing power is available or used.

The capacity coefficient presented above assumes that the RTs used are all observed during trials in which a correct response was generated. This measure, therefore, excludes information regarding response accuracy. For application reasons it would be beneficial to have a measure that takes into account both response time and accuracy.

Capacity using both RT and accuracy. Table 1 is a truth table relating to the accuracy values assuming an "OR" rule. In this table, the left column indicates correct/incorrect recognition for the auditory modality. The second column shows the same for the visual modality. The third column (labeled "Winner") indicates whether the recognition occurred first in either modality. Lastly, the fourth column (labeled "Accuracy") indicates correct or incorrect recognition accuracy based on the information in the other columns (Altieri et al., 2014). This can be utilized along with the C_I(t) function to illustrate integration possibilities and probable processing models. (i.e. race model predictions). Race model predictions are mathematically calculated average results from a hypothetical "normal" distribution.

Table 1

Truth Table

Auditory	**Visual**	**Winner**	**Accuracy**
Correct	Correct	Auditory	Correct
Correct	Correct	Visual	Correct
Correct	Incorrect	Auditory	Correct
Correct	Incorrect	Visual	Incorrect
Incorrect	Correct	Auditory	Incorrect
Incorrect	Correct	Visual	Correct
Incorrect	Incorrect	A/V	Incorrect

Note. Truth table outlining accuracy scores assuming an "OR" stopping rule.

$$C\_I(t)=\frac{\log\left[\begin{aligned}&\int_0^t P_A(T_{AC}=t'<T_{AI})dt' * P_V(I)+\int_0^t P_V(T_{VC}=t'<T_{VI})dt' * P_A(I)\\&+\int_0^t P_A(T_{AC}=t'<T_{AI})dt' * \int_{t'=t}^{\infty} P_V(T_{VC}=t'<T_{VI})dt' + \int_0^t P_V(T_{VC}=t'<T_{VI})dt' * \int_{t'=t}^{\infty} P_A(T_{AC}=t'<T_{AI})dt'\\&+\int_0^t P_A(T_{AC}=t'<T_{AI})dt' * \int_0^t P_V(T_{VC}=t'<T_{VI})dt'\end{aligned}\right]}{\log\left[\int_0^t P_{AV}(T_{AVC}=t'<T_{AVI})dt'\right]}$$

Figure 4. Capacity measure including accuracy values (numerator) for auditory only trials (summation with subscripts of A), visual only trials (summation with subscript of V) and audiovisual trials (A and V summations added together) along with RTs to calculate capacity (C_I(t)) as compared to race model predictions (denominator) (Altieri, Townsend & Wenger 2014).

Figure 4 shows the measure which includes both RTs and accuracy to calculate capacity. The numerator factors in model predictions from auditory and visual trials. In the following part of the equation, $\int_0^t P_A(T_{AC} = t' < T_{AI})dt' * P_V(I)$, the subscript "A" represents auditory *only* trials and in the subscript "V" represents visual *only* trials. The subscript "t" found throughout denotes time; taking into account the speed along with accuracy. The subscript 'AV' in the quantity in the denominator represents audiovisual stimuli—auditory and visual trials presented together. The denominator allows for comparison to race model predictions.

Altieri et al. (2014) described results for an example participant to display the difference and utility of the two measures C(t) and C_I(t). The participant volunteered in a speeded word recognition task along with an adjustment of three different signal-to-noise (S/N) ratios; their C(t) and C_I(t) results are shown in Figure 5. S/N ratios allow for simulation of signal degradation that can happen in natural conversational environments. The results for the C_I(t) are displayed in the right panel. The layout allows for analysis of efficiency in terms of RT and accuracy separately. The individual points indicate the C(t) and C_I(t) value across time for the three different S/N ratios. The results indicate that as the signal becomes increasingly degraded the efficiency, in terms of speed and accuracy, improved for the audiovisual condition. For the "clear" auditory condition, the results showed, not surprisingly, that visual information did not have any effect on either speed or accuracy. This is evidenced by the fact that the C(t) and C_I(t) scores were less than 1 for a large range of RTs.

