NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS
DEPARTMENT OF PHILOSOPHY AND HISTORY OF SCIENCE
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS
DEPARTMENT OF PSYCHOLOGY
DEPARTMENT OF PHILOLOGY

Lydia-Kalliopi Liapi

A.M. 15M08

Behavioral validation of inverse effectiveness and the interactions with the temporal rule of multisensory integration

Thesis submitted
In partial fulfillment of the requirements for the Masters degree by the
Interdepartmental Graduate Program in Cognitive Science

Advisory committee:
Konstantinos Moutoussis, Associate Professor, National and Kapodistrian University of Athens
Argiro Vatakis, Researcher, Cognitive Systems Research Institute
Petros Maragos, Professor, National Technical University of Athens

Athens, Greece
January 2018

The thesis is approved

Konstantinos Moutoussis.....

Argiro Vatakis.....

Petros Maragos .....

Abstract

One of the most fundamental principles of multisensory integration is that of inverse effectiveness (IE), whereby multisensory gain is maximized when the unisensory components of an event evoke weak neuronal responses. Behavioral investigations of IE are often limited to the use of speech stimuli and focus mostly on the artificial degradation of the auditory stream. Here, we first examined IE behaviorally by implementing naturalistic degradation in both streams of an audiovisual speech signal. We used auditory, visual, and audiovisual presentations of three syllables (/ba/, /fa/, /tha/) at different levels of noise and noise combinations, while participants performed a syllable-identification task. Multisensory gain was calculated with four different indices. For the Contrast index, gain was minimized when the auditory stream was of the highest noise independent of visual noise for the syllable /ba/, while no differences were noted for /fa/ and /tha/. For the Absolute Difference (in %) index, combinations of high auditory and low visual noise levels led to a maximum gain for /ba/ and /fa/, while for /tha/, maximum gain was obtained when both streams were of low noise. For the Multisensory Integration and Absolute Difference indices, gain was minimized when the auditory stream was of the highest noise for /ba/ but not for /fa/ and /tha/. Thus, the IE effect was verified only for some of the indices and stimuli utilized but not for all, thus adding to the discussion of the validity of the IE at a behavioral level. Subsequently, we examined how IE interacts with the temporal rule (i.e., signals presented close in time are more likely to be integrated) of multisensory integration. In Exp. 2, we presented the audiovisual stimuli that led to the highest and lowest multisensory gain from Exp. 1 at different stimulus onset asynchronies. Participants' temporal order judgments showed a higher asynchrony tolerance when high gain stimuli were presented as compared to low gain audiovisual pairs. Taken together, these findings suggest that the magnitude of multisensory gain and the width of the temporal window of integration interact as a function of the effectiveness levels of the auditory and visual streams of the speech event.

Keywords: inverse effectiveness; temporal window; multisensory integration; synchrony perception; audiovisual speech.

Acknowledgements

I would like to sincerely thank Prof. Konstantinos Moutoussis, Dr. Argiro Vatakis and Prof. Petros Maragos for their guidance and excellent collaboration throughout the process of writing this thesis. My special thanks go to Dr. Argiro Vatakis for her constant support, encouragement and patience. Her assistance, her scientific enthusiasm and her knowledge made this study possible and it was truly a pleasure working with her. I would also like to thank Vassilis for all the help he provided by opening the lab, my friends, for making this process much more manageable, my brother and my parents, for their limitless support and encouragement. Finally, I would like to thank everyone who participated in this study for their contribution.

Contents

List of tables

List of figures

Behavioral validation of inverse effectiveness and the interactions with the temporal rule of multisensory integration

1. Introduction

The representation of the external world is accomplished through information captured by our sensory systems. This information must be integrated (i.e., multisensory integration) in order to perceive the multiple, distinct, and unified events that occur in the physical world (e.g., Nahanni, 2014; Talsma, Senkowski, Soto-Faraco, & Woldorff, 2010; Tsilionis & Vatakis, 2014; Vatakis, 2013; Vatakis, Ghazanfar, & Spence, 2008). Extensive research on multisensory integration has led to the formation of three fundamental principles governing this process, those of the: spatial rule, temporal rule, and principle of inverse effectiveness. Each of these principles states, respectively, that multisensory gain is maximized when the unisensory stimulus components occur close in space, in time, and are weakly effective (Holmes & Spence, 2005). These findings were based on single cell recordings in cats' superior colliculus (e.g., Meredith, Nemitz, & Stein, 1987; Meredith & Stein, 1983; Meredith & Stein, 1986), whereby, during exposure to a multisensory stimulus, the firing rate of a single superior colliculus neuron could be lower than, equal to, or greater than the sum of the firing rate of the same neuron during exposure to the unisensory components of the same underlying event (subadditive, additive, or superadditive responses, respectively). Multisensory gain (or multisensory enhancement) refers to the latter case (i.e., superadditive response) and reflects the relative increase of a neuron's firing rate from unisensory to multisensory conditions (Stanford & Stein, 2007; Stein & Stanford, 2008). The principle of inverse effectiveness -the focus of this thesis- has been extended from single cell recordings in animals (e.g., Meredith & Stein, 1987) to neuroimaging (e.g., James, Stevenson, & Kim, 2009; Nath & Beauchamp, 2011; Stevenson & James, 2009; Stevenson et al., 2009) and electrophysiological studies with humans (e.g., Stevenson et al., 2012). However, its behavioral validation remains a matter of debate due to the lack of consensus regarding the levels of noise that lead to the maximization of multisensory gain (e.g., Diederich & Colonius, 2004; Erber, 1969; Stevenson et al., 2012; Sumby & Pollack, 1954; but see Holmes, 2007; Ross et al., 2007; Ma et al., 2009).

