Published in final edited form as:

Int J Audiol. 2016 April ; 55(4): 206–214. doi:10.3109/14992027.2015.1120895.

Normative Data on Audiovisual Speech Integration Using Sentence Recognition and Capacity Measures

Nicholas Altieri and Daniel Hudock

Department of Communication Sciences and Disorders, Idaho State University

Abstract

Objective—The ability to use visual speech cues and integrate them with auditory information is important, especially in noisy environments and for hearing-impaired (HI) listeners. Providing data on measures of integration skills that encompass accuracy and processing speed will benefit researchers and clinicians.

Design—The study consisted of two experiments: First, accuracy scores were obtained using CUNY sentences, and capacity measures that assessed reaction-time distributions were obtained from a monosyllabic word recognition task.

Study Sample—We report data on two measures of integration obtained from a sample comprised of 86 young and middle-age adult listeners:

Results—To summarize our results, capacity showed a positive correlation with accuracy measures of audiovisual benefit obtained from sentence recognition. More relevant, factor analysis indicated that a single-factor model captured audiovisual speech integration better than models containing more factors. Capacity exhibited strong loadings on the factor, while the accuracy-based measures from sentence recognition exhibited weaker loadings.

Conclusions—Results suggest that a listener's integration skills may be assessed optimally using a measure that incorporates both processing speed and accuracy.

Decades of research have consistently demonstrated that speech cues obtained from the visual modality can provide redundant and complementary information to the auditory modality; this is true even in normal-hearing listeners (e.g., Grant et al, 1998; Sumby & Pollack, 1954). The ancillary effect of visual speech across multiple auditory signal-to-noise ratios was first described by Sumby and Pollack (1954) in their seminal study: Most normal-hearing listeners receive the greatest benefit from visual cues in poor listening conditions. The benefits of being able to see a talker's face bring to bear issues concerning how effectively listeners "integrate" or utilize cues across different modalities to recognize speech. Aside from variability in integration efficiency among normal-hearing listeners (Altieri & Hudock, 2014a), factors such as ageing (e.g., Sommers et al., 2005), and varying degrees of high and low-frequency hearing-loss may adversely affect a listener's ability to benefit from visual speech cues (e.g., Altieri & Hudock, 2014b; Bergeson & Pisoni, 2004; Erber, 2003).

Normative measures on visual-only speech recognition have already been reported using a sample of eighty-four participants (Altieri et al., 2011). This study will go further by reporting normative data on audiovisual speech perception skills using a similar sample of adults. Two types of measures of audiovisual performance will be assessed: audiovisual benefit from open-set sentence recognition (e.g., Sommers et al., 2005), and a reaction-time (RT) measure known as “capacity” that will be assessed using a closed-set word identification task (Altieri et al., 2014). This study represents a continuation of Altieri and Hudock (2014a); the authors compared a subset of the capacity data reported in this study to the low and high-frequency pure-tone thresholds of each listener. These results suggest that a listener’s ability to integrate auditory and visual speech, measured using capacity, is negatively associated with auditory sensory function (see Erber, 2003).

Accuracy and Capacity Measures of Integration

Assessments of audiovisual integration have used techniques across neural and behavioral domains; these have been used to examine how effectively listeners can combine auditory and visual signals (e.g., Stevenson et al., 2014). Behavioral measures often include deviations from computational model predictions. Models such as the Fuzzy Logical Model of Perception (FLMP; Massaro, 2004) and the Pre-labeling Model of integration (PRE; Braida, 1991) use algorithms to obtain optimal audiovisual accuracy predictions. The predictions are derived from confusion matrices indicating error rates obtained from auditory and visual-only trials (Grant et al., 1998).

Other behavioral measures quantify audiovisual benefit using an approach that does not rely on model predictions: such approaches essentially compare audiovisual accuracy relative to baseline predictions obtained from auditory and visual-only scores using sentence or monosyllabic word recognition tasks (e.g., Sommers et al., 2005; Tye-Murray et al., 2007). For example, a recently modified measure, known as “capacity”, involves comparing distributions of RTs obtained from audiovisual trials to the baseline predictions of “independent race models” (Altieri & Townsnd, 2011; Townsend & Nozawa, 1995). This method assumes that separate sources of information—in this case auditory and visual cues—are processed independently (Townsend & Nozawa, 1995). The logic of using capacity as an integration measure goes as follows: RTs are obtained from trials in which both auditory and visual information are presented, as well as from trials where only auditory or visual information is available. Second, integrated hazard functions ($H(t)$) are empirically calculated from audiovisual, auditory, and visual-only trials (Altieri et al., 2014). These functions can be calculated by first obtaining a distribution of RTs and then computing the cumulative sum (i.e., $F(t)$ which sums to 1). The integrated hazard function can then be generated by taking a logarithmic transformation: $H(t) = -\log(1-F(t))$. Independent race model predictions are computed by summing the integrated hazard functions derived from auditory and visual-only trials. Capacity is equal to the $H(t)$ obtained from audiovisual trials divided by the sum of $H(t)$ s from auditory and visual-only trials, which constitutes

$$\text{independent predictions: Capacity}(C(t)) = \frac{H_{AV}(t)}{H_A(t) + H_V(t)}$$