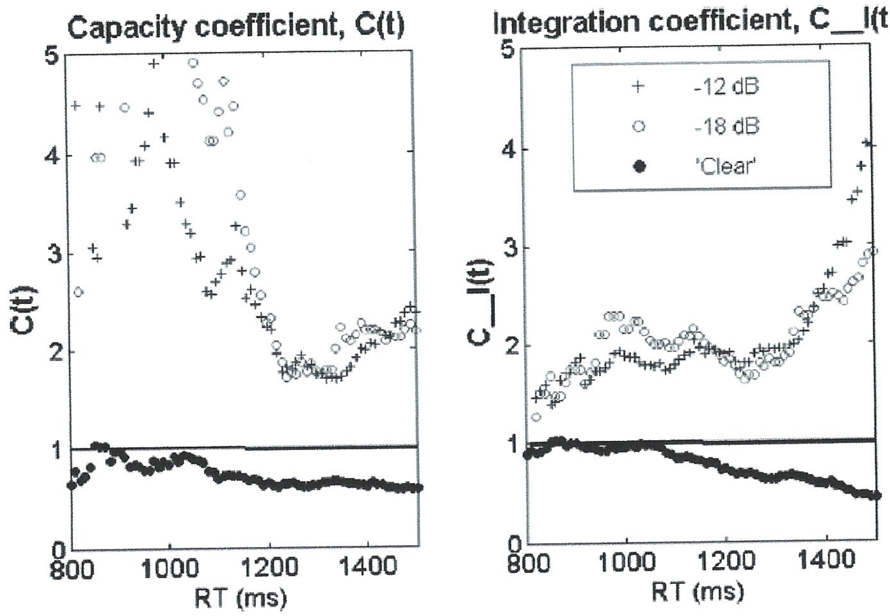


Figure 5. Example of one individual's accuracy scores compared to response times under separate S/N ratios (This figure appeared in Altieri, Townsend, & Wenger 2014).

Calculations including accuracy measures and response times (RT) utilize the (capacity) integration assessment measure, C_I(t) (Altieri et al., 2014). This is utilized to incorporate both speed and accuracy into a capacity measure using modified F(t) distribution functions. Similar to the RT-only measure, this capacity coefficient can be compared to independent, parallel, race model predictions as a benchmark. The logic for including RTs and accuracy in a singular measure will be reviewed here. The underlying theory and details can be found in Townsend and Altieri (2012).

Figure 5 displays a modified capacity measure that uses both speed and accuracy and is designed to display a snapshot of the listeners true integration skills. The RT-only capacity measure from Townsend and Nozawa (1995; C(t)) only measures benefit from multimodal input as determined by processing speed. For example, to illustrate the benefit of using both C(t) and C_I(t), imagine a normal-hearing listener who displays ceiling level audiovisual performance in accuracy, but shows only slightly limited benefit from multiple signals in terms of C(t). Utilizing C_I(t) for this individual may still show "good" integration skills because their accuracy scores are just omitted from the RT-only measure of C(t). C_I(t) then becomes more beneficial for listeners with poor audio-only or visual-only skills, hearing loss, or listeners in a difficult listening environment. C_I(t) too can be useful when speed-accuracy tradeoffs occur; imagine a listener who is very fast on audiovisual trials as compared with unisensory conditions but their accuracy becomes lower than predicted.

Next, Table 2 from Altieri, Townsend and Wenger (2014) shows accuracy scores from each condition (auditory only and AV) and S/N ratio, along with race model predictions. For the "clear" S/N ratio the obtained AV were very near to race model predictions; however for the degraded situations the obtained AV scores were higher than race model predictions for accuracy. Again race model predictions are mathematically calculated average results from a hypothetical "normal" distribution.

Table 2

Accuracy Scores and Race Model Predictions

S/N Ratio	Auditory Only	Obtained AV	Race Model
Clear	.98	.98	.99
-12 dB	.69	.95*	.90
-18 dB	.36	.90*	.80

Note. Auditory only, AV and race model predictions for the separate S/N ratios. [a]The "*" indicates accuracy scores that were greater than model predictions (Altieri, Townsend & Wenger, 2014).

The RT results, C(t), in the left panel of Figure 5, during the "clear" S/N ratio indicate inefficient integration skills. This is due to the fact that the listener, similar to other normal hearing individuals, did not benefit from visual information under optimal listening conditions. These results are consistent with the C_I(t) integration coefficient panel on the right. The decrement observed in both panels approximating race model predictions, results from a processing slowdown during less-than-optimal listening conditions.