Behavioral investigations of inverse effectiveness have mainly focused on the use of audiovisual (AV) speech and the implementation of noise mainly in the auditory channel, while maintaining the visual signal clear (e.g., Erber, 1969; Ross et al., 2007; Ma et al., 2009; Sumby & Pollack, 1954). Numerous studies following this methodological approach have demonstrated findings that were in accordance with the principle of inverse effectiveness (e.g., Bernstein, Auer Jr., & Takayanagi, 2004; Grant & Seitz, 2000; Erber, 1969, 1971; Ewersten & Nielsen, 1971; O' Neil, 1954; Sumby & Pollack, 1954). For example, in the seminal study by Sumby and Pollack (1954), participants were asked to identify words presented at various levels of auditory noise. They manipulated the signal-to-noise ratio (SNR) by implementing a standard level of noise in the speech signal, produced with the use of a gas tube source, and by varying the levels of intensity of the clear auditory signal. The presentation of each word was either auditory or AV and the researchers tested for any behavioural enhancement that could be attributed to the simultaneous presentation of the visual component of the speech stimuli. They also examined whether this enhancement would vary depending on the noise level of the auditory channel. Their findings were in accordance with the predictions of inverse effectiveness, since the enhancement in participants' accuracy scores was more pronounced as the noise levels in the auditory channel increased. Accordingly, O'Neil (1954) asked participants to identify speech stimuli (vowels, consonants, words, or phrases), presented visually, aurally, or audiovisually. In the two latter cases, the auditory component was presented at one out of four different SNRs. O'Neil found that the contribution of the visual signal led to enhanced performance irrespectively of the SNR level, but this enhancement was more pronounced when stimuli were presented at lower (i.e., high level of noise) as compared to higher SNRs, a finding that is in line with the predictions of inverse effectiveness (see also Bernstein, Auer Jr., & Takayanagi, 2004; Grant & Seitz, 2000; Erber, 1969, 1971; Ewersten & Nielsen, 1971). Taken together, these findings suggest that for AV speech stimuli, the levels of multisensory gain increase as the auditory noise levels increase.

A number of more recent studies examining speech perception in noise have demonstrated that maximum multisensory gain levels are reached at certain, intermediate rather than high levels of noise. For instance, in a study conducted by Ross and his colleagues (2007), participants were asked to identify monosyllabic words presented at various SNRs. The presentation of syllables was auditory or AV and the

magnitude of the multisensory gain was calculated with the use of several mathematical indices based on participants' accuracy scores across experimental conditions. Results showed that the maximum levels of multisensory gain were obtained for different levels of auditory noise, depending on the specific index used. For example, the % Gain index showed maximum gain for maximum levels of noise, in accordance with the predictions of the principle of inverse effectiveness, while the Absolute Gain (in %) index showed maximum gain for intermediate levels of noise (see also Ma et al., 2009). The latter finding relates to the phenomenon of stochastic resonance that occurs in nonlinear systems and predicts signal detection facilitation when a specific level of noise is implemented (Moss, Ward, & Sannita, 2004). Furthermore, Nahanni (2014) conducted an experiment whereby participants had to identify monosyllabic words presented at six SNRs in auditory and AV presentations. Visual presentations were also included but the visual signal was not degraded. Six different gain indices were used with different outcomes. The Difference Score and Integration Enhancement metrics resulted in the gain peaking at two low SNRs, but not at the lowest; the Visual Enhancement and Normalized Integration Enhancement metrics indicated that gain increased as the auditory signal became clearer (i.e., at higher SNRs), and, finally, the % Gain for Auditory Enhancement and Integration Enhancement metrics showed that gain decreased as the SNRs increased, with the former reaching its peak at the lowest SNR, in accordance with the predictions of the principle of inverse effectiveness. Together, these findings emphasize the fact that different gain indices for the same data can lead to different gain patterns, thus, highlighting the potential source of inconsistency in previous inverse effectiveness studies.

The manipulation of the visual modality has been neglected in most studies of inverse effectiveness (e.g., Erber, 1969, 1971; Ma et al., 2009; Nahanni, 2014; O' Neil, 1954; Ross et al., 2007; Sumby & Pollack, 1954), given the notion that the auditory modality is the most informative sensory channel when it comes to speech perception. However, speech perception is highly affected by the available, concurrent visual cues made during speech production (Callan, Callan, Kroos, & Vatikiotis-Bateson, 2001; Massaro, 2004; Massaro & Cohen, 1983, McGurk, & McDonald, 1976; Vatakis, Maragos, Rodomagoulakis, & Spence, 2012). The contribution of the visual modality to speech perception was firstly pointed out by perceptual phenomena, such as the McGurk effect, which was originally observed when participants were presented with the utterance of

the syllable /ba/, while synchronously watching the talker articulating the syllable /ga/. In these cases, the illusory percept of being presented with the syllable /da/ was evoked (McGurk, & McDonald, 1976), demonstrating how viewing a speaker's articulatory movements can affect the final percept of an AV speech event. Additionally, Massaro and Cohen (1983) conducted an experiment in which they used auditory and visual presentations of the syllables /ba/ and /da/, with five levels of visual and auditory ambiguity, and every AV combination of ambiguity levels between the two sensory streams. Participants' identification scores were examined in order to determine the amount of one stream's influence on the final percept, as a function of the ambiguity of the other sensory stream. By analyzing individual data for AV presentations, Massaro and Cohen found that the amount of one sensory stream's influence on the final percept is greater when the information provided by the other stream is ambiguous or neutral. The interplay between the available auditory and visual cues on the final percept of an AV speech event has been characterized as the complementarity of vision and audition in speech perception (Massaro, 2010, Chap.10). This interplay suggests that both modalities affect the final perceptual outcome in a dynamic fashion. That is, the information provided by each modality is weighted, depending on the available information provided by the other modality. Thus, the visual modality, along with the auditory, shapes the final perceptual outcome for speech stimuli and the manipulation of both sensory channels should be applied to examine speech perception in noise.