Generally, if capacity is greater than 1, then audiovisual recognition is faster and more efficient than independent race model predictions (Townsend & Nozawa, 1995). This scenario is also referred to as “super-capacity” (“efficient integration” and “super-capacity” may be used interchangeably). If capacity equals 1 for some RT range, then it equals independent race model predictions: In these cases we can also say that capacity is “unlimited”. Finally, if capacity is less than 1, it is lower than independent predictions and we may refer to it as “limited capacity” or “inefficient integration”.

Capacity is ideal for use in speech perception studies. In order for audiovisual speech recognition to be considered “efficient” recognition must occur quickly relative to changes in perceptual workload (auditory-visual vs. uni-sensory). Importantly, capacity has recently been augmented to include the contribution of both RTs with corrections made for accuracy in one unitary measure (Altieri et al., 2014; Townsend & Altieri, 2012). Another important contribution of the capacity functions is that they avoid ceiling effects often seen in accuracy-based measures. This is especially true when accuracy is high in both audiovisual and auditory-only settings.

In addition to scores from sentence recognition, normative data on audiovisual integration skills will be provided by: $C(t)$ which is RT only (Altieri & Wenger, 2011; Townsend & Nozawa, 1995), and $C_I(t)$, which is the accuracy-corrected capacity measure incorporating both RT distributions and accuracy. This latter function is beneficial since integration skills rely on processing language quickly while also maintaining accuracy. This paper will furnish comprehensive RT and accuracy data establishing what constitutes a “good” or “poor” integrator relative to the adult population, show the extent to which capacity performance may be predictive of traditional accuracy measures of audiovisual benefit obtained from sentence recognition (e.g., Altieri & Hudock; 2014a; Tye-Murray et al., 2007), and utilize factor analysis to reveal which integration measure—capacity or audiovisual accuracy—best accounts for integration skills.

Methods

This study represents a continuation of Altieri and Hudock (2014a); consequently, the methodology in this study was identical. Methods for calculating capacity were furnished by Altieri and Hudock (2014a) in the supplementary materials and also by Altieri et al. (2014) who provided Matlab code.

In Experiments 1 and 2, the auditory signal was degraded using an eight-channel sinewave cochlear implant (CI) simulator (AngelSim: <http://www.tigerspeech.com/>), and presented at a comfortable level of ~ 68 dB. Simulated speech was used to avoid ceiling performance. Simulating CIs using eight channels has been shown to yield accuracy levels of approximately 70% in normal-hearing listeners in sentence recognition tasks; this is similar to the accuracy afforded by multi-talker babble background noise (Bent, Buchwald, & Pisoni, 2009). A motivation for using CI speech instead of babble is that babble often leads to higher thresholds than other noise, likely due to the variation in phonemic cues (e.g., Hall, Grose, Buss, & Dev, 2002; also Altieri & Hudock, 2014a). Additionally, using vocoded

speech can have translational value for researchers interested in comparing our results to CI users.

Participants

Data obtained from 86 adults (58 females) between 18 and 60 years-old with normal or corrected-to-normal vision were included. Other requirements were that all participants must be native speakers of American English, have no reported history of cognitive deficit, or clinically diagnosed hearing loss. (Listeners with hearing aids did not participate in the study). Data for one participant was lost due to computer error, and mean replacement was used to fill in missing data. The mean participant age was 30.93 years old ($SD = 12.04$ years old; range 18–60 years old). Table 1 contains information regarding the number listeners in each age range, and mean low (250–1000 Hz) and high-frequency (2000–8000) pure-tone thresholds (dB SPL) in their better ear (Altieri & Hudock, 2014a).

To obtain a representative sample of young-middle age adult listeners, e-mails were sent out to at Idaho State University (ISU) (faculty, staff, graduate, and undergraduate students) and the Pocatello Idaho community over e-mail. The ISU campus and surrounding area is an ideal location to obtain such a sample of adult listeners. The ISU campus comprises a non-traditional student body with diverse age groups with a substantial number of students above age 35 that come from diverse backgrounds (e.g., stay-at-home mothers and returning veterans). Participants were paid 10-dollars per hour. The ISU Institutional Review Board approved this study.

Experiment 1: CUNY Sentence Recognition—CUNY stimulus materials consisted of 75 sentences presented to each listener. These included 25 audiovisual, 25 auditory, and 25 visual-only sentences spoken by a female talker. Each set included word lengths of 3, 5, 7, 9, and 11 (Boothroyd, Hanin, & Hnath, 1985; five sentences each). The order of audiovisual, auditory, and visual-only block presentation was randomized across participants to avoid order effects.