For the more degraded S/N ratios (-12 and -18 dB) processing speed regularly violated the race model predictions. In fact, as the signal became more degraded, integration efficiency increased. These violations of race model predictions showed that increased audiovisual integration efficiency, during the deteriorated listening conditions, was displayed through enhancement of both speed and accuracy. Regardless, C_I(t) results did differ from C(t) results in significant ways. Both panels in Figure 5 suggest super capacity for the faster RTs. The difference is observed in the C_I(t) panel; the range of scores is high, but lower than the C(t) (~2 as opposed to > 5 respectively). This indicates that accuracy moderated integration efficiency when it is taken into account using the C_I(t) function. These differences are important for displaying the variation of integration skills as measured by these two functions. These data show a normal hearing listener who displays efficient integration in the RT domain, but less efficient integration as measured by C_I(t) due to a suboptimal gain in accuracy. This can be compared to a hearing-impaired individual with poor auditory skills who may show significantly higher gain in the accuracy domain. Ultimately the utilization of the C_I(t) function can bring to light valuable information through the use of accuracy measures that RT measures alone cannot. These measures together compose a set of comprehensive tools to measure integration. To be clinically useful, however, data is needed on how people, on average, integrate information using these measures.

Obtaining Normative Measures

Normative data are useful for characterizing the distribution of what is considered "normal" in a certain population. Normative data collection requires specific details concerning the methodology for collection and the definition of the target population (O'Connor, 1990). Increased knowledge about the role of audiovisual information during speech perception will allow professionals to better assess, diagnose, and treat individuals with any deficit in either the auditory or visual sensory areas. Integration skills may differ across individuals but this information should be predictive of receptive communication skills. For example, two individuals can have similar auditory-only and visual-only accuracy scores but display substantially different scores for audiovisual accuracy and speed. Treatment strategies should therefore vary according to any difference observed in any given individual. Normative data sampling is the next logical step for comparison of a hearing-impaired population against same-age normal-hearing peers. Normative data also allows for new methodologies to be tested for validation purposes by sampling across a normal population.

To accurately accrue a normative sample using Altieri and colleagues (2014) methodology, careful attention was given to their methodological approach, and reproduction of those conditions for the sample. A normative data sample for audiovisual integration in speech perception allows for comparison of scores from individuals with suspected impairment in either area to be compared with a normal group and allow for more appropriate intervention techniques and aids to be utilized and calibrated respectively. The major measurement used to obtain normative data on integration ability was the new capacity assessment measure that utilizes both RT, and accuracy. Utilization of RTs and accuracy is important because RT and accuracy have not been used before Altieri and colleagues proposed their measure. It is this measure that is implemented into this study's data collection to ascertain a normative data sample against which hearing-impaired individuals can be compared. This paper is focused on obtaining a sample from 76 subjects, calculating a capacity score, obtaining a peak capacity score, and comparing them to sentence recognition accuracy scores.

This study used Altieri and colleagues' (2014) methodology for calculating capacity from word recognition tasks and response times from a sample of adults ranging in age from 20 years to 75 years. This data includes RT's and accuracy measures and an open set sentence recognition task. These results are compared to further understand integration skills based on the capacity coefficient measure versus sentence recognition. Ideally a positive correlation should be present between these two data sets. Hearing threshold measures are important for comparison of individuals who display impairment against a truly normative sample set. The future goal is to implement this new capacity measure of integration, compare it to previous accuracy measures and sentence recognition data, and assess how integration changes as a function of audiometric configuration.

Chapter 2: Experimental Methods

Participants

Participants included 76 adults obtained from the Idaho State University campus and the Pocatello, Idaho community. All participants reportedly had normal or corrected vision, normal-hearing, and were reported to be native speakers of American English. Each participant was paid an institutional review board (IRB) approved $10. Volunteers participated in all three tasks.

Audiometric Testing

Participants of both experiments (i.e. word recognition and sentence processing) first had their hearing thresholds taken separately for each ear at 250, 500, 1000, 2000, 4000, and 8000 Hz. For each frequency, thresholds were obtained using a continuous tone for approximately 1000ms. This was done utilizing standard audiometric procedures to obtain threshold. For example when a listener identified a tone correctly the sound level was reduced by 10dB and on an incorrect response and increase of 5dB was made until a threshold was determined for all frequencies in both ears. A Benson Medical NEXT stand-alone audiometer was used to obtain hearing thresholds.

Single Word Recognition Task

Materials. Word recognition stimulus materials included audiovisual movie clips of two different female talkers from the Hoosier Multi-Talker Database (Sherffert, Lachs, & Hernandez, 1997). For the word recognition task two sets of English, monosyllabic words were used utilizing two female talkers and included: “mouse,” “job,” “gain,” “tile,” “shop,” “boat,” “date,” and “page.”

