Visual speech information: A help or hindrance in perceptual processing of dysarthric speech

Stephanie A. Borrie[a)]
Department of Communicative Disorders and Deaf Education, Utah State University, Logan, Utah 84322

(Received 2 June 2014; revised 24 October 2014; accepted 21 January 2015)

This study investigated the influence of visual speech information on perceptual processing of neurologically degraded speech. Fifty listeners identified spastic dysarthric speech under both audio (A) and audiovisual (AV) conditions. Condition comparisons revealed that the addition of visual speech information enhanced processing of the neurologically degraded input in terms of (a) acuity (percent phonemes correct) of vowels and consonants and (b) recognition (percent words correct) of predictive and nonpredictive phrases. Listeners exploited stress-based segmentation strategies more readily in AV conditions, suggesting that the perceptual benefit associated with adding visual speech information to the auditory signal—the AV advantage—has both segmental and suprasegmental origins. Results also revealed that the magnitude of the AV advantage can be predicted, to some degree, by the extent to which an individual utilizes syllabic stress cues to inform word recognition in AV conditions. Findings inform the development of a listener-specific model of speech perception that applies to processing of dysarthric speech in everyday communication contexts.
 [http://dx.doi.org/10.1121/1.4913770]

[MAH] Pages: 1473–1480

I. INTRODUCTION

Everyday speech communication typically takes place face-to-face. This statement is particularly applicable for encounters involving individuals with neurological speech disorders, whereby the telephone is perceived as a difficult communication medium and frequently avoided (Dickson *et al.*, 2008). Accordingly, the task of perceiving dysarthric speech is largely a multisensory phenomenon involving the detection, integration, and interpretation of both auditory (A) and visual (V) speech information. Although Mattys *et al.* (2013, p. 1) have recognized that "improving the validity of speech-recognition models requires an understanding of the conditions in which speech is experienced on a daily basis," no plausible models that address perceptual processing of dysarthric speech in audiovisual (AV) contexts currently exist.

Speech perception involves more than just the auditory signal. A rich body of literature, spanning decades and disciplines, examines the contribution of visual speech information for recognizing spoken language and provides substantial evidence to support the claim that simultaneous auditory and visual speech input provides a perceptual benefit—the AV advantage—over that achieved from just the auditory information. This AV advantage is evidenced in typical settings (e.g., Davis and Kim, 2004; Helfer, 1997) but is most pronounced in suboptimal or adverse conditions, including listener limitations of hearing impairment (e.g., Grant *et al.*, 1998), environmental degradation related to background noise (e.g., Ross *et al.*, 2007), and source degradation associated with synthetic speech (Reynolds *et al.*, 1997) and esophageal speech (Evitts *et al.*, 2010). Further, this AV advantage can be clinically significant, with reports documenting intelligibility improvements of up to 80% for listeners recognizing speech in noise in AV versus A conditions (Sumby and Pollack, 1954).

The perceptual benefit of adding visual speech information is considered to stem from redundant and complementary segmental and suprasegmental information provided by the additional sensory modality (Grant and Seitz, 2000; Jesse and Massaro, 2010; Klucharev *et al.*, 2003). Under adverse listening conditions, this complementary information may become particularly beneficial, providing cues that constrain possible interpretations of the ambiguous auditory information and subsequently reducing the cognitive demands associated with processing speech (Grant and Seitz, 2000). For example, listeners may struggle to identify an indistinct auditory production of the phoneme /p/, but cues afforded in the visual correlates of this sound (i.e., lip closure) increase the likelihood that the phoneme will be recognized successfully. The existing body of work provides evidence that simultaneously perceiving both auditory and visual speech information provides a perceptual enhancement, over and above that afforded by the auditory modality alone. The current study seeks to determine if this claim extends to perceptual processing of neurologically degraded speech.

Darley *et al.* (1969b, p. 246) defined dysarthria as "a collective name for a group of speech disorders resulting from disturbances in muscular control over the speech mechanism due to damage of the central or peripheral nervous system." Dysarthria manifests itself in a number of segmental (e.g., phoneme omissions, distortions, and substitutions) and suprasegmental (e.g., intonation, vocal intensity, and rate-rhythm) deficits that render the acoustic signal "unsuitable, in varying ways and degrees, for language perception" (Weismer, 2006, p. 320). Disturbed muscular control has perceptual consequences for the availability of auditory information, but also for the affordance of visual

a)Author to whom correspondence should be addressed. Electronic mail: stephanie.borrie@usu.edu

speech cues provided in a speaker's face (e.g., lip and jaw movement). These consequences are dependent upon the localization of neural damage. Damage to the cerebellar control circuit, for example, is associated with incoordination and reduced muscle tone whereas basal ganglia pathology is characterized by decreased mobility and range of movement. Together, these deficits make dysarthric speech difficult for listeners to decipher (Klasner and Yorkston, 2005; Liss *et al.*, 2000).

If visual input disambiguates degraded auditory input, then perceptual detriments may increase when the visual cues themselves are also degraded, as is the case for individuals with dysarthria. Thus, the AV advantage, evidenced in other adverse listening conditions, may not apply to dysarthric speech. In the case of perceptual processing of neurologically degraded speech, there could actually be an AV *disadvantage*—reduced speech perception in AV conditions relative to A conditions. An AV disadvantage was recently documented in studies in which listeners were required to perform an additional task (e.g., tactile pattern recognition) while recognizing speech in noise under A and AV conditions (Fraser *et al.*, 2010; Rudner and Rönnberg, 2004). The results of these studies suggest that increased processing effort may in fact reverse the classic AV advantage reported in the existing literature.