The degradation of the auditory speech stream is commonly introduced with the use of various SNR levels in studies conducted to investigate the perception of speech in noise (e.g., Bernstein, Auer Jr., & Takayanagi, 2004; Erber, 1969; Ewersten & Nielsen, 1971; Nahanni, 2014; Ross et al., 2007, 2011). SNRs are manipulated by either maintaining the auditory signal clear, while implementing various noise manipulations in the auditory stimulus, or by implementing a standard level of noise as the intensity or sound pressure of the auditory signal is manipulated. This noise-introducing technique is widely accepted for the manipulation of intelligibility of speech stimuli (Nahanni, 2014) and researchers often implement artificial noise in the auditory stream in order to test their hypotheses. For example, Sumby and Pollack (1954) used a gas tube source to introduce a standard level of noise in the speech stream, Erber (1971) used ambient noise from a video tape recorder, Erber (1969), Ewersten and Nielsen (1971), and Bernstein, Auer, and Takanayagi (2004) used simple white noise to mask the auditory

signal, while Ross et al. (2007, 2011) and Nahanni (2014) used a standard level of pink noise for the degradation of the auditory channel. The implementation of artificial noise, however, is a potential methodological drawback due to the lack of naturalism of the final auditory stimulus. The findings reported above might not be easily generalized in natural environments and, therefore, a more naturalistic approach should be developed for the investigation of speech perception in noise.

Given the degradation of mainly one of the sensory streams of an event, the artificial manipulation of SNRs, and the conflicting gain data in the inverse effectiveness literature, in Experiment 1 we examined whether the principle of inverse effectiveness would be demonstrated behaviorally with the use of AV speech stimuli that were manipulated in both sensory streams by a more naturalistic approach. More specifically, we used visual, auditory, and AV presentations of 3 syllables that were manipulated in an equivalent manner, by implementing different levels of naturalistic noise in both the auditory and visual streams of an AV speech signal. Visual noise was introduced by manipulating the visual saliency of the articulatory movements made by the talker, while auditory noise was introduced with the use of a vocoder. Both manipulations were selected in order to provide the naturalistic degradation of each sensory stream (see Section 2.1.3). While the principle of inverse effectiveness is well established on a neuronal level, the aim of Experiment 1 was to examine whether it would be demonstrated behaviorally with the use of AV speech stimuli. Due to the highly conflicting findings in studies that have examined the effect of noise in speech perception (e.g., Erber, 1969, 1971; O' Neil, 1954; Sumby & Pollack, 1954, but see Ma et al., 2009; Nahanni, 2014; Ross et al., 2007) we aimed to determine whether high or intermediate levels of noise implemented in the two sensory channels would lead to maximization of multisensory gain levels regarding participants' accuracy in identifying AV speech stimuli.

2. Methods

2.1 Experiment 1

2.1.1 Participants

Twenty-one female undergraduate students, aged between 19-51 years (Mean age = 22.2 years) took part in the experiment and received course credit for their

participation. All of them were naïve as to the purpose of the experiment and reported having normal hearing and normal or corrected-to-normal vision. The duration of the experiment was approximately 50 minutes.

2.1.2 Apparatus

The experiment took place in a dark, quiet room. The visual stimuli were presented on a 15-inch, CRT monitor (1600 x 1200 pixel resolution, 60 Hz refresh rate), while the auditory stimuli were delivered via loudspeakers (Creative Inspire 265), placed on both sides of the monitor. The experiment was programmed using Presentation (version 17.0, Neurobehavioral Systems Inc.). The participants responded using a standard computer keyboard, by pressing 1 out of 3 specified keys in every trial with their right hand.

2.1.3 Stimuli

The speech stimuli were brief videos (2160 ms in duration) with visual only (VO), auditory only (AO), or AV presentations of the syllables /ba/, /fa/, and /tha/ at various levels of visual and auditory noise. The video processing was conducted using Adobe Premiere Pro CC 2015. Each video file consisted of 66 frames (frame size = 720 x 576 pixels, depth = 24 bits) and each recording had a sampling frequency of 48000 Hz. The volume was approximately 65 dB, in relation to the participants' seating position.

The consonants /b/, /f/, and /th/ were selected due to their high levels of visual saliency (i.e., the articulatory movements that are made as these consonants are pronounced; Vatakis et al., 2012). The auditory and visual physical characteristics for each consonant can be assessed by their manner and place of articulation, respectively (i.e., /b/ is a stop consonant, /f/ and /th/ are both fricatives in terms of the manner of articulation, whereas /b/ is bilabial, /f/ is labiodental, and /th/ is dental in terms of the place of articulation). The open vowel /a/ was selected because of the high contrast in the visible movements made for its articulation, especially following the articulation of a consonant (Vatakis, et al., 2012).

Initially, AV speech recordings from a professional actress were captured, as the actress pronounced (both uttered and articulated) each of the 3 syllables (i.e., /ba/, /fa/, and /tha/). Visual noise, in this case, was introduced by having the actress articulate each syllable with varying levels of saliency of her articulation. With this

manipulation however, we could not parameterize the levels of visual noise and, consequently, we could not make sure that the actress's articulatory movements were equivalent at each visual noise level between syllables. Thus, the visual components of these recordings were not used any further, while the auditory recordings were used to model a 3D animated avatar using the Synface Technology (Synface, n.d.; an application originally developed to provide real-time lip-reading support to people with hearing impairments during telephone conversations). The final experimental videos utilized consisted of this animated 3D avatar and the original speech recordings from the professional actress.

The visual noise was implemented by manipulating the levels of jaw opening during the articulation of each syllable's vowels (ranging from 100-20%, with 100% of the movement at the clearest visual presentation and 20% of the movement at the most ambiguous visual presentation of the syllables) and the levels of lip closure, labiodental closure, and tongue tip motion during the articulation of each syllable's consonants (ranging from 100-80%, with 100% of the movement at the clearest visual presentation and 80% of the movement at the most ambiguous visual presentation of the syllables) by the avatar. This manipulation resulted in 6 different levels of saliency for the articulatory movements made for each syllable (i.e., 6 levels of visual noise for each syllable).