Audio, visual, and audiovisual files were edited and presented using E-prime software version 2.0. The audio files were sampled at a rate of 48 kHz at a size of 16 bits. The duration of the auditory, visual, and audiovisual files ranged from approximately 800 to 1000 ms. The stimuli were selected and edited in such a way as to minimize differences between onset of facial movement and vocalization between clips. To avoid ceiling performance, the audio stimuli were degraded using an 8-channel cochlear implant (CI) simulator (sinewave vocoded) from TigerCIS version 1.08.01.

Procedure. For the word recognition task, participants were seated 14" to 18" in front of a Dell computer equipped with Beyer Dynamic-100 headphones. Each trial began with a "fixation" dot in the center of the screen to indicate that a new trial will begin. Participants then began the trial by pressing the space bar on the keyboard. The trial stimuli included auditory-only, visual-only or audiovisual stimuli, which were presented in different blocks. A standard keyboard was used —the numbers -1 through 8- were labeled with the words. This was the same methodology used by Altieri et al., 2014. The word-number associations did not change over the course of the task (i.e. 1 was always "mouse," etc.). The participants were then instructed to respond as quickly and accurately as possible by pressing the corresponding number on the keyboard. Reaction times were measured from stimulus onset. On auditory-only trails participants were required to base their response on auditory information, and on visual-only trials participants were required to lip-read. Participants received 25 practice trials at the onset of the task that were not included in the data analysis. The stimuli presented in the word recognition task were consistent with stimuli used by Altieri and Townsend (2011) and Altieri et al. (2014).

Sentence Processing Task

Materials. Sentence stimuli consisted of 25 sentences obtained from a database of pre-recorded audiovisual City University of New York (CUNY), English, sentences spoken by a female talker (Boothroyd, Hnath-Chisolm, Hanin, & Kishon-Rabin, 1988). The set of 25 sentences were subdivided into the following word lengths: 3, 5, 7, 9, and 11 words with five sentences for each length. This was done because sentence length naturally varies in everyday conversation. Sentences were presented randomly for each participant and no cues were provided in regard to sentence length or semantic content. The sentence materials are shown in the Appendix A.

Procedure. As with the word task participants were seated 14" to 18" in front of a Dell computer equipped with Beyer Dynamic-100 headphones. Each trial began with a "fixation" dot in the center of the screen to indicate that a new trial will begin. Participants then began the trial by pressing the space bar on the keyboard. The trial stimuli included auditory-only, visual-only or audiovisual stimuli, which were presented in different blocks. Immediately after the presentation of the stimulus sentence a dialog box appeared in the center of the screen instructing the participant to type in the words they thought the talker said by using a keyboard. Each sentence was given to the participant only once. No feedback was provided on any of the test trials.

Scoring for the sentence task was carried out in the following manner: if the participant correctly typed a word in the sentence, then that word was scored as "correct." The proportion of words correct was scored across sentences. For example, for the sentence "Is your sister in school," if the participant typed "Is the…" only the word "is" would be scored as correct making the proportion correct = 1/5 = 0.20. Word order was not a critical criterion for a word to be scored as accurate. The sentence stimuli and procedures presented in this task were consistent with those used by Altieri, Pisoni and Townsend (2011). Completion of the two tasks took, on average, approximately one hour.

Chapter 3: Results

Sentence Processing Task

This experiment was designed to simulate a more conversational based measure utilizing our two capacity measures for RT and accuracy. Figure 6 represents mean accuracy scores for the various modalities and the gain that individuals received from having the visual component available during the audiovisual presentations as compared to the auditory only presentations.

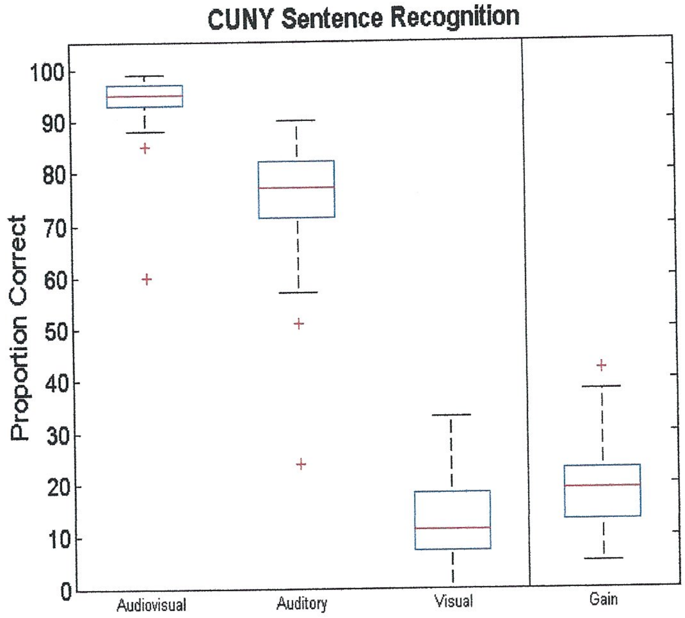


Figure 6. Mean accuracy scores for sentence recognition represented for audiovisual situation, auditory only, and visual only. The far left column indicates the gain that participants received from the visual signal (as opposed to auditory only) when comparing the audiovisual and auditory only situations. The small red crosses above and below the groupings indicate individual outliers.