Such a hypothesis is supported by a renowned perceptual phenomenon, the McGurk effect, which purports that visual speech information can interfere with accurate perception of auditory information. A dramatic illustration of the cross-modal influence of visual information on processing auditory information was documented when listeners reported hearing the phoneme /t/, a sound that was neither seen nor heard but derived by synthesizing information from both the auditory (e.g., /pa/) and visual (e.g., /ka/) signals (McGurk and MacDonald, 1976). The McGurk effect has been observed in both adult and child populations (e.g., Burnham and Dodd, 2004) and is robust under a variety of experimental manipulations (e.g., Jordan and Sergeant, 2000). This auditory illusion suggests that the perceptual system "fuses" auditory and visual speech information together to ascertain a speech percept. Incongruent auditory and visual speech information may challenge the perceptual integrity of the incoming signal, and as such, adversely affect successful speech perception.

Despite the validity of examining speech perception in face-to-face interactions, very few studies have explored the influence of visual speech information in perceptual processing of neurologically degraded speech. Conclusions arising from studies that have compared intelligibility scores for listeners exposed to dysarthric speech under A versus AV conditions are largely inconclusive. While some studies with dysarthric stimuli have yielded results that are consistent with the AV advantage (e.g., Garcia *et al.*, 1992; Schumeyer and Barner, 1996), other studies have failed to detect its presence (e.g., Keintz *et al.*, 2007). The suggestion that these equivocal findings may be related to severity of degradation is disputable. Hunter *et al.* (1991) observed the AV advantage when listeners were tasked with perceiving speech associated with moderate dysarthria but not when they were asked to distinguish the speech of severe dysarthria. Conversely, Hustad and Cahill (2003) reported the AV advantage for only one of five speakers, and this advantage was associated with a speaker who exhibited the most severe intelligibility deficit. These studies suggest that visual information may play a role in deciphering dysarthric speech, however, further investigation is warranted. Specifically, research is needed to explore the cognitive-perceptual mechanisms that underlie perceptual processing in different listening contexts. Listeners use a variety of segmental and suprasegmental cues to help them recognize speech (see Mattys *et al.*, 2005). Whether they deploy different perceptual strategies to recognize dysarthric speech under A versus AV conditions could provide critical information for the development of speech recognition models with this population.

Finally, current models of speech recognition with normal-hearing listeners have been routinely developed upon group averages. While individual differences have been examined, reflection on individual variability has been minimal, and researchers have largely treated these differences as a type of error or as a justification to remove the data from the analysis. This illusion of homogeneity among normal-hearing listeners is further confounded because the intelligibility of typical speech generally hovers around ceiling. However, as the signal and/or listening conditions become increasingly degraded, performance variability among listeners begins to emerge (e.g., Andersson *et al.*, 2001; Bernstein *et al.*, 2000). This variability reflects the nature of speech perception, which is in fact a complex cognitive-perceptual task that some listeners are better equipped to tackle than others. The neurologically degraded speech signal and the task of processing multisensory input may reveal and amplify differences in perceptual processing that are not apparent when listening to clearer speech in a single sensory modality.

The purpose of the current study was to investigate the influence of visual speech information on perceptual processing of dysarthric speech. The following research questions were addressed: (1) does the addition of visual speech information enhance phoneme acuity (identification of vowels and consonants) and word recognition (intelligibility of words in phrases) of dysarthric speech; (2) are the cognitive-perceptual processes that underlie word recognition in A conditions the same as those that underlie word recognition in AV conditions (i.e., do listeners deploy different perceptual strategies when visual information is added); and (3) can the effects of adding visual speech information be predicted by individual segmental and suprasegmental cue use profiles? It was hypothesized that visual speech information may disrupt perceptual processing of dysarthric speech and, further, that this interference would be reflected in cognitive-perceptual strategies recruited for lexical segmentation. It was also hypothesized that individual acuity profiles and speech segmentation errors may be useful in predicting the degree to which an individual will or will not benefit from adding visual speech information. Spastic dysarthric speech, associated with bilateral upper motor neuron damage and underlying muscle hypertonia, was used as the entry point into investigations with visual speech information and neurologically degraded speech.

II. METHODS

A. Participants

Fifty young, healthy adults (32 females and 18 males) aged 19 to 40 years old [mean (M) = 25.48; standard deviation (SD) = 4.19] participated in the experiment. All participants were native speakers of American English and reported no significant history of language, learning, hearing, or cognitive disabilities, and no prior contact with persons having motor speech disorders. Participants who used glasses or contact lenses in daily life were asked to also wear them during the experiment. Participants were recruited from undergraduate classes at Arizona State University and received course credit for their participation in the study.

B. Speech stimuli

One male native speaker (26 years old) of American English, with a moderate spastic dysarthria secondary to a traumatic brain injury provided the speech stimuli for the present study. Perceptual ratings from three experts in motor speech disorders confirmed that his speech was characterized by a strained-strangled and harsh vocal quality, excess and equal stress, slow rate, and imprecise articulation—all of which are considered cardinal features of spastic dysarthria, according to the Mayo Classification System (Darley *et al.*, 1969a,b). Speech intelligibility on a random selection of 20 predictive phrases was rated to be 55%,[1] according to perceptual judgments from two naive listeners who transcribed the sentences in auditory conditions.