Each trial began with the presentation of a black dot on a solid gray background, which cued the participant to press the space bar on the keyboard to begin the trial. Stimulus presentation began with the female talker speaking one of the sentences. After the talker finished speaking the sentence, a dialog box appeared in the center of the monitor instructing the participant to type in the words they thought the talker said. Each sentence was given to the participant only once and feedback was not provided on any of the trials.

Scoring was carried out in a manner identical to the protocol described by Altieri et al. (2011). Whenever the participant correctly typed a word, the word was scored as “correct”. The proportion of words correct was scored in each sentence. Word order was not a criterion for a word to be scored as accurate, and typed responses were manually corrected for misspellings. For example, in the sentence “Is your sister in school?” suppose the participant typed “Is the ...”. Only the word “Is” would be scored correct, making the total proportion correct equal to 1/5 or .20. Upon inspection of the data, participants rarely switched word order in their responses.

Experiment 2: Closed Set Monosyllabic Word Recognition—We selected eight high-frequency words with variable phonemic onsets and offsets. The stimulus words included: *Mouse, Job, Tile, Gain, Shop, Boat, Page, and Date*. We employed a closed-set discrimination task for the following reasons. First, we required a test from which we could obtain RT distributions. A closed-set speeded-response task is thus viable for calculating capacity, whereas an open-set task typically only affords measures of accuracy. An alternative way to calculate RTs using an open-set format could require participants to make a higher-order judgment (e.g., “Respond ‘YES’ if the word refers to something living, and ‘NO’ if it refers to something non-living”; e.g., Winneke and Phillips, 2011). However, this method does not allow for the researcher to assess recognition errors, or develop confusion matrices for model fitting which are common in many audiovisual speech studies (e.g., Braida, 1991; Grant et al., 1998; Massaro, 2004). Therefore, for capacity calculations, the closed-set format employed here is the suggested approach (see Altieri & Townsend, 2011; Altieri & Wenger, 2013; Altieri et al., 2014). Finally, this task was originally developed to replicate accuracy and enhancement levels obtained in Sumbly and Pollack’s original research when they employed a set size of eight words (Altieri & Townsend, 2011).

Two female talkers with the highest intelligibility ratings were selected from the Hoosier Multi-Talker Database (Sherffert, Lachs, & Hernandez, 1997). Bradlow, Toretta, and Pisoni (1996) showed that female talkers tend to be more intelligible than male talkers; interestingly, this gender difference is preserved under different forms of stimulus degradation, such as CI simulation and multi-talker babble (Bent et al., 2009). There were a total of 128 audiovisual trials (8 words $\times$ 2 talkers $\times$ 2 recordings $\times$ 4 repetitions): 128 auditory-only trials, and 128 visual-only trials, for a total of 384 experimental trials. Participants received 48 practice trials at the onset of the experiment to assist with learning the response mappings. Feedback was provided (“Correct” vs. “Incorrect”) during practice, although not in the actual experiment. Auditory stimuli were played at the same comfortable listening volume for each listener (~ 68 dB SPL) over Headphones. Responses were collected through button-press on a standard keyboard. Each of the buttons, 1 through 8, was arranged linearly labeled with a word from the stimulus set. The position of the labels was identical for all participants. Participants were instructed to press the button labeled with the word that they judged the talker to have said “as quickly and accurately as possible”. On auditory-only trials, listeners were required to base their response only on auditory information, and on visual-only trials, they were required to base their responses only on visual cues.

Monosyllabic audiovisual, auditory, and visual-only words were presented in random order to participants. Stimuli were not blocked by modality as they were in Experiment 1. This was done to avoid modality-order effects that can arise, for example, if a block of easier audiovisual trials preceded more difficult auditory or visual-only trials. Order effects are more likely to arise in Experiment 2 because the same stimuli are repeated in audiovisual, auditory and visual-only conditions. An inspection of our data with previous versions of this study in which blocking was used (e.g., Altieri & Wenger, 2013; Altieri et al., 2014) did not reveal any substantial differences in capacity.

Results

Sentence Recognition

Figure 1 shows box-plots of the audiovisual, auditory-only, and visual-only sentence recognition accuracy. The middle line indicates the median, the surrounding lines indicate the 25th and 75th percentiles, and the whiskers indicate 1.5 times the interquartile range. Outliers are denoted by plus “+” signs. Results demonstrated a mean audiovisual recognition accuracy of 95.1% (SD = 2.9%), an auditory-only accuracy of 75.8% (SD = 8.6%), and a visual-only accuracy of 12.7% (SD = 6.9%). The visual-only results replicated those reported by Altieri et al. (2011). A repeated measures ANOVA using arcsine transformed accuracy scores suggested significant differences in accuracy across conditions ($F(2,85) = 3081, p < .0001$). Paired-samples t-tests demonstrated more accurate audiovisual accuracy compared to auditory-only, and more accurate auditory-only accuracy compared to visual-only ($p < .0001$, for both).