The auditory noise was introduced with the use of a vocoder, while the actress uttered the 3 syllables. It is common for studies on the topic of inverse effectiveness to implement white or pink noise filters in the auditory stream (e.g., Ross et al., 2007). However, with the use of the vocoder the final stimulus produced is based on the characteristics of the speech signal (such as amplitude, frequency, pitch; Carney, 2012) and these spectral characteristics are replaced with noise (Alexanderson & Beskow, 2014). For this reason, the vocoder was used in order to create more naturalistic conditions of auditory noise. Twelve levels of auditory noise were created and the syllables were recorded at a different frequency band for each noise level.

We conducted two pilot studies (with 10 and 11 participants for the first and second pilot study, respectively), in order to select the least and most behaviorally ambiguous levels of auditory and visual noise that would be included in Experiment 1. In the first pilot study, participants were presented with 12 AO stimuli for each syllable

(each representing a different level of auditory noise; 36 AO stimuli in total) and 6 VO stimuli for every syllable (each representing a different level of visual noise; 18 VO stimuli in total) and were asked to identify the presented syllable in each trial. The levels of auditory and visual noise that we used in Experiment 1 were selected by examining the differences in participants' accuracy scores, across the various levels of noise, separately for each stimulus stream. We also made sure that the physical levels of noise, implemented in each sensory channel, reflected participants' accuracy scores (i.e., the clearest stimuli led to the highest accuracy scores, the 2nd clearest led to the 2nd highest accuracy score, and accordingly for the 3rd and 4th level of noise). The second pilot study was conducted in order to verify the results of the first pilot study. Based on the results of the two pilot studies, 4 auditory and 4 visual noise levels were selected, ranging between 1 and 4, with 1 being the clearest and 4 being the most ambiguous presentation for each sensory stream.

The final AV videos were created by pairing all possible combinations of auditory and visual noise for every syllable. The peaks in the auditory and visual signals were used as a reference to synchronize the two streams of the AV speech event.

2.1.4 Design

Participants were seated at approximately 60 cm distance from the monitor. A practice block was performed prior to the beginning of the experiment, to familiarize participants with the experimental procedure and to make sure that the instructions we provided were clear. The practice block consisted of one randomized loop that included the AO, VO, and AV stimuli with the lowest levels of noise (i.e., the clearest stimuli; 9 stimuli in total were presented). The practice block was followed by 4 experimental blocks; two unisensory blocks that consisted of VO and AO stimuli, and two multisensory blocks that consisted of AV stimuli. Given that behavioral (as well as neural) responses to multisensory stimuli are characterized by increased levels of variability (Baum, Colonius, Thelen, Micheli, & Wallace, 2016), we made the distinction between unisensory and multisensory blocks in order to minimize the magnitude of trial-to-trial variability in AV conditions and potentially attentional shifting effects (Lukas, Phillipp, & Koch, 2010) within a block. The order of blocks was randomized for each participant.

For the unisensory blocks, the experimental conditions were 24 in total (i.e., 4 noise levels for each of the two sensory channels and for each of the three syllables). There were 4 repetitions within each unisensory block, resulting in a total of 96 experimental trials. Each unisensory block lasted approximately 8 minutes. For the multisensory blocks, the experimental conditions were 48 in total (i.e., all possible combinations between the four levels of auditory and visual noise for each of the three syllables). There were 4 repetitions within each multisensory block, resulting in a total of 192 experimental trials. Each multisensory block lasted approximately 15 minutes. In total, the experiment consisted of 576 experimental trials.

2.1.5 Procedure

Participants received verbal instructions prior to the beginning of the experiment. Specifically, they were informed that three syllables will be presented to them during the course of the experiment and that their task was to identify the presented syllable in each trial by pressing key 1, 2, and 3 when the syllables /ba/, /fa/, and /tha/ were presented, respectively. They were also informed that within a single block, the presentation could either be unisensory (i.e., AO or VO) or multisensory (i.e., AV). Prior to the beginning of each block, they were informed about the type of the impending block (i.e., unisensory or AV block). Finally, they were asked to remain focused on the task and provide responses that were as accurate as possible. The task was self-paced and after the presentation of each stimulus, participants had to provide a response in order to proceed to the next trial. Participants were allowed to take breaks between the experimental blocks.

2.1.6 Results and Discussion

We first obtained and averaged data from each participant separately across experimental conditions. Based on the average scores obtained in each experimental condition for each participant, we analyzed participants' accuracy scores and calculated the levels of multisensory gain, with the use of the appropriate mathematical indices (see section 2.6.1.2). For every analysis reported in this section, Bonferroni-corrected t tests (where $p < .05$ prior to correction) were used for all post hoc comparisons.

2.1.6.1 Accuracy scores

First, we analyzed participants' accuracy scores over the various experimental conditions by conducting three separate analyses. The first analysis was conducted on the basis that the auditory modality is thought to be the most dominant modality in speech perception (e.g., Ross et al., 2007). On this basis, apart from examining participants' performance in VO and AO conditions, we also examined how participants' accuracy scores in AV conditions would vary as a function of auditory noise. Therefore, we analyzed the data obtained from every unisensory condition, but for AV only the combinations that consisted of the lowest level of visual noise paired with every level of auditory noise. We performed a repeated-measures analysis of variance ANOVA with the factors of: Syllable (3 levels: /ba/, /fa/, /tha/), Modality (3 levels: VO, AO, AV), and Noise (4 levels: 1, 2, 3, 4). This analysis revealed a main effect of Syllable [$F(2, 40) = 27.814$, $p < 0.001$, $\eta^2_p = 0.582$], with the accuracy scores for the syllable /fa/ being significantly higher (M = 0.850) than those for /ba/ and /tha/ (M = 0.787 and 0.707, respectively) and /ba/ higher than /tha/. A main effect of Modality was also found [$F(2, 40) = 412.593$, $p < 0.001$, $\eta^2_p = 0.954$], with accuracy being higher for AV presentations (M = 0.950) as compared to VO or AO presentations (M = 0.824 and 0.559, respectively) and for VO as compared to AO. Finally, a main effect of Noise was obtained [$F(2.053, 41.059) = 251.629$, $p < 0.001$, $\eta^2_p = 0.926$], with participants performing better at the 1st and 2nd level of noise (i.e., clearest levels; M = 0.885 and 0.875, respectively) as compared to the 3rd and 4th level of noise (M = 0.790 and 0.574, respectively).