Table 3 displays mean accuracy and standard deviations across auditory, visual-only, and audiovisual trials. It also provides mean gain score showing the gain individuals received by having the visual component during audiovisual trials as compared to the maximum of the auditory and visual-only accuracy scores. Upon observation, all the gain scores from having the visual modalities are positive, indicating that even with normal hearing individuals visual information aids in understanding as opposed to just the auditory signal. This evidence supports Altieri and Townsend (2011) that revealed that further degradation of the auditory signal results in higher gains by having the visual modality present.

Table 3

Mean Accuracy Scores Across Modalities

	AV	A-only	V-only	AV Gain
Mean	94.40	75.60	12.60	**18.80**
SD	4.96	10.50	7.04	**8.33**
Mean + 1.5*SD	100.00	91.30	23.20	**31.30**
Mean – 1.5*SD	87.00	59.90	2.03	**6.30**

Note. Mean accuracy and standard deviations (SD) across modalities. The far left column indicates gain scores associated with these measures (i.e. mean, SD, etc.).

Single Word Recognition Task

This experiment was designed to measure reaction times (RT) utilizing the C(t) equation independent from accuracy scores as a comparison tool to correlate C_I(t) scores with previous research, which only utilized RTs as a measure. Figure 7 displays three graphs associated with the word recognition task. Outliers can be noted as the red crosses above and/or below the main body for each modality. Figure 7 (A) and (B) indicate capacity based on RT (C(t)) and accuracy (C_I(t)) respectively. Figure 7 (C) represents this comparison showing mean capacity for the word recognition task and compares these scores to traditional RT only measures. Figure 7(A) shows a general trend that RTs for audiovisual and auditory modalities are fairly consistent, while the visual-only modality shows a higher response time value; this may be due to the fact that the participants were all normal or near-normal hearing individuals and typically rely more on auditory information when processing speech (Altieri & Townsend, 2011). Similarly, in Figure 7(B), accuracy scores for the audiovisual and auditory-only modalities are high and consistent (~98%) while visual-only scores are lower (~75%). In Figure 7(C) capacity scores for the C_I(t) function appear more tightly knit than the C(t) grouping. While the means remain similar the grouping and the outliers for the RT-only measure are less consistent than those which include accuracy. These data show that inclusion of RTs and accuracy (as opposed to just RTs alone) allows for a more accurate picture of integration skills for this task. One outlier does appear for the C_I(t) function in Figure 7(C) indicating that the individual is poor at integrating when both response time and accuracy measures are taken, but it seems the exception that proves the rule in this case.