Next, two measures of audiovisual benefit obtained from visual cues over and above those provided by the auditory modality alone (i.e., “gain”) are displayed in Figure 1. The first was a normalized measure of visual gain, which was computed by comparing audiovisual accuracy to auditory-only accuracy divided by the total possible benefit ($[AV - A] / [100 - A]$; $M = 78.1\%$, $SD = 13.4\%$; e.g., Tye-Murray et al., 2007). A normalized measure was used because many auditory and audiovisual accuracy scores were near ceiling. The second measure compared audiovisual accuracy to predictions assuming independence across modalities ($AV - [A + V - A \times V]$; $M = 16.5\%$, $SD = 7.7$). In this latter measure, the predictions for independent processing are given in the brackets. The term “ $A + V$ ” denotes the sum of auditory and visual accuracy probability correct, while “ $A \times V$ ” represents the intersection or multiplication of auditory and visual probability correct when independence is assumed.

Monosyllabic Word Recognition

Accuracy—The analysis of mean accuracy revealed a similar pattern reported in the mean RT analysis, with high audiovisual ($M = 99\%$; $SD = 1.7\%$) and auditory-only performance ($M = 98\%$; $SD = 5.4\%$), but lower visual-only performance ($M = 73\%$; $SD = 9.6\%$). Results from a repeated-measures ANOVA using arcsine transformed accuracy scores indicated differences in accuracy across conditions ($F(2,85) = 1327, p < .0001$). This was driven by lower accuracy in the visual-only condition compared to both the auditory ($t(85) = 28.4, p < .0001$) and audiovisual conditions ($t(85) = 28.2, p < .0001$).

Reaction Time (RT)—The analysis of mean RTs in the eight-word forced choice experiment revealed a mean of 1,845 ms ($SD = 271$ ms) on audiovisual trials, 1,835 ms ($SD = 270$ ms) on auditory-only trials, and 2,469 ms ($SD = 448$ ms) on visual-only trials. Results from a repeated-measures ANOVA indicated differences in RT across conditions in the speeded word recognition task ($F(2,85) = 429, p < .0001$). Critically, this finding was driven by slower responses in the visual condition compared to either the auditory-only ($t(85) = 17.38, p < .0001$) or audiovisual ($t(85) = 19.89, p < .0001$), as shown by paired samples t-tests comparing these conditions. The paired samples t-test comparing mean RTs between

the audiovisual and auditory-only conditions was non-significant ($t(85) = 1.51, p = .14$). This finding suggests that on average, the normal-hearing adult population does not respond to auditory plus visual speech cues significantly faster than auditory-only cues. In terms of capacity, this pattern of results should predict that capacity will range between limited to unlimited across most processing times.

Capacity—In line with previous studies examining capacity (e.g., Altieri & Hudock, 2014a; 2014b; Altieri & Townsend, 2011; Altieri & Wenger, 2013) only correct responses were analyzed in the functions. Capacity functions can be calculated using either correct or incorrect responses (Townsend & Altieri, 2012).² Accuracy often approached ceiling levels in audiovisual and auditory conditions, so using incorrect responses was not viable. Normative capacity scores were computed by averaging continuous capacity values across time points for each participant. For each participant, RTs faster or slower than ± 3 standard deviations from the mean were removed prior to calculating capacity. (Similar to Houpt et al., (2014), data were retained for time points which provided stable integrated hazard function estimates.). Averaged data were computed only across time points that contained data from at least 90% of participants; this prevented atypical participants from skewing fast or slower processing times in the capacity functions.

The panel on the right shows the RT-only capacity measure of $C(t)$, while the left panel in Figure 2 plots $C_I(t)$; the measure that incorporates both RT and accuracy. The dotted lines above and below the mean depict departures of ± 1.5 standard deviations. The averaged capacity data suggest limited to slightly limited capacity ranging from .5 to approximately 1. This suggests that adult listeners may benefit slightly from visual information for some RTs, although the actual benefit is often less than predicted by independent race models. Differences between $C(t)$ and $C_I(t)$, which exhibited more limited capacity, particularly for slow RTs, could be due to the accuracy correction. Because unisensory accuracy was generally near ceiling, the audiovisual benefit in terms of accuracy was weak and therefore negatively impacted $C_I(t)$.

Towards a Comprehensive Assessment Approach

For statistical computation and general summary purposes, it is advantageous to obtain single-point estimates of the continuous capacity functions. One summary statistic that has been used in previous studies is *maximum capacity* because it provides an estimate of an individual listener's total integration ability (Altieri & Hudock, 2014a; 2014b; Altieri & Wenger, 2013). Other potentially useful but previously unexplored estimates include median and mean capacity, which provide measures of central tendency of the capacity function.