AV speech stimuli were collected in a sound-attenuated booth with a Shure KSM 32 microphone and a Canon XA10 video camera, positioned to capture a view of the speaker's head and shoulders, against a plain black backdrop. Speech output elicited during the speech tasks was recorded digitally to a memory card at 48 kHz (16 bit sampling rate) and stored as individual .mts files. Samples included (a) 13 medial vowel targets surrounded by the initial consonant /b/ and final consonant /t/ (e.g., bait, but, bit), (b) 16 medial consonants surrounded by the vowel /a/ (e.g., aba, ata, aga), (c) 40 predictive (P) phrases (e.g., the cat is black), and (d) 40 nonpredictive (NP) phrases (e.g., amend estate approach), for a total of 109 speech stimuli files. NP phrases, taken from Liss *et al.* (1998) and modeled on those of Cutler and Butterfield (1992), consisted of phrases that were syntactically plausible but semantically anomalous to control for the contribution of semantic and contextual knowledge to intelligibility. NP phrases were all six-syllables in length, alternating strong (S) and weak (W) syllables, such that half of the phrases contained a SWSWSW phrasal stress pattern and the other half contained a WSWSWS phrasal stress pattern. These alternating syllabic stress patterns enable errors in speech segmentation (lexical boundary errors) to be interpreted relative to a perceptual strategy hypothesis, the Metrical Segmentation Strategy (MSS). The MSS predicts that, when perceiving the spoken English language, listeners will likely exploit strong syllables to determine word onsets in connected speech (Cutler and Butterfield, 1992). These phrases have been utilized in a number of studies examining the MSS hypothesis in perceptual processing of dysarthric speech (Borrie *et al.*, 2012b, 2013; Borrie *et al.*, 2012a; Liss *et al.*, 2000). Speech stimuli were presented via a PowerPoint presentation displayed on a laptop positioned directly in front of the speaker. The speaker was encouraged to use his "normal speaking" voice while reading the stimuli aloud and looking directly into the video camera. All .mts files were then opened into Adobe Premiere Pro. Audio portions of the stimuli files were imported for editing into Adobe Audition, where each file underwent noise reduction, was converted to mono, and was normalized to −3 dB. Audio portions were then imported back into Adobe Premiere Pro and realigned with their video portion. Edited speech stimuli files were then converted into .avi and .wav files using Prism Video File Converter.

C. Procedure

The experiment was conducted in a quiet room using sound-attenuating headphones (Sennheiser HD 280 pro). Participants were told that they would complete a series of phoneme acuity and word recognition tasks, in which they would be required to identify vowels, consonants, and words in phrases under two different conditions—one where they would only hear the speaker and the other where they would both hear and see the speaker. They were informed that task-specific instructions would be delivered via the computer program. This process was employed to ensure identical stimulus presentation instructions across participants. The experiment was presented via a laptop computer, using Presentation® software (Neurobehavioral Systems, 2014).

During the *acuity* task, participants were presented with all vowel and consonant stimuli under A and AV conditions (58 tokens). The presentation order of the 58 tokens was randomized for each participant. For vowels, participants were informed that they would be presented with some real words that would all start with a "b" and end with a "t" but would contain different vowels in the middle (e.g., "bait" or "bat"). For consonants, participants were informed that they would be presented with some nonsense words that would all start and end with "a" but would contain different consonants in the middle (e.g., "aba" or "aka"). Stimuli were presented one at a time, and participants were instructed to identify the presented consonant or vowel by selecting their response from a monitor display of all possible labels. Participants were afforded as much time as needed to make their selection; however, once they selected a response, the program automatically progressed to present the next stimulus. Participants received no feedback regarding the accuracy of their response.

Immediately following the acuity task, participants completed a *recognition* task in which they were presented with the P and NP phrases (160 tokens). The presentation order of the 160 tokens was randomized for each participant. Participants were informed that needed to attend closely to a series of short phrases and try to determine what was being said. They were told that all the phrases contained real English words but that some of them would not make sense (e.g., had eaten junk and train). Phrases were presented one

at a time, and following each presentation, participants were instructed to use the keyboard to type out exactly what they thought was being said. Participants were encouraged to make a guess at any words they did not recognize and to use an "X" to represent any part of a provided phrase where a guess could not be made. Participants were afforded as much time as needed to type their response and were prompted to press the Enter key to move on to the next phrase. Participants received no feedback about the accuracy of their responses. All speech stimuli were presented binaurally through headphones at a comfortable listening level of 65 dB.

D. File analysis

The presentation generated a total of eight files for each of the 50 participants: (1) vowels–A, (2) vowels–AV, (3) consonants–A, (4) consonants–AV, (5) P phrases–A, (6) P phrases–AV, (7) NP phrases–A, and (8) NP phrases–AV. Thus, the total data set consisted of 450 data files for analysis. All vowel and consonant files were analyzed for a measure of percent phonemes correct (PPC), and all phrase files were analyzed for a measure of percent words correct (PWC). Words were defined as correct if they accurately matched the intended target or differed only by tense (*-ed*) or plurality (*-s*). In addition, word substitutions between "a" and "the" were also coded as correct.

The NP phrase files were also analyzed for lexical boundary errors (LBEs). LBEs were defined as an incorrect insertion or deletion of a lexical boundary, occurring either before a strong or weak syllable; abbreviated, the errors were designated with two letters that derived from their type (insert or delete) and their location (weak or strong syllable). This analysis resulted in four types of possible errors: (1) the insertion of a lexical boundary before a strong syllable (IS); (2) the insertion of a lexical boundary before a weak syllable (IW); (3) the deletion of a lexical boundary before a strong syllable (DS); or (4) the deletion of a lexical boundary before a weak syllable (DW) (see Liss *et al.*, 2000, for more details on LBE). LBE proportions for each error type were calculated as a percent score for each condition. According to the MSS, in which listeners exploit rhythmic cues to identify word boundaries, listeners will more likely make IS and DW error types (predicted errors) as opposed to IW and DS error types (unpredicted) (Cutler and Butterfield, 1992). In addition to the LBE proportion comparisons, IS/IW and DW/DS ratios based on the sum of group errors were calculated, again for each condition. Ratio values are considered to reflect the strength of adherence to predicted MSS error patterns (see Liss *et al.*, 1998).

Finally, AV advantage, a metric that accounts for the perceptual benefit of adding visual speech information to the auditory signal, was calculated for each participant by subtracting A scores from AV scores. This was done for all stimuli, including consonants, vowels, P phrases, and NP phrases. A multiple regression analysis was then used to examine whether the dependent variable, AV advantage, could be predicted by six candidate variables, including vowels–A, vowels–AV, consonants–A, consonants–AV, LBE ratio–A, and LBE ratio–AV.[2] All variables were included in the initial model, and a backward stepwise regression was subsequently completed. Model fit was analyzed with an overall regression F statistic. Individual variables with regression coefficients significant at the 0.05 level were retained in the model.