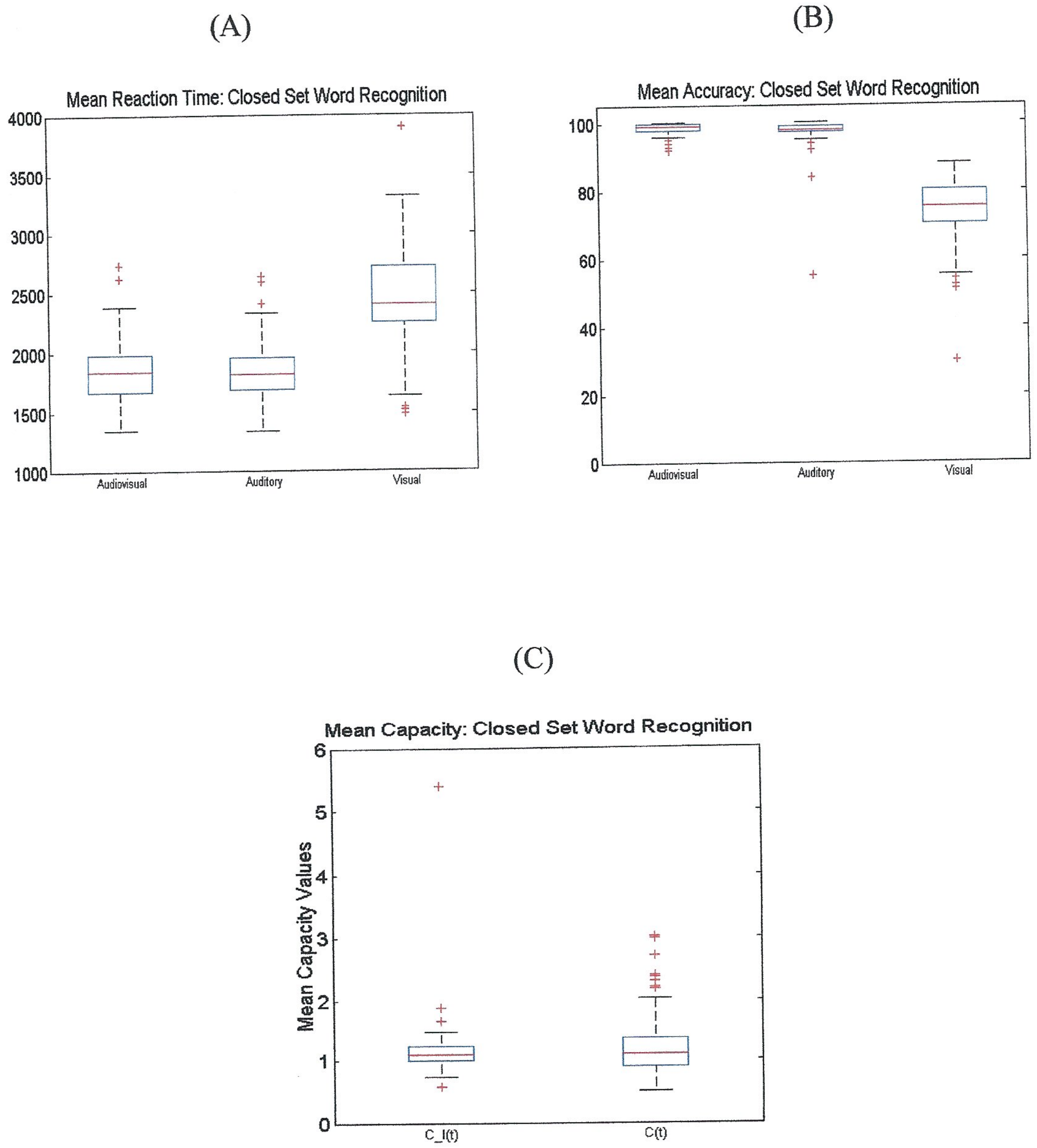


Figure 7. Display of graphs discussing values associated with the word recognition task. (A): displays mean reaction times for the various modalities. (B): shows mean accuracy scores for all three modalities. (C): shows mean capacity values for both measures, C(t) and C_I(t). Red crosses above and below groupings are individual outliers.

It is important to utilize peak capacity values for individuals when comparing scores to a normative group. First, a peak score allows a single point of data to be used as a comparison to the group; this allows for less variability in the results of comparisons. Second, a peak score gives researchers and clinicians a summary measure of the capability of a single participant.

Table 4 shows the average auditory, visual, and audiovisual accuracy levels and standard deviations across the participants for the RT only (C(t)) equation. Consistent with predictions, the mean audiovisual recognition accuracy scores approximated ceiling levels at 98%. Similar mean scores for the auditory only situation, the third column, were observed at 97%. This is not surprising considering the sample subjects all reported normal or near-normal hearing. The visual-only modality mean scores are listed at ~73% for this population. Normal listeners use auditory information more readily than visual information, with the exception of when the audio signal is distorted significantly (Altieri & Townsend, 2011).

Table 4

Average C(t) Scores and SDs

	AV	A-only	V-only
Mean	98.00	97.00	73.10
SD	1.70	5.40	10.10
Mean + 1.5*SD	100.00	100.00	88.20
Mean – 1.5*SD	96.00	89.00	58.00

Note. Average RT-only (C(t)) scores for the various modalities.

Table 5 lists mean scores utilizing both equations with standard deviation (SD) scores demonstrating the ability for comparison of individual scores to that of this normative sample. In this table and experiment maximum capacity scores are used; these max scores provide a snapshot of what a participant is capable of doing; with auditory only, visual only and redundant information. Capacity, again, can be understood as limited, unlimited, and super capacity depending on the scores and where and how they fall in line with predictive models. A positive correlation between these two measures was observed. A gain was observed for both RTs and accuracy during audiovisual trials versus unimodal presentations across tasks. The correlation statistics, r = .27, p = .018, df = 74, indicate a significant positive correlation between gain scores and peak capacity values. However, a non-significant negative correlation was observed for C(t) scores.