Correlations were carried out between capacity and audiovisual gain measures from sentence recognition to begin establishing which capacity measure(s) may be significantly correlated with traditional gain scores from open-set sentence recognition. (A strong positive correlation between one or more capacity measures and gain scores ($> .90$) would indicate that the tests are redundant.). Table 2 shows Pearson correlation coefficients between $C(t)$,

²Research has shown that RT-only $C(t)$ functions are robust even for an error rate of 30 % (Townsend & Wenger, 2004).

$C_I(t)$, and the two audiovisual benefit or gain scores from sentence recognition. Additionally, we displayed correlations from Experiment 2: These gain scores were low because mean audiovisual and auditory-only accuracy were near ceiling. In future studies, these measures may be very important, especially for hearing-impaired listeners (see Altieri & Hudock, 2014b). All tests were one-tailed due to the assumption that positive correlations or trends would be present; however, we did use a more conservative α of .008 to adjust for multiple comparisons. Results indicate that maximum $C_I(t)$ may be weakly to moderately correlated with the audiovisual benefit measure from sentence recognition. This suggests that while maximum $C_I(t)$ is correlated with accuracy gain, the measures do not appear redundant.

To provide an overall summary, Figure 3 shows box and whiskers plots for mean, median, and maximum $C_I(t)$, and $C(t)$ obtained from the monosyllabic word recognition experiment. The middle lines indicate the median, the surrounding lines indicate the 25th and 75th percentiles, the whiskers indicate 1.5 times the interquartile range, and outliers are shown as “+” sign. The capacity values in the figure indicate that many adult listeners are capable of unlimited (≈ 1), and in some cases, super-capacity (i.e., $C(t) > 1$), which denotes efficient integration.

To investigate the extent to which word recognition in the auditory and visual-only modalities were predictors for capacity we carried out Pearson correlations between the six capacity measures and mean unisensory accuracy scores. These results are shown in Table 3. Results indicate a negative relationship between auditory and visual-only accuracy, and capacity; that is, poorer unisensory recognition skills are associated with better integration scores indexed by higher capacity. However, the only significant correlations that emerged were between mean and maximum $C(t)$, and visual-only word recognition. Because there was an outlier for auditory-only CUNY sentence recognition (see Fig. 1), we also carried out the correlations without this outlier. This listener was in the oldest age group and age may have been a factor influencing their ability to recognize CI speech. However, this listener’s capacity and gain scores were not outliers, and the removal of this data point did not affect the significance of the correlations in Table 2 or 3.

In summary, while a positive correlation was observed between maximum capacity and $AV - [A + V - A \times V]$ (negative correlations for visual-only accuracy and capacity), the correlations were only moderate. Maximum $C_I(t)$ thus fails to adequately account for the variability observed in integration measures obtained from open-set sentence recognition. It may be that the capacity and accuracy measures are tapping into different factors. Another approach to developing a comprehensive assessment of integration can be accomplished by identifying the underlying factors (e.g., processing speed or accuracy) contributing to audiovisual recognition.

Factor Analysis

This section discusses results of exploratory and subsequent confirmatory factor analyses carried out on the following eight measures of audiovisual integration: two measures of AV benefit from the CUNY sentence recognition experiment, plus maximum, median, and mean $C(t)$, and the same three measures for $C_I(t)$ from the word recognition experiment.

Exploratory Factor Analysis—First, exploratory factor analysis using principal components analysis (PCA) was carried out to determine the number of dimensions that best explained the data structure. We initially predicted that two dimensions would best explain the variance in the data: one corresponding to capacity (RT) and the other to accuracy and benefit scores from sentence recognition.

To test this hypothesis, a scree plot and Horn's (1965) parallel analysis methods were used to determine the number of factors to be used in the subsequent confirmatory factor analysis. Scree plots display the individual factors on the abscissa and the eigenvalues corresponding to those components (in descending order) on the ordinate. When the drop in eigenvalues creates an elbow, this methodology suggests dropping all factors after that elbow. However, choosing the elbow can be subjective especially in cases in which there is more than one elbow. Another approach, known as "Parallel" or "Horn's analysis", compares the actual eigenvalues to those obtained from simulated uncorrelated variables. All factors are retained that are larger than the eigenvalues obtained from the random data (cf. Ford et al., 1986; and Comrey & Lee, 1992). Contrary to our initial expectations the scree plot, Horn's analysis, and model-fit ($X^2(20) = 222.40$, $p < .0001$) indicated that a single-factor model provided the best account of the data structure. What may account for the low factor correlations between the capacity and audiovisual gain scores on one hand, and the single factor model on the other (observed in the Scree plot, Horn's analysis, and statistical results)? One possibility is that the difference between tasks partially accounted for the low correlations. Our interpretation is that only one measure of the latent variable was obtained; this result, while mathematically plausible, was unexpected given our initial hypothesis regarding separate factors for speed and accuracy.