A significant three-way interaction between Syllable, Modality, and Noise was also obtained [$F(4.300, 85.996) = 22.793$, $p < 0.001$, $\eta^2_p = 0.533$] with /ba/, at the 3rd and 4th level of noise, being more accurately detected for AV presentations (M = 0.976 and 0.825 for the 3rd and 4th level of noise, respectively) as compared to VO or AO presentations (M = 0.838 and 0.568 for the 3rd, and 0.222 and 0.090 for the 4th level of noise, respectively) and for VO as compared to AO (see Figure 1A). For the syllable /fa/, at the 1st, 2nd, and 3rd level of noise, performance was significantly worse for AO presentations (M = 0.870, 0.849, and 0.617 for the 1st, 2nd and 3rd level, respectively) as compared to AV or VO presentations (M = 0.986 and 0.981 for the 1st level, 0.973 and 0.976 for the 2nd level, and 0.973 and 0.924 for the 3rd level, respectively). At the 4th level of noise, accuracy was lower for AO presentations (M = 0.400) as compared to AV

or VO presentations (M = 0.976 and 0.671, respectively) and for VO as compared to AV (see Figure 1B). For the syllable /tha/, at the 1st, 2nd, and 3rd level of noise, performance was worse for AO presentations (M = 0.233, 0.262, and 0.370 for the 1st, 2nd, and 3rd level, respectively) as compared to AV or VO presentations (M = 0.957 and 0.944 for the 1st level, 0.944 and 0.946 for the 2nd level, and 0.976 and 0.865 for the 3rd level, respectively). At the 4th level of noise, accuracy for VO and AO presentations was lower (M = 0.544 and 0.476, respectively) than for AV presentations (M = 0.962; see Figure 1C). We also found that the interaction between Syllable and Modality was significant [$F(2.694, 53.885) = 28.800$, $p < 0.001$, $\eta^2_p = 0.590$], as was the interaction between Syllable and Noise [$F(3.461, 69.223) = 69.003$, $p < 0.001$, $\eta^2_p = 0.775$], and Modality and Noise [$F(3.611, 72.226) = 51.614$, $p < 0.001$, $\eta^2_p = 0.721$].

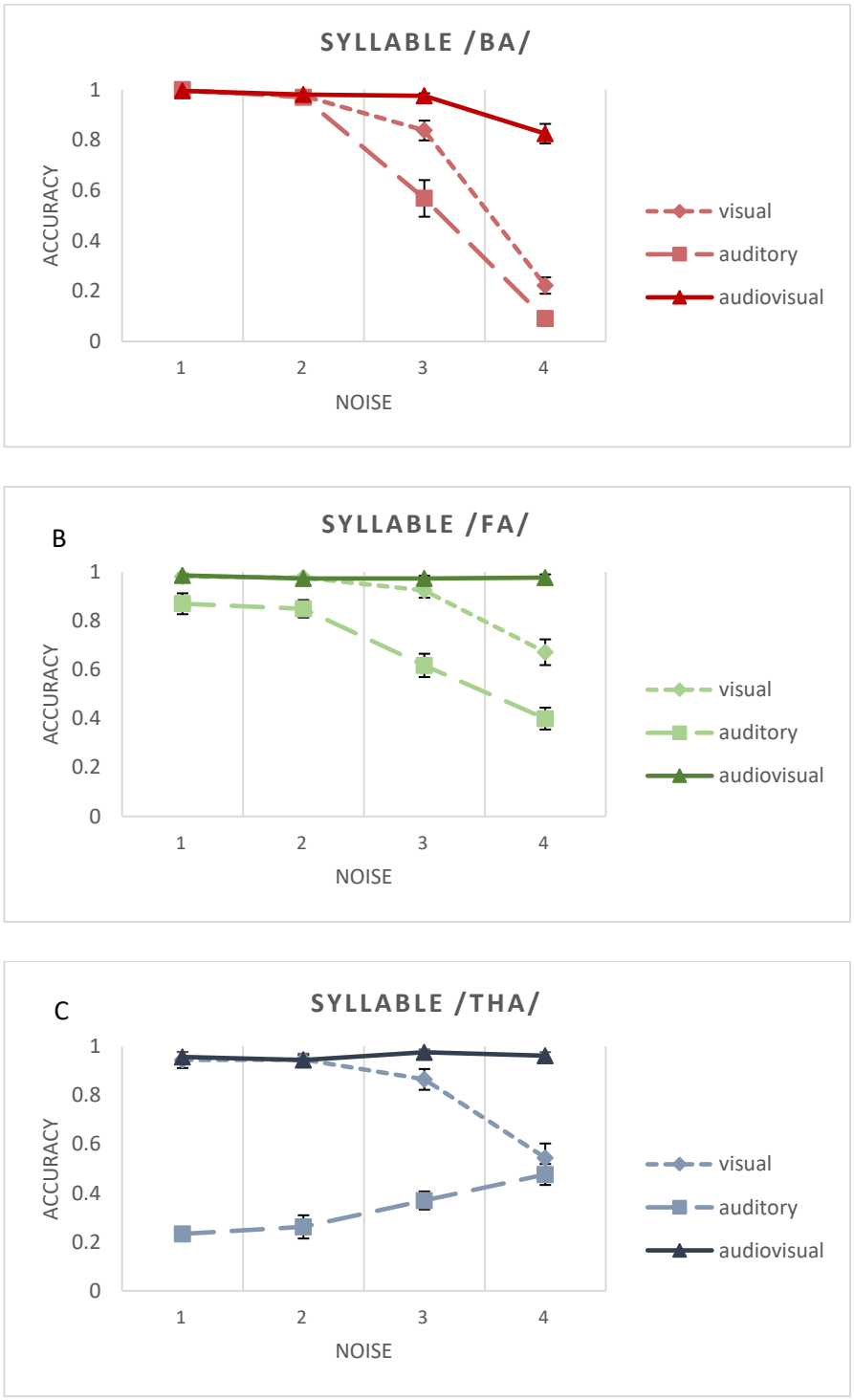

Figure 1. Mean accuracy scores, for the visual only (VO), auditory only (AO), and audiovisual (AV) conditions with the clearest visual component. Accuracy scores are averaged across participants for each modality and plotted as a function of noise for the syllables A) /ba/, B) /fa/, and C) /tha/. Error bars represent the standard error of the mean.