Table 5

Mean Scores for C_I(t) and C(t)

	Max(C_I(t))	Max(C(t))	AV_RT	A_RT	V_RT
Mean	1.18	1.25	1851	1870	2472
SD	0.54	0.55	255	265	439
Mean + 1.5 SD	2.00	2.08	2234	2268	3132
Mean – 1.5 SD	0.37	0.43	1469	1472	1812

Note. Mean scores and SD comparisons for both equations.

Chapter 4: Discussion and Conclusion

This novel capacity approach for comprehensively assessing for audiovisual processing in speech perception is a critical tool for allowing greater understanding of integration of multimodal stimuli. This normative sample data set allows for comparison of individuals against a group of their normal-hearing peers to compare integration capability using accuracy, C(t), and the unified C_I(t) capacity-assessment measure (Altieri et al., 2014). This information can then perhaps be used clinically to adjust hearing aids, assist in treatment plans and formulation of treatment measures, etc. The C_I(t) function is a critical component to this sample; by utilizing both RTs and accuracy it yields a more accurate picture of integration abilities. Capacity measures should be used alongside accuracy only measures for integration ability.

The various tasks were designed to be useful clinically and valid ecologically. The word recognition task, although not very ecologically valid, may be of use to audiologists in a clinical manner through use of specific phoneme recognition, and as a good indicator of processing speed. The sentence recognition task yields more ecological validity as a representation of conversational speech integration skills. Processing speed cannot be measured for this task however, so an individual may display high integration scores for the audiovisual modality during this task but take an inordinate amount of time to process; which would be considered inefficient conversationally.

Notwithstanding the experimenter's drive to attain a representative sample of normal or near normal hearing adults against which to compare scores, perhaps various populations and/or locations could be sampled, utilizing this capacity measure to obtain a more representative sample of English speaking adults in general.

Future Applications

Future applications of this information can include comparisons to hearing impaired individuals, be it either high frequency or low frequency loss (Altieri & Hudock, in press). Also this information could be crucial to future applications for hearing aid-users. Utilizing this new capacity equation to measure integration skills will allow for more accurate and personal hearing aid adjustment.

The use of the C_I(t) function for measurement of integration skills may also allow for better clinical evaluations and treatment plans for individuals who have sensory decline; this can include auditory or visual. Knowledge of how efficiently an individual utilizes visual information during conversation (or even single word tasks) can allow for adjustments to be made in lifestyle, conversational partner interaction behaviors, or assistive devices. The more that is understood about audiovisual integration in speech perception will only allow for further advances for application.

References

Altieri, N., Pisoni, D.B., & Townsend, J.T. (2011, July). Some normative data on lip-reading skills [Letter to the editor]. *Journal of the Acoustical Society of America, 130*(1), 1-4.

Altieri, N., & Townsend, J.T. (2011). An assessment of behavioral dynamic information processing measures in audiovisual speech perception. *Frontiers in Psychology, 2*(238), 1-15. doi:10.3389/fpsyg.2011.00238

Altieri, N., Townsend, J.T., and Wenger, M.J. (2014). A measure for assessing the effects of audiovisual speech integration. *Behavior Research Methods, 46*(2), 406-15. doi:10.3758/s13428-013-0372-8.

Arnold, D.H., Tear, M., Schindel, R., & Rosebloom, W. (2010). Audiovisual speech cue combination. *PLoS ONE 5*, e10217. doi:10.1371/journal.pone.0010217

Bernstein, L.E. (2005). Phonetic perception by the speech perceiving brain. *The Handbook of Speech Perception*, 79-98, D.B. Pisoni, & R.E. Remez (Eds.). Malden, MA.

Blank, H., & von Kriegstein, K. (2013). Mechanisms of enhancing visual-speech recognition by prior auditory information. *NeuroImage, 65*, 109-118. doi:10.1016/j.neuroimage.2012.09.047

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

Conrey, B., & Pisoni, D.B. (2006). Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. *Journal o the Acoustical Society of America,* 119, 4065. doi:10.1121/1.2195091

E-prime (Version 2.0) [Computer software]. Sharpsburg, PA: Psychology Software Tools Inc.

Erber, N. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Sport and Health Research, 12*, 423-425.

Gelfand, S.A. (2009). *Essentials of audiology* (3rd ed.). New York, NY: Thieme Medical Publishers.