Confirmatory Factor Analysis—Next, confirmatory factor analysis was carried out using the Matlab (R2012a) function *factoran* (Matlab Statistics Toolbox) to determine the loadings or coefficients of each variable on the latent factor. This was done to determine which variables correlate with the factor, thereby showing which of the dependent variables provide the best explanatory power for the variance in the CUNY sentence and capacity data. Results from the confirmatory factor analysis, using the single-factor model with varimax rotation, are displayed in Table 4. Similar results were reported for an oblique factor rotation (i.e., promax; see Comrey and Lee (1992) for justification of rotational procedures). (Refer to Ford et al. (1986) for a review of Factor Analysis procedures).

An examination of the factor coefficients indicates that the capacity measures provide a better summary of integration skills compared to the accuracy benefit scores from sentence recognition. Mean and median $C_I(t)$ showed the strongest association with very high factor loadings of .90 and .94 respectively (see Ford et al., 1986, for discussion on strength of loadings). Maximum $C_I(t)$ was also strong providing a loading of .78. Finally, the $C(t)$ capacity measures—the mean and median—both yielded strong factor loadings of .87. This pattern of results suggests that integration skills could be measured in the most comprehensive manner using a measure that computes both speed and accuracy. Therefore, the single factor model suggests that audiovisual integration is best accounted for by a combination of speed and accuracy performance (which is captured by $C_I(t)$), rather than one separate factor for speed and another factor for accuracy.

The factor analysis results suggested that the best way to examine whether capacity differed across age, is to test whether mean and median $C_I(t)$ values were significantly correlated with listener's age. First, mean $C_I(t)$ was weakly although significantly correlated with age using an α level of .05 ($r(83) = .23$). The mean $C_I(t)$ averaged across listeners in each age group was: 18–25 years old = .69; 26–35 years old = .78; 36–49 years old = .79; 50–60 = .79. Next, median $C_I(t)$ scores were also positively correlated with listener age, although the correlation was weaker ($r(83) = .19$). The group averaged median $C_I(t)$ for each age group was: 18–25 years old = .69; 26–35 years old = .79; 36–49 years old = .79; and 50–60 years old = .76. We carried out a second factor analysis on all listeners under 35 years old ($N = 59$): exploratory factor analysis once again revealed the existence of only one factor to explain the data. Second, confirmatory factor analysis showed an identical pattern of factor loadings, with median and mean $C_I(t)$ being strongest followed by mean and median $C(t)$.

The capacity methodology should prove advantageous because RTs provide valuable information regarding recognition ability. This ability may be influenced by a variety of factors, including a listener's hearing ability or age (Ratcliff et al., 2001). While accurately decoding a speaker's message is important and should be accounted for, so is the ability to rapidly decode the information. The measures of $C(t)$, and particularly $C_I(t)$, mathematically capture "efficiency as a function of workload": a concept fundamental for human information processing (Townsend & Altieri, 2012; Townsend & Nozawa, 1995). This is done by incorporating RT distributions in the former, and a combination of RT distributions with a correction for accuracy in the latter measure.

General Discussion

A major contribution of this study is that researchers now have access to specific benchmarks for assessing auditory-visual speech recognition abilities. This was accomplished by obtaining data on statistically motivated measures of capacity that compare recognition skills against independent model predictions; thus far, these have only been applied to small groups of participants and selected samples of listeners (Altieri & Hudock, 2014a; 2014b). Additionally, we replicated results revealing the difficulty of visual-only speech recognition (mean $\approx 12\%$ correct; Altieri et al., 2011).

This report provided a major step toward making audiovisual speech tests clinically viable: Specifically, these data should be useful for comparing a listener's audiovisual integration skills with that of the general adult population. One caveat is that our results may depend in part on the talkers because different talkers afford different levels of intelligibility (e.g., Bradlow et al. 1996). Hence, different talkers (e.g., males) could have influenced auditory and visual-only accuracy, and perhaps even contributed to differences in capacity scores. Additionally, the method of stimulus degradation may have impacted accuracy. Bent and colleagues (2009) nonetheless observed that CI speech and multi-talker babble yield similar levels of accuracy, although the rate of adaptation may slightly differ when multiple sessions are used. Relatedly, Bent, Loebach, Phillips, and Pisoni (2011) showed that bottom-up acoustical properties can be sufficient for adaptation to CI speech. Thus, while signal degradation occurs via different mechanisms in babble versus CI speech, accuracy and

adaptation effects appear similar enough to render vocoded speech useful across many experimental settings.

A significant contribution of this study was the factor analysis results indicating that only one underlying factor appears to explain audiovisual speech integration skills. This was true when data from both sentence recognition and monosyllabic word recognition were assessed. Perhaps most significantly, capacity loaded most strongly on the integration factor: especially the mean and median $C_I(t)$ which measure both speed and accuracy. The significance of these results lies in the fact that capacity characterizes a listener's ability to integrate audiovisual sources of information better than traditional measures of benefit that only compute accuracy. A useful application would be to determine where an individual listener falls on the standardized distribution by testing them on a closed set speeded response task similar to the one used in this study. A recent case study report indicates differences between listeners with different audiometric configurations in terms of capacity values (Altieri & Hudock, 2014b); however, these data had not been compared to a standardized data set, and only maximum $C_I(t)$ and $C(t)$ were computed.