E. Reliability

Fifty percent of all files were randomly selected and reanalyzed by the original judge (intra-judge) and by a second trained judge (inter-judge) to obtain reliability estimates for the coding of the dependent variables of PPC, PWC, LBEs, and PSR. Reliability analysis confirmed that the agreement rate between the reanalyzed data and the original data was high (all correlations $r > 0.95$).

III. RESULTS

A. Acuity analysis

Figure 1 shows the mean percent phoneme correct (PPC), separated by vowels and consonants, for acuity of phonemes in A and AV presentation conditions. Paired *t*-tests were calculated to examine whether the acuity of vowels and consonants was influenced by the presentation condition. Results showed that acuity of both vowels, $t(49) = 5.53$, $p < 0.001$, $d = 0.86$, and consonants, $t(49) = 8.07$, $p < 0.001$, $d = 1.25$, was more successful in AV conditions (vowels, M = 77.84%, SD = 11.98%; consonants, M = 67.51%, SD = 8.28%) than in A conditions (vowels, M = 69.08%, SD = 12.08%; consonants, M = 54.75%, SD = 11.88%). Thus, an AV advantage was observed for phoneme acuity in both vowels (M = 8.76%) and consonants (M = 12.77%) phrases.

B. Recognition analysis

Figure 2 shows the mean percent words correct (PWC), separated by NP phrases and P phrases, for word recognition in A and AV presentation conditions. Paired *t* tests were calculated to examine whether the recognition of words in NP phrases and P phrases was influenced by the presentation

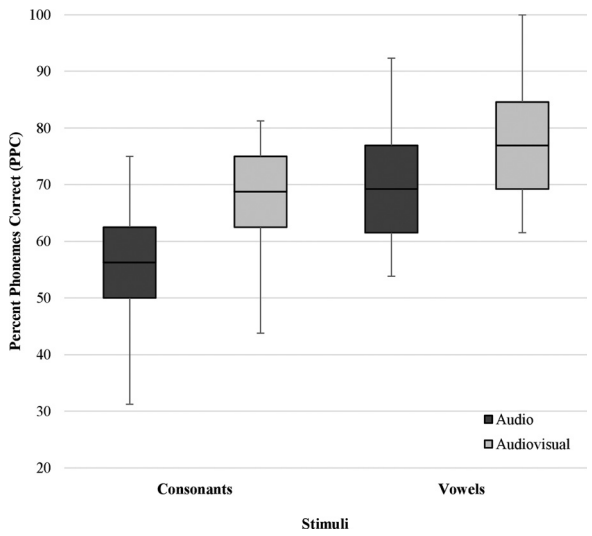

FIG. 1. Percent phonemes correct (PPC) for listeners ($n = 50$) identifying consonants and vowels under audio (A) and audiovisual (AV) conditions.

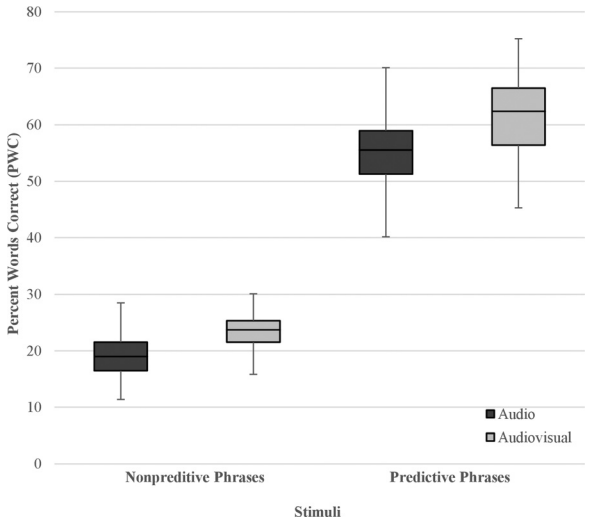

FIG. 2. Percent words correct (PWC) for listeners ($n = 50$) recognizing words in nonpredictive and predictive phrases under audio (A) and audiovisual (AV) conditions.

condition. Results showed that recognition of both NP phrases, $t(49) = 12.85$, $p < 0.001$, $d = 1.02$, and P phrases, $t(49) = 9.58$, $p < 0.001$, $d = 0.98$, was more successful in AV conditions (NP phrases, M = 23.61%, SD = 4.61%; P phrases, M = 61.61%, SD = 7.12%) than in A conditions (NP phrases, M = 19.46%, SD = 4.53%; P phrases, M = 54.49%, SD = 8.12%). Thus, an AV advantage was observed for word recognition in both NP (M = 4.15%) and P (M = 7.12%) phrases.

C. Suprasegmental analysis

An analysis of LBEs in the NP phrases is summarized in Table I. As shown, listeners exhibited a similar number of LBEs when processing dysarthric speech under A and AV conditions. A paired t test revealed no significant difference between the total number of LBEs in A and AV conditions, $t(49) = 1.02$, $p = 0.95$. Table I details the LBE category proportions for the listeners recognizing spastic dysarthric speech under A and AV conditions. Contingency tables, categorized by error type (i.e., insertion/deletion) and error location (i.e., before strong/weak syllable), were constructed using the total number of LBEs exhibited under A and AV conditions to determine whether the variables were related. A within condition, χ^2 test revealed a significant interaction between the variables of type (insert/delete) and location (strong/weak) for the errors under A conditions, χ^2 (1, N = 4) = 65.57, $p < 0.001$, and AV conditions, χ^2 (1, N = 4) = 54.16, $p < 0.001$. Error types were not evenly distributed for listeners recognizing dysarthric speech in A and AV conditions. In both conditions, erroneous lexical boundary insertions occurred more often before strong (IS) than before weak syllables (IW), and erroneous lexical boundary deletions occurred more often before weak (DW) than before strong syllables (DS). Such error patterns conform to MSS predictions (Cutler and Butterfield, 1992), suggesting that listeners utilized syllabic stress cues to segment the spastic dysarthric speech.