Goebel, R., & van Atteveldt, N. (2009) Multisensory functional magnetic resonance imaging: a future perspective. *Experimental Brain Research, 198*, 153-164. doi:10.1007/s00221-009-1881-7

Grant, K.W., & Seitz, P.F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *Journal of the Acoustical Society of America, 104*, 2438-2450.

Grant, K.W., Walden, B.E. & Seitz, P.F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America, 5*(1), 2677-90.

Massaro, D.W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Lawrence Erlbaum.

Massaro, D.W., Cohen, M.M., & Smeele, P.M. (1996). Perception of asynchronous and conflicting visual and auditory speech. *Journal of the Acoustical Society of America, 100,* 1777-86.

Massaro, D.W. (2004). From multisensory integration to talking heads and language learning. The Handbook of Multisensory Processes, 53-83, G.A. Calvert, C. Spence, & B.E. Stein (Eds.). Hillsdale, NJ.

McGrath, M., & Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *Journal of the Acoustical Society of America, 77*(2), 678-684.

McGurk, H., & MacDonald, .J.W. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748.

O'Connor, P.J., (1990) Normative data: Their definition, interpretation, and importance for primary care physicians. *Family Medicine Journal, 22*(4), 307-311.

Ostwald, D., Porcaro, C., Mayhew, S.D., & Bagshaw, A.P. (2012) EEG-fMRI Based information theoretic characterization of the human perceptual decision system. *PLoS ONE 7*(4). doi:10.1371/journal.pone.0033896

Rosenblum, L.D. (2005). Primacy of multimodal speech perception. *The Handbook of Speech Perception*, 51-78, D.B. Pisoni, & R.E. Remez (Eds.). Malden, MA.

Rosenblum. L.D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science, 17*(6), 405-409. doi:10.1111/j.1467-8721.2008.00615.x

Raij. T., Ahveninen, J., Lin, F.-H., Witzel, T., Jääskeläinen, I.P., Letham, B.,...Belliveau, J.W. (2010). Onset timing of cross-sensory activations and multisensory interactions in auditory and visual sensory cortices. *European Journal of Neuroscience, 30*(10), 1772-1782. doi:10.1111/j.1460-9568.2010.07213.x

Seikel, J.A., King, D.W., & Drumright, D.G. (2005). *Anatomy & physiology for speech, language, and hearing* (3rd ed.). Clifton Park, NY: Thomson Delmar Learning.

Sekiyama, K., & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics, 21*, 427-444.

Sherffert, S., Lachs, L., & Hernandez, L.R. (1997). The hoosier audiovisual multi-talker database. *Research on Spoken Language Processing Progress Report, 21*, Bloomington, IN.

Stein, B.E., Stanford, T.R., Ramachandran, R., Perrault, T.J.Jr., & Rowland, B.A. (2009). Challenges in quantifying multisensory integration: Alternative criteria, models, and inverse effectiveness. *Experimental Brain Research, 198*(2-3), 113-126. doi:10.1007/s00221-009-1880-8

Stumby, W.H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustic Society of America, 26*, 12-15.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. *The Psychology of Lip-Reading*, 3-50.

TigerCIS (Version 1.08.01) [Computer software]. Los Angeles, CA: Innovative Speech Software.

Townsend, J.T., & Altieri, N. (2012) An accuracy-response time capacity assessment function that measures performance against standard parallel predictions. *Psychological Review, 119*(3), 500-516. doi:10.1037/a0028448

Townsend, J.T., & Wenger, M.J. (2004). A theory of interactive parallel processing: New capacity measures and predictions for a response time inequality series. *Psychological Review, 111*(4), 1003-1035. doi:10.1037/0033-295X.111.4.1003

Appendix A

What will we make for dinner when our neighbors come over
Is your sister in school
Does your boss give you a bonus every year
Do not spend so much on new clothes
What is your recipe for cheesecake
Is your nephew having a birthday party next week
What is the humidity
Let the children stay up for Halloween
He plays the bass in a jazz band every Monday night
How long does it take to roast a turkey
Which team won
Take your vitamins every morning after breakfast
People who invest in stocks and bonds now take some risks
Those albums are very old
Aren’t dishwashers convenient
Is it snowing or raining right now
The school will be closed for Washington’s Birthday and Lincoln’s Birthday
Your check arrived by mail
Professional musicians must practice at least three hours everyday
Are whales mammals
Did the basketball game go into overtime
When he went to the dentist he had his teeth cleaned
We’ll plant roses this spring
I always mail in my loan payments on time
Sneakers are comfortable