In order to compare the pattern of results reported above with those obtained when the visual channel was degraded in the AV conditions, we conducted a similar analysis with the AV combinations consisting of the lowest level of auditory noise paired with each level of visual noise. This analysis revealed a main effect of Syllable [$F(2, 40) = 36.580$, $p < 0.001$, $\eta^2_p = 0.647$], with the accuracy scores for /tha/ being significantly lower (M = 0.669) than those for the syllables /ba/ and /fa/ (M = 0.800 and 0.842, respectively). A significant main effect of Modality was also obtained [$F(2, 40) = 363.810$, $p < 0.001$, $\eta^2_p = 0.948$], with accuracy being higher for AV presentations (M = 0.929) as compared to VO or AO presentations (M = 0.824 and 0.559, respectively) and for VO as compared to AO. Finally, a main effect of Noise was obtained [$F(1.935, 38.709) = 266.870$, $p < 0.001$, $\eta^2_p = 0.930$], with accuracy being significantly higher at the 1st and 2nd level of noise (M = 0.885 and 0.880, respectively) as compared to the scores obtained at the 3rd and 4th level of noise (M = 0.782 and 0.536, respectively).

A significant three-way interaction between Syllable, Modality, and Noise was obtained [$F(4.834, 96.680) = 39.971$, $p < 0.001$, $\eta^2_p = 0.667$] with the syllable /ba/, at the 3rd and 4th level of noise, being more accurately detected for AV presentations (M = 0.987 and 0.996, respectively) as compared to VO or AO presentations (M = 0.838 and 0.568 for the 3rd level, and 0.222 and 0.090 for the 4th level, respectively), and for VO as compared to AO (see Figure 2A). For the syllable /fa/, at the 1st, 2nd, and 3rd level of noise, performance was worse for AO presentations (M = 0.870, 0.849 and 0.617, for the 1st, 2nd, and 3rd level of noise, respectively) as compared to AV or VO presentations (M = 0.986 and 0.981 for the 1st level, 0.990 and 0.976 for the 2nd level, and 0.990 and 0.924 for the 3rd level, respectively). At the 4th level of noise, performance for AO presentations was worse (M = 0.400) than for AV or VO presentations (M = 0.851 and 0.671, respectively) and for VO as compared to AV (see Figure 2B). For the syllable /tha/, at the 1st, 2nd, and 3rd level of noise, accuracy was lower for AO presentations (M = 0.233, 0.262 and 0.370, respectively) as compared to AV or VO presentations (M = 0.957 and 0.944 for the 1st level, 0.976 and 0.946 for the 2nd level, and 0.876 and 0.865 for the 3rd level, respectively; see Figure 2C). We also found significant interactions between Syllable and Modality [$F(2.671, 53.421) = 24.126$, $p < 0.001$, $\eta^2_p = 0.547$], Syllable and Noise [$F(3.047, 60.948) = 27.528$, $p < 0.001$, $\eta^2_p = 0.579$], and Modality and Noise

[$F(3.524, 70.484) = 26.724$, $p < 0.001$, $\eta^2_p = 0.572$]. The three-way interaction between these factors reflects these two-way interactions.