Summary and Conclusion

In summary, our data should prove useful because clinical audiologists do not typically obtain visual-only or audiovisual speech recognition measures. When they do obtain these measures, they do not have a standard procedure or data-set to make comparisons against. These data will be a valuable tool for researchers who work with suspected hearing-impaired listeners to determine their strengths and weaknesses across the domains of accuracy and processing speed, and also across auditory and visual modalities. Auditory recognition skills matter in face-to-face settings; however, audiologists should take into consideration audiovisual integration, visual-only speech perception, and cognitive skills in a comprehensive diagnostic and treatment protocol (Remensnyder, 2012). Suggestions for making our approach clinically useful include reducing the number of experimental trials. Reducing the number of stimulus items in the set, particularly for very young or elderly listeners, may also be beneficial. We recommend pilot testing in a clinical setting using these changes before they are implemented by audiologists.

Acknowledgments

The project described was supported by an internal University Research Office Grant awarded to the first author at Idaho State University, the INBRE Program, NIH Grant Nos. P20 RR016454 (National Center for Research Resources) and P20 GM103408 (National Institute of General Medical Sciences.)

References

- Altieri N, Hudock D. Variability in audiovisual speech integration skills assessed by combined capacity and accuracy measures. *Int J Audiol.* 2014a; 53:710–718. [PubMed: 24806080]
- Altieri N, Hudock D. Hearing impairment and audiovisual speech integration ability: A case study report. *Front Psychol.* 2014b; 5:1–10. [PubMed: 24474945]
- Altieri N, Pisoni DB, Townsend JT. Some normative data on lip-reading skills. *J Acoust Soc Am.* 2011; 130:1–4. [PubMed: 21786870]

- Altieri N, Townsend JT, Wenger MJ. A dynamic assessment function for measuring age-related sensory decline in audiovisual speech recognition. *Behav Res Methods*. 2014; 46:406–415. [PubMed: 23943582]
- Altieri N, Wenger M. Neural dynamics of audiovisual integration efficiency under variable listening conditions. *Front Psychol*. 2013; 4:1–15. [PubMed: 23382719]
- Bent T, Buchwald, Pisoni DB. Perceptual adaptation and intelligibility of multiple talkers for two types of degraded speech. *J. of the Acoustical Society of America*. 2009; 126:2660–2669.
- Bergeson, TR.; Pisoni, DB. Audiovisual speech perception in deaf adults and children following cochlear implantation. In: Calvert, G.; Spence, C.; Stein, BE., editors. *The Handbook of Multisensory Processes*. Cambridge, MA: MIT Press; 2004. p. 749-771.
- Boothroyd, A.; Hanin, L.; Hnath, T. A sentence test of speech perception: Reliability, set equivalence, and short term learning (internal report RCI 10). New York: City University of New York; 1985.
- Bradlow AR, Toretta GM, Pisoni DB. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Comm*. 1996; 20:255–272.
- Comrey, AL.; Lee, HB. *A first course in factor analysis* (2nd edition). Hillsdale, NJ: Lawrence Erlbaum Associates; 1992.
- Erber NP. Use of hearing aids by older people: influence of non-auditory factors (vision, manual dexterity). *Int J Audiol*. 2003; 42:2S21–2S26. [PubMed: 12918625]
- Ford JK, MacCallum RC, Tait M. The application of exploratory factor analysis in applied psychology: a critical review and analysis. *Pers Psychol*. 1986; 39:291–314.
- Grant KW, Walden BE, Seitz PF. Auditory-visual speech recognition by hearing impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *J Acoust Soc Am*. 1998; 103:2677–2690. [PubMed: 9604361]
- Hall JW, Grose JH, Buss E, Dev MB. Spondee recognition in a two-talker masker and a speech-shaped noise masker in adults and children. *Ear Hear*. 2002; 23:159–165. [PubMed: 11951851]
- Horn J. A rationale and test for the number of factors in factor analysis. *Psychometrika*. 1965; 30:179–185. [PubMed: 14306381]
- Houpt JW, Townsend JT, Donkin C. A new perspective on visual word processing efficiency. *Acta Psychol*. 2014; 145:118–127.
- Massaro, DW. From multisensory integration to talking heads and language learning. In: Calvert, GA.; Spence, C.; Stein, BE., editors. *The Handbook of Multisensory Processes*. Cambridge, MA: The MIT Press; 2004. p. 153-176.
- Ratcliff R, Thapar A, McKoon G. The effects of aging on reaction time in a signal detection task. *Psychol Aging*. 2001; 16:323–341. [PubMed: 11405319]
- Remensnyder LS. Audiologists as gatekeepers: And it's not just for hearing loss. *Audiol Today*. 2012:25–31.
- Sherffert, S.; Lachs, L.; Hernandez, LR. Research on Spoken Language Processing Progress Report No. 21. Bloomington, IN: Speech Research Laboratory, Psychology, Indiana University; 1997. The Hoosier audiovisual multi-talker database.
- Sommers M, Tye-Murray N, Spehar B. Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear Hear*. 2005; 26:263–275. [PubMed: 15937408]
- Stevenson RA, Sarko DK, Nidiffer AR, Ghose D, Fister J, et al. Identifying and quantifying multisensory integration: A tutorial review. *Brain Topogr*. 2014; 27:707–730. [PubMed: 24722880]
- Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. *J Acoust Soc Am*. 1954; 26:12–15.
- Townsend JT, Altieri N. An accuracy-response time capacity assessment function that measures performance against standard parallel predictions. *Psychol Rev*. 2012; 119:500–516. [PubMed: 22775497]
- Townsend JT, Nozawa G. Spatio-temporal properties of elementary perception: An investigation of parallel, serial and coactive theories. *J Math Psychol*. 1995; 39:321–360.