Table I also details the sum IS/IW ratio (number of insertions before strong syllables relative to those before weak syllables) and the sum DW/DS ratio (number of deletions before weak syllables relative to those before strong syllables) values for listeners recognizing spastic dysarthric speech under A and AV conditions. Ratio values reflect strength of adherence to predicted error pattern relative to syllabic stress—values of "1" indicate that insertions and deletions occurred as frequently before strong and weak syllables. Values greater than "1" indicate that insertions occurred more frequently before strong syllables and that deletions occurred more frequently before weak syllables.[3] Under A conditions, insertion errors occurred 1.4 times more often before strong than before weak syllables, and deletion errors occurred 1.4 times more often before weak than before strong syllables. Under AV conditions, insertion errors occurred 2.3 times more often before strong than before weak syllables, and deletion errors occurred 2.5 times more often before weak than before strong syllables. While the error patterns indicate that listeners made use of suprasegmental information in both A and AV conditions, the ratio values suggest that listeners were better able to exploit these cues when the visual speech information was added to the acoustic signal.

D. Prediction analysis

The magnitude of the effect of adding visual speech information to the auditory signal—the AV advantage/disadvantage—is depicted in Fig. 3. While group means reflect the AV advantage (vowels, 8.76%; consonants, 12.77%; NP phrases, 4.15%; P phrases, 7.12%), the box plot also illustrates substantial individual variance in the ability to benefit from adding visual speech information in processing of dysarthric speech. A multiple regression analysis was used to determine whether individual acuity profiles (detection of vowels and consonants) and reliance of suprasegmental information (IS/IW LBE ratio values) could predict the magnitude of the AV advantage/disadvantage in PP phrases.[4] The results of the regression indicated that, when entered together, variables (vowel–A, vowel–AV, consonants–A, consonants–AV, LBE ratio–A, LBE ratio–AV) did not

TABLE I. Number of lexical boundary errors shown by condition (A versus AV).[a]

Condition	Total Number	%IS	%IW	%DS	%DW	IS/IW Ratio	DW/DS Ratio
Audio (A)	2021	40.52	29.99	12.07	17.42	1.4	1.4
Audiovisual (AV)	2011	47.64	21.03	8.95	22.38	2.3	2.5

[a]Error numbers are listed for Total, Type (expressed as percentages), and Error Ratios. "IS," "DS," "IW," and "DW" refer to types of lexical boundary errors defined as insert boundary before strong syllable, delete boundary before strong syllable, insert boundary before weak syllable, and delete boundary before weak syllable, respectively.

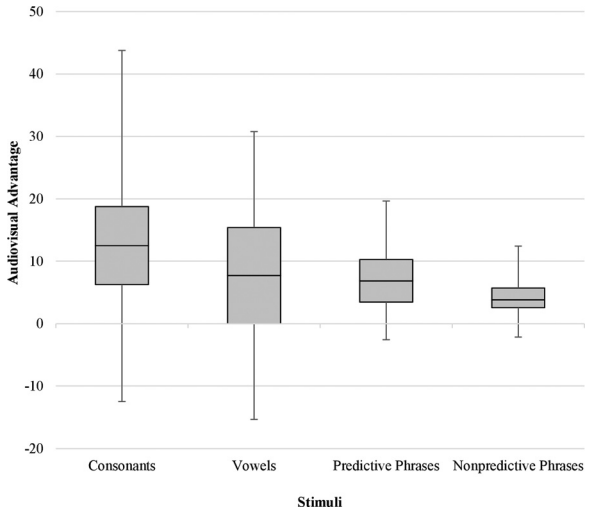

FIG. 3. Audiovisual advantage—calculated by subtracting audio (A) scores from audiovisual (AV) scores for listeners ($n = 50$) identifying consonants and vowels, and recognizing words in predictive and nonpredictive phrases.

significantly account for the variance in AV advantage/disadvantage scores in either NP phrases, $F(6, 43) = 1.54$, $p = 0.207$, $R^2 = 0.042$. However, when applying a backward elimination, and the least significant predictor at each iteration was dropped, LBE ratio–AV was, independently, a significant predictor variable for the AV advantage, $F(1, 48) = 6.09$, $p = 0.017$, $R^2 = 0.232$. That is, the model explained 23.2% of the variance in AV advantage. A positive relationship between LBE–AV ratio scores and AV advantage suggests that individuals who adhered more to the predicted error patterns—and therefore exploited syllabic stress cues to a greater degree—appeared to benefit more from adding visual speech information. Individual variation in the ability to benefit from adding visual information, however, cannot be explained by the ability to detect phonemes under A or V conditions.

IV. DISCUSSION

The purpose of this study was to investigate the influence of visual speech information on perceptual processing of dysarthric speech. This study provides evidence that for the type and severity of signal degradation in the stimuli, the addition of visual cues appeared to offer perceptual benefits to acuity and recognition of dysarthric speech. These results, and how they inform our understanding of speech processing in adverse conditions, specifically neurological signal degradation, are discussed in detail in the ensuing paragraphs.