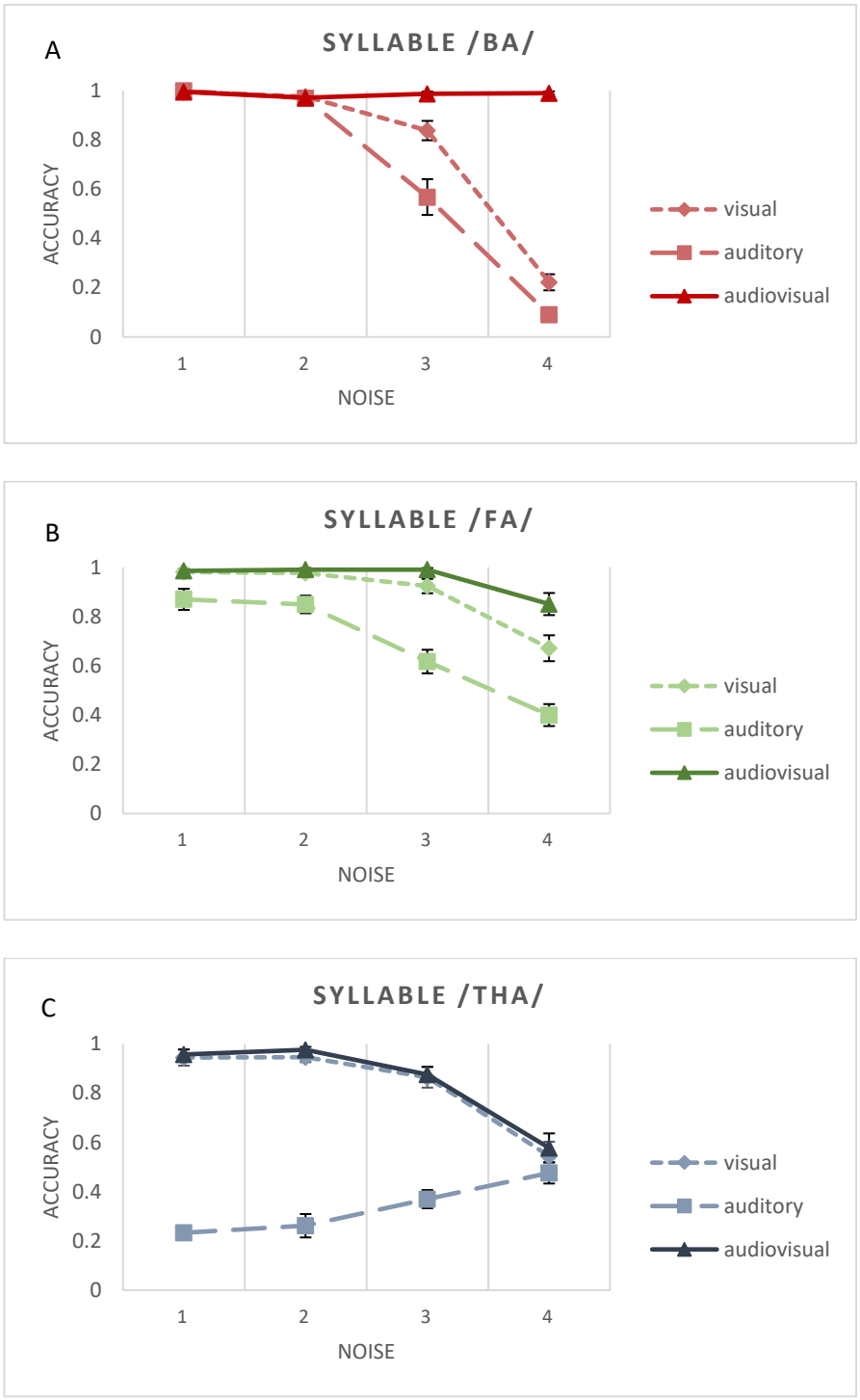

Figure 2. Mean accuracy scores, for the VO, AO, and AV conditions with the clearest auditory component. Accuracy scores are averaged across participants for each modality and plotted as a function of noise for the syllables A) /ba/, B) /fa/, and C) /tha/. Error bars represent the standard error of the mean.

In order to examine how participants' performance varied across every combination of auditory and visual noise for all syllables presented, a final analysis was performed on the data obtained only from the AV conditions. A repeated-measures ANOVA with the factors of Syllable (3 levels: /ba/, /fa/, /tha/), Visual Noise (4 levels: 1, 2, 3, 4), and Auditory Noise (4 levels: 1, 2 ,3, 4) was conducted. This analysis revealed a main effect of Syllable [$F(2, 40) = 11.155$, $p < 0.001$, $\eta^2_p = 0.358$], with the accuracy for the syllable /ba/ being significantly lower (M = 0.812) than that for /fa/ and /tha/ (M = 0.905 and 0.866, respectively). A significant main effect of Visual Noise was also obtained [$F(1.907, 38.136) = 165.424$, $p < 0.001$, $\eta^2_p = 0.892$], with performance being more accurate at the 1st level of visual noise (M = 0.960) as compared to the 2nd, 3rd, and 4th level of visual noise (M = 0.944, 0.866, and 0.673, respectively) as well as for the 2nd as compared to the 3rd and 4th level of visual noise and the 3rd as compared to the 4th. Finally, a main effect of Auditory Noise was obtained [$F(1.517, 30.345) = 89.216$, $p < 0.001$, $\eta^2_p = 0.817$], with performance being more accurate at the 1st and 2nd level of auditory noise (M = 0.929 and 0.922, respectively) as compared to the 3rd and 4th level of auditory noise (M = 0.863 and 0.730, respectively) and at the 3rd as compared to the 4th.

A significant three-way interaction between Syllable, Visual Noise, and Auditory Noise was obtained [$F(6.750, 135.003) = 17.103$, $p < 0.001$, $\eta^2_p = 0.461$], with accuracy for the syllable /ba/, at the 1st, 2nd, 3rd, and 4th level of visual noise, being significantly lower at the 4th level of auditory noise (M = 0.825, 0.621, 0.330, and 0.084, respectively), as compared to the 1st, 2nd, and 3rd level of auditory noise (M = 0.995, 0.981, and 0.976, for the 1st level of visual noise, 0.971, 0.981, and 0.959 for the 2nd level of visual noise, 0.986, 0.976, and 0.843 for the 3rd level of visual noise, and 0.990, 0.976, and 0.497 for the 4th level of visual noise, respectively). Also, at the 3rd and 4th level of visual noise, accuracy for the 3rd level of auditory noise was lower than that obtained for the 1st and 2nd level of auditory noise (see Figure 3A). For the syllable /fa/, at the 3rd level of visual noise, accuracy was significantly higher for the 1st level of auditory noise (M = 0.986) as compared to the 3rd and 4th level of auditory noise (M = 0.878 and 0.854, respectively). At the 4th level of visual noise, accuracy was significantly higher at the 1st level of auditory noise (M = 0.851) as compared to the 4th level of auditory noise (M = 0.640; see Figure 3B). No differences were found for the syllable /tha/. Finally, a significant two-way interaction between Syllable and Visual Noise [$F(2.544, 50.881) = 84.150$, $p <$

0.001, η^2_p = 0.808] and Visual Noise and Auditory Noise [F(3.637, 72.739) = 23.276, p < 0.001, η^2_p = 0.538] was obtained. These two-way interactions are reflected in the results of the three-way interaction obtained. No interaction between Syllable and Visual Noise was obtained [F(1.978, 2.906) = 2.906, p = 0.067, η^2_p = 0.127].