- Townsend JT, Wenger MJ. A theory of interactive parallel processing: New capacity measures and predictions for a response time inequality series. *Psychol Rev.* 2004; 111:1003–1035. [PubMed: 15482071]
- Tye-Murray N, Sommers M, Spehar B. Audiovisual integration and lip-reading abilities of older adults with normal and impaired hearing. *Ear Hear.* 2007; 28:656–668. [PubMed: 17804980]
- Winneke AH, Phillips NA. Does audiovisual speech offer a fountain of youth for old ears? An event-related brain potential study of age differences in audiovisual speech perception. *Psychol Aging.* 2011; 26:427–438. [PubMed: 21443357]

Figure 1.

Boxplots showing the recognition accuracy scores from the CUNY sentence recognition study. Audiovisual (AV), auditory-only (A), and visual-only (V) are included, as are two measures of audiovisual benefit.

Figure 2.

Capacity values across RT values averaged across participants. The left panel shows accuracy corrected capacity, $C_I(t)$, while the right shows $C(t)$. The dotted lines above and below the thick line (representing mean capacity), indicate ± 1.5 standard deviations.

Figure 3.
Boxplots showing the distribution of capacity values from the monosyllabic word recognition experiment.

Table 1

Number of listeners in each age range and their mean (SD) pure-tone thresholds.¹

Age (years)	N	Mean LF Threshold (SD)	Mean HF Threshold (SD)
18–25	42	9.9 dB (5.5)	7.6 dB (9.5)
26–35	17	6.7 dB (4.7)	4.1 dB (4.7)
36–49	18	9.4 dB (4.5)	11.7 dB (4.5)
50–60	9	10.9 dB (5.4)	16.1 dB (8.9)

¹Averaging across spectra is not a precise representation of auditory tuning curves or hearing acuity. It is a simplification based on previous work assessing auditory-visual benefit in light of low versus high-frequency hearing loss (cf. Erber, 2003).

Table 2

Pearson correlations between capacity and AV gain for Experiments 1 and 2.

Capacity	Exp1 $_{[AV-A] / [100-A]}$	Exp1 $_{AV - [A + V - A \times V]}$	Exp2 $_{[AV-A] / [100-A]}$	Exp2 $_{AV - [A + V - A \times V]}$
Mean $C(t)$	.15	.08	-.07	.01
Med. $C(t)$	.01	.01	-.07	.05
Max. $C(t)$	.08	.07	-.09	-.06
Mean $C_I(t)$	.19	.19	.18	.05
Med. $C_I(t)$	.17	.19	.01	.01
Max. $C_I(t)$	.23	.30*	.04	.03

The asterisk “*” denotes significance at $p = .008$ (corrected $\alpha = .05/6 = .008$ for experiments 1 and 2).

Table 3

Pearson correlation coefficients for capacity and unisensory scores.

Capacity	Mean Auditory	Mean Visual
Mean $C(t)$	-.17	-.36*
Median $C(t)$	-.12	-.24
Max $C(t)$	-.20	-.32*
Mean $C_{-I}(t)$	-.11	-.02
Median $C_{-I}(t)$	-.09	.01
Max $C_{-I}(t)$	-.12	-.06

The asterisk “*” indicates significance at $p = .008$ (Corrected $\alpha = .05/6 = .008$).

Table 4

This table displays the factor loadings (i.e., coefficients) of each variable on the underlying factor corresponding to *integration*.

Integration Measure	Factor Loading
$[AV - A] / [100 - A]$	.21
$AV - [A + V - A \times V]$	.17
Mean $C_I(t)$	.90*
Median $C_I(t)$	.94*
Max $C_I(t)$	.78*
Mean $C(t)$	.87*
Median $C(t)$	.87*
Max $C(t)$	.45

* Indicates a strong loading of ($> .50$).