According to the current findings, the addition of visual speech information appears to enhance perceptual processing of spastic dysarthric speech of moderate severity. Accuracy of phoneme identification and word recognition was, on average, significantly higher in AV versus A conditions. These findings are consistent with existing literature on nonmotor-impaired speech, which overwhelmingly report on the perceptual benefit of adding visual speech information when processing the signal in adverse or suboptimal listening conditions (e.g., Grant and Seitz, 1998; Ross *et al.*, 2007). An earlier study of the perception of spastic dysarthria did not consistently reflect intelligibility improvements in AV versus A conditions (Hustad and Cahill, 2003), but their study included only speakers with a mild or severe intelligibility deficit. A summary of findings from published studies reported by Hustad *et al.* (2007) suggests that listeners perceiving dysarthric speech of moderate severity seem to benefit from adding visual speech information (see Hustad *et al.*, 2007, for summary table; Hunter *et al.*, 1991; Keintz *et al.*, 2007; Garcia and Dagenais, 1998, for relevant studies). The results of the current study, therefore, add support for the idea that the perception of moderately severe dysarthric speech may be improved by providing the listener with the associated visual speech cues. While we hypothesized that in the case of dysarthric speech, this AV advantage may not transpire—that impaired visual speech information may increase cognitive load and negate any perceptual benefit expected from adding the additional input—it appears that listeners were able to utilize cues from the visual signal to constitute this sensory modality as beneficial, to at least some degree, for speech processing of this type and severity of signal degradation.

It could be hypothesized that the visual information afforded by speakers with spastic dysarthria carries less meaningful cues for recognizing spoken language than that of healthy speech. This tentative speculation arises from the current finding of relatively small accuracy gains by adding the visual signal. On average, acuity of vowels and consonants in A versus AV conditions improved by 9% and 12%, respectively, and recognition of NP and P phrases in A versus AV conditions improved by an average of 4% and 7%, respectively. So while visual speech appears to have aided recognition of dysarthric speech, the perceptual benefit afforded by the additional information is less advantageous than that expected for recognition of non-disordered speech.

The finding that word recognition was greater in AV versus A conditions raises questions regarding the types of cues listeners are extracting from the visual signal that aided in the challenging task of deciphering neurologically degraded speech. Existing literature documents that visual speech information affords important cues for the identification of segmental information (e.g., Lansing and McConkie, 1994; Marassa and Lansing, 1995; Preminger *et al.*, 1998). The current findings of improved phoneme acuity in AV versus A conditions suggests that the visual signal may offer some level of beneficial segmental information, over and above that afforded in the acoustic signal alone. However, the results of this study also suggest that listeners appear to exploit suprasegmental cues to a much greater degree when provided with the additional visual speech input.

A body of more recent work has demonstrated that when recognizing dysarthric speech under audio only conditions, listeners use syllabic stress contrast cues as a cognitive-perceptual strategy to segment the degraded acoustic information (Borrie *et al.*, 2012a; Borrie *et al.*, 2012b, 2013; Liss *et al.*, 1998, 2000). The lexical boundary error analysis performed in the current study shows that listeners employed this cognitive-perceptual strategy to decipher

spastic dysarthric speech in both A and AV conditions. However, this stress-based segmentation strategy was exploited more readily when visual input was available, as evidenced in higher LBE ratio values in AV versus A conditions. Thus, it appears that the visual cues associated with spastic dysarthric speech afforded additional suprasegmental information that listeners could reliably extract and use to inform speech segmentation decisions.

The finding that visual speech information carries useful suprasegmental information has been reported elsewhere. Jesse and McQueen (2014), for example, showed that visual speech information carried cues that signal the presence of primary lexical stress in spoken-word recognition of ones' native language. Others have demonstrated that the visual correlates of lip movement relate well with fluctuations of the second speech format (F2), providing information about signal energy within the auditory signal (Grant and Seitz, 2000). Visual speech information may extend beyond that afforded by the articulators, to include head and eyebrow movements, both of which can offer cues about stress on lexical items (Lansing and McConkie, 1999; Munhall *et al.*, 2004).

Measures of AV advantage—a metric that accounts for the perceptual benefit of adding visual speech information—reveals substantial individual variance in ones' ability to benefit from the additional sensory modality when processing dysarthric speech. For example, when visual speech information was provided alongside the auditory signal, accuracy of consonant identification improved by 44% for some individuals, whereas other individuals ($n = 9$) displayed no condition difference, and others still ($n = 4$) exhibited an accuracy decrease of up to 8%. Similar findings, although to a lesser extreme, are evident with accuracy of word recognition in both NP and P phrases. The prediction model analysis reveals that while accuracy of vowel or consonant acuity does not appear to predict whether or not someone will benefit from additional visual input with spastic dysarthric speech, reliance of stress-based segmentation strategies does. That is, individual adherence to predicted LBE patterns in AV conditions was observed to account for 23% of the variance in the AV advantage data. Given that the benefit of adding visual speech information appears to have a suprasegmental locus, it is perhaps not surprising that an individuals' aptitude to exploit syllabic stress cues can predict, to at least some degree, their ability to benefit from this additional sensory modality when processing dysarthric speech.

Reliance of stress-based segmentation strategies, however, does not explain all of the variance evident in the AV advantage data. What factors, then, are also important in predicting an individual's ability to benefit from adding visual input to process dysarthric speech? A recent study found that the size of the AV benefit for listener recognizing speech in the presence of a competing speaker was modulated by individual speech reading abilities, or visual acuity (Jesse and Janse, 2012). Other studies have documented cognitive processing speed, attentional focus control, and inhibitive control as factors regulating the size of the AV benefit when recognizing speech in adverse conditions (Humes *et al.*, 2006; Tun *et al.*, 2002; Tun and Wingfield, 1999). Further, theoretical models that predict AV recognition in nonmotor-impaired speech (e.g., Blamey *et al.*, 1989; Oden and Massaro, 1978) account for an estimation of AV *integration*, as an attempt to explain the benefit that is observed over and above that which would be expected from ones' ability to detect speech in each of the unimodal input modalities. Such factors warrant investigation with dysarthric speech. The results of this initial study, in combination with subsequent studies that explore visual acuity, cognitive abilities, AV integration skills, and employ a more comprehensive segmental level analysis, will contribute to the development of an outcome model that predicts ones' ability to process dysarthric speech in A and AV conditions. Further, extending this work to other types and severities of dysarthria will offer insight into how different segmental and suprasegmental deficits influence signal processing. Clinically, such a model is significant. A tool that helps to predict the benefits from visual speech input, based on unique listener and speaker profiles, may offer a valuable resource for identifying source of breakdown and affords potential targets for intervention to improve listener processing of dysarthric speech.

V. CONCLUSION

The results obtained in the present study provide evidence that visual speech information may aid processing of spastic dysarthric speech of moderate severity. The findings implicate an increase in the availability of both segmental and suprasegmental cues as the locus of this perceptual benefit. However, individual differences in the ability to benefit from the additional sensory input point to the need to parse out the contribution of multiple listener factors that may account for variance observed. Such findings could inform the development of a prediction-based outcome model that incorporates consideration of unique listener profiles to infer one's ability to decipher dysarthric speech in everyday communication contexts.

ACKNOWLEDGMENTS

This research was supported by the Neurological Foundation of New Zealand Grant 1119-WF and was conducted during a postdoctoral fellowship at Arizona State University. Gratitude is extended to Elizabeth Fall and Denae Faller for their assistance with data collection and analysis.

[1]Baseline intelligibility greater than 88% is considered to result in a ceiling effect, whereby visual information may offer no additional benefit (Keintz *et al.*, 2007).

[2]Given that ratio values were highly correlated, IS/IW and DW/DS ratios were combined into a single LBE ratio for A and AV conditions.

[3]The greater the positive distance from "1," the greater the strength of adherence to the predicted error pattern.

[4]PP phrases were used instead of NP phrases, due to greater variance in the associated AV advantage data (see Fig. 3).

Andersson, U., Lyxell, B., Rönnberg, J., and Spens, K. E. (**2001**). "Cognitive correlates of visual speech understanding in hearing-impaired individuals," J. Deaf Studies Deaf Educ. **6**(2), 103–116.

Bernstein, L. E., Demorest, M. E., and Tucker, P. E. (**2000**). "Speech perception without hearing," Percept. Psychophys. **62**(2), 233–252.

Blamey, P., Cowan, R., Alcantara, J., Whiteford, L., and Clark, G. (**1989**). "Speech perception using combinations of auditory, visual, and tactile information," J. Rehab. Res. Dev. **26**, 15–24.
Borrie, S., McAuliffe, M., Liss, J., Kirk, C., O'Beirne, G., and Anderson, T. (**2012a**). "Familiarisation conditions and the mechanisms that underlie improved recognition of dysarthric speech," Language Cognit. Process. **27**(7–8), 1039–1055.
Borrie, S., McAuliffe, M., Liss, J., O'Beirne, G., and Anderson, T. (**2012b**). "A follow-up investigation into the mechanisms that underlie improved recognition of dysarthric speech," J. Acoust. Soc. Am. **132**(2), EL102–EL108.
Borrie, S., McAuliffe, M., Liss, J., O'Beirne, G., and Anderson, T. (**2013**). "The role of linguistic and indexical information in improved recognition of dysarthric speech," J. Acoust. Soc. Am. **133**(1), 474–482.
Burnham, D., and Dodd, B. (**2004**). "Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect," Dev. Psychobiol. **45**(4), 204–220.
Cutler, A., and Butterfield, S. (**1992**). "Rhythmic cues to speech segmentation: Evidence from juncture misperception," J. Mem. Language **31**(2), 218–236.
Darley, F., Aronson, A., and Brown, J. (**1969a**). "Clusters of deviant speech dimensions in the dysarthrias," J. Speech Lang. Hear. Res. **12**, 462–496.
Darley, F., Aronson, A., and Brown, J. (**1969b**). "Differential diagnosis patterns of dysarthria," J. Speech Lang. Hear. Res. **12**, 246–269.
Davis, C., and Kim, J. (**2004**). "Audio-visual interactions with intact clearly audible speech," Q. J. Exp. Psychol. **57**, 1103–1121.
Dickson, S., Barbour, R. S., Brady, M., Clark, A. M., and Paton, G. (**2008**). "Patients' experiences of disruptions associated with post-stroke dysarthria," Int. J. Language Commun. Disord. **43**(2), 135–153.
Evitts, P. M., Portugal, L., Van Dine, A., and Holler, A. (**2010**). "Effects of audio-visual information on the intelligibility of alaryngeal speech," J. Commun. Disord. **43**(2), 92–104.
Fraser, S., Gagné, J.-P., Alepins, M., and Dubois, P. (**2010**). "Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues," J. Speech Lang. Hear. Res. **53**(1), 18–33.
Garcia, J. M., and Dagenais, P. (**1998**). "Dysarthric sentence intelligibility: Contribution of iconic gestures and message predictiveness," J. Speech Lang. Hear. Res. **41**, 1282–1293.
Garcia, J. M., Dagenais, P. A., Terrel, P., and Mallory, A. (**1992**). "Normal and dysarthric speech intelligibility using audio versus videotaped presentation," in *Annual Convention of the American Speech-Language-Hearing Association*, San Antonio, TX.
Grant, K., and Seitz, P. (**1998**). "Measures of auditory-visual integration in nonsense syllables and sentences," J. Acoust. Soc. Am. **104**(4), 2438–2450.
Grant, K., and Seitz, P. (**2000**). "The use of visible speech cues for improving auditory detection of spoken sentences," J. Acoust. Soc. Am. **108**(3), 1197–1208.
Grant, K., Walden, B. E., and Seitz, P. (**1998**). "Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration," J. Acoust. Soc. Am. **103**(5), 2677–2690.
Helfer, K. S. (**1997**). "Auditory and auditory-visual perception of clear and conversational speech," J. Speech Lang. Hear. Res. **40**(2), 432–443.
Humes, L., Lee, J., and Coughlin, M. (**2006**). "Auditory measures of selective and divided attention in young and older adults using single-talker competition," J. Acoust. Soc. Am. **120**, 2926–2937.
Hunter, L., Pring, T., and Martin, S. (**1991**). "The use of strategies to increase speech intelligibility in cerebral palsy: An experimental evaluation," Br. J. Disord. Commun. **26**, 163–174.
Hustad, K. C., and Cahill, M. (**2003**). "Effects of presentation mode and repeated familiarization on intelligibility of dysarthric speech," Am. J. Speech Lang. Pathol. **12**(2), 198–208.
Hustad, K. C., Dardis, C. M., and McCourt, K. A. (**2007**). "Effects of visual information on intelligibility of open and closed class words in predictable sentences produced by speakers with dysarthria," Clin. Linguist. Phon. **21**(5), 353–367.
Jesse, A., and Janse, E. (**2012**). "Audiovisual benefit for recognition of speech presented with single-talker noise in older listeners," Language Cognit. Processes **27**, 1167–1191.
Jesse, A., and Massaro, D. (**2010**). "The temporal distribution of information in audiovisual spoken-word identification," Atten Percept. Psychophys. **72**, 209–225.
Jesse, A., and McQueen, J. M. (**2014**). "Suprasegmental lexical stress cues in visual speech can guide spoken-word recognition," Q. J. Exp. Psychol. **67**(4), 793–808.
Jordan, T. R., and Sergeant, P. (**2000**). "Effects of distance on visual and audiovisual speech recognition," Language Speech **41**(1), 107–124.
Keintz, C. K., Bunton, K., and Hoit, J. D. (**2007**). "Influence of visual information on the intelligibility of dysarthric speech," Am. J. Speech Language Pathol. **16**(3), 222–234.
Klasner, E. R., and Yorkston, K. M. (**2005**). "Speech intelligibility in ALS and HD dysarthria: The everyday listener's perspective," J. Med. Speech Language Pathol. **13**(2), 127–139.
Klucharev, V., Mottonen, R., and Sams, M. (**2003**). "Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception," Cognit. Brain Res. **18**(1), 65–75.
Lansing, C. R., and McConkie, G. W. (**1994**). "A new method for speech reading research: Eyetracking observers' eye movements," J. Acad. Rehab. Audiol. **27**, 25–43.
Lansing, C. R., and McConkie, G. W. (**1999**). "Attention to facial regions in segmental and prosodic visual speech perception tasks," J. Speech Lang. Hear. Res. **42**(3), 526–539.
Liss, J., Spitzer, S., Caviness, J., Adler, C., and Edwards, B. (**1998**). "Syllabic strength and lexical boundary decisions in the perception of hypokinetic dysarthric speech," J. Acoust. Soc. Am. **104**(4), 2457–2466.
Liss, J., Spitzer, S., Caviness, J., Adler, C., and Edwards, B. (**2000**). "Lexical boundary error analysis in hypokinetic and ataxic dysarthria," J. Acoust. Soc. Am. **107**(6), 3415–3424.
Marassa, L., and Lansing, C. R. (**1995**). "Visual word recognition in two facial motion conditions: Full-face versus lips-plus-mandible," J. Speech Hear. Res. **38**, 1387–1394.
Mattys, S., Seymour, F., Attwood, A., and Munafò, M. (**2013**). "Effects of acute anxiety induction on speech perception: Are anxious listeners distracted listeners?," Psychol. Sci. **24**(8), 1606–1608.
Mattys, S., White, L., and Melhorn, J. F. (**2005**). "Integration of multiple speech segmentation cues: A hierarchical framework," J. Exp. Psychol. Gen. **134**(4), 477–500.
McGurk, H., and MacDonald, J. (**1976**). "Hearing lips and seeing voices," Nature **264**, 746–748.
Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., and Vatikiotis-Bateson, E. (**2004**). "Visual prosody and speech intelligibility: Head movement improves auditory speech perception," Psychol. Sci. **15**(2), 133–137.
Neurobehavioral Systems (**2014**). "Presentation® software (version 0.70) [computer program]," www.neurobs.com (Last viewed January 5, 2014).
Oden, G., and Massaro, D. (**1978**). "Integration of featural information in speech perception," Psychol. Rev. **85**, 172–191.
Preminger, J., Lin, H., Payen, M., and Levitt, H. (**1998**). "Selective masking in speechreading," J. Speech Hear. Res. **41**, 564–575.
Reynolds, M., Fucci, D., and Bond, Z. (**1997**). "Effect of visual cuing on synthetic speech intelligibility: A comparison of native and nonnative speakers of English," Percept. Motor Skills **84**, 695–698.
Ross, L., Saint-Amour, D., Leavitt, V., Javitt, D., and Foze, J. (**2007**). "Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments," Cereb. Cortex **17**, 1147–1153.
Rudner, M., and Rönnberg, J. (**2004**). "Perceptual saliency in the visual channel enhances explicit language processing," Iranian Audiol. **3**, 16–26.
Schumeyer, R. P., and Barner, K. E. (**1996**). "The effect of visual information on word initial consonant perception of dysarthric speech," in *Proceeding of Fourth International Conference on Spoken Language Processing* (IEEE, New York), Vol. 1, pp. 46–49.
Sumby, W., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**, 212–215.
Tun, P., O'Kane, G., and Wingfield, A. (**2002**). "Distraction by competing speech in young and older adult listeners," Psychol. Aging **17**, 453–467.
Tun, P., and Wingfield, A. (**1999**). "One voice too many: Adult age differences in language processing with different types of distracting sounds," J. Gerontol. B **54**, P317–P327.
Weismer, G. (**2006**). "Philosophy of research in motor speech disorders," Clin. Linguist. Phon. **20**(5), 315–349.