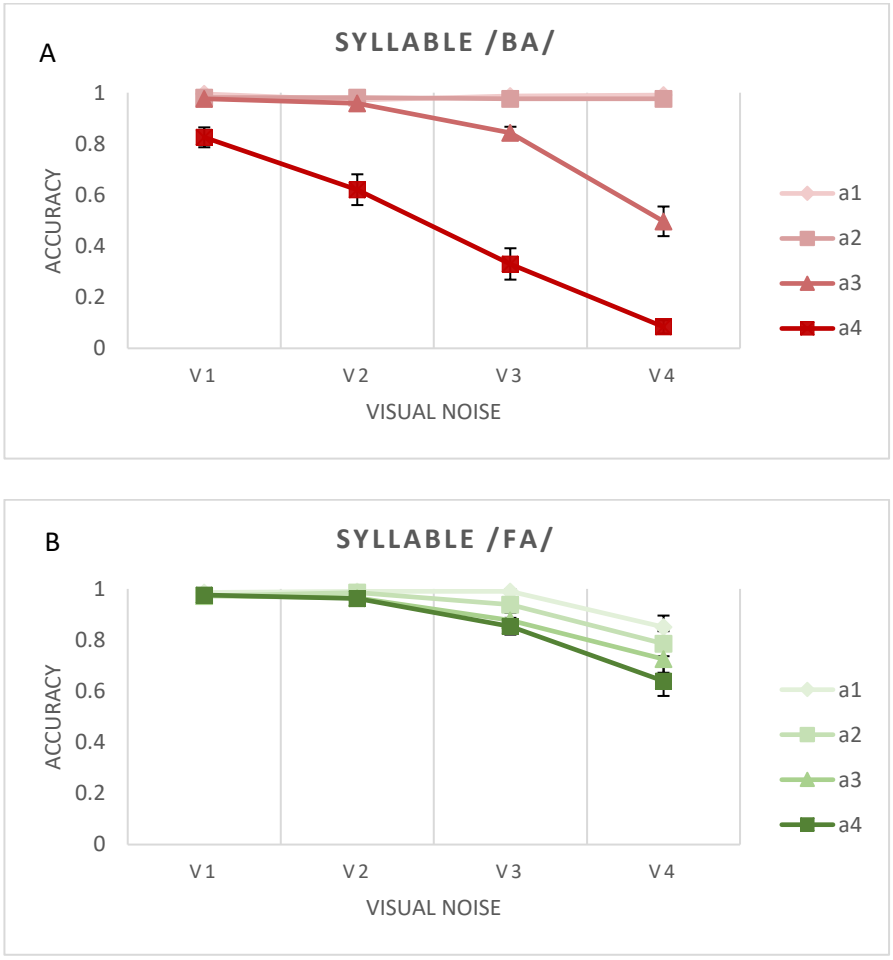

Figure 3. Mean accuracy for every AV condition averaged across participants. Accuracy scores are plotted as a function of visual noise for each level of auditory noise for the syllables A) /ba/ and B) /fa/. Error bars represent the standard error of the mean.

The results from the first and second analyses reported in this section (including VO, AO, and AV combinations with the clearest visual and auditory components, respectively) were highly similar. In both cases, highest accuracy was obtained for /fa/, then for /ba/, and finally for /tha/. AV performance was in both cases better than performance for unisensory conditions, with participants being more accurate for VO, as compared to AO conditions. Both analyses indicated that accuracy was higher for the

lowest levels of noise (i.e., 1st and 2nd) as compared to the highest (i.e., 3rd and 4th). This effect was obtained due to the fact that performance for unisensory conditions deteriorated as the noise levels increased, whereas participants' performance was nearly optimal for AV presentations, when either of two sensory streams was degraded as the other remained intact. The only exception was a decline in AV performance obtained when the visual channel was degraded for the syllable /tha/, pronounced for the 3rd and mostly, for the 4th level of noise. This suggests that participants relied mostly on the visual channel for the identification of the syllable /tha/, which is supported by the fact that this effect was not pronounced when the auditory channel was degraded. However, this can only be supported under the condition that one of the streams remains intact.

For the third analysis (including AV combinations across every level of visual and auditory noise), participants found it increasingly harder to identify the syllables /ba/ and /fa/ as the auditory noise levels increased. This effect was observed across every level of visual noise and was mostly pronounced when the visual stream was least informative (i.e., of highest noise). The degradation of the visual stream affected performance mostly for /ba/, suggesting that participants relied more on the visual stream when the syllable /ba/ was presented, as compared to when /fa/ was presented. For the syllable /tha/, accuracy scores obtained were lower than for the other two syllables and performance was not modulated depending on the combinations of auditory and visual noise, suggesting that noise did not affect performance when /tha/ was presented audiovisually. The different patterns of accuracy obtained between syllables indicate that the reliability of an AV speech event is modulated by the auditory and visual characteristics of each speech stimulus.

2.6.1.2 Multisensory Gain

In order to examine whether the data obtained were in accordance with the principle of inverse effectiveness, we calculated the magnitude of multisensory gain across the various experimental conditions with the use of four different indices. Given that there is no consensus regarding the most appropriate mathematical index for the calculation of multisensory gain, we used multiple indices in order to compare our results with previous findings and to examine whether past inconsistencies could be due to the different indices used for the calculation of multisensory gain. Specifically, we

utilized the Multisensory Integration (MSI), Contrast, Absolute Difference, and Absolute Difference in % indices. The MSI index is one of the most widely used indices of multisensory gain, originally used to calculate the magnitude of multisensory gain by Meredith and Stein (1986a, 1986b; see also e.g., Alvarado et al., 2009; Bell, Corneil, Munoz, & Meredith, 2003; Frens & van Opstel, 1998; Ghazanfar, Maier, Hoffman, & Logothetis, 2005; Jiang et al., 2001; Meredith, Nemitz, & Stein, 1987; Perrault, Vaughan, Stein, & Wallace, 2005; Rowland & Stein, 2007; Stanford, Quessy, & Stein, 2005; Stein & Meredith, 1994; Stein et al., 2009; Wallace et al., 2006; Xu et al., 2012; Xu et al., 2017; Xu et al., 2015; Yu, Rowland, & Stein, 2010; Yu, Xu, Rowland, & Stein, 2013). It reflects multisensory enhancement as the difference between the magnitude of the multisensory response and the most effective unisensory response, normalized by the magnitude of the most effective unisensory response, and calculated by the formula: MSI = [AV – max (VO, AO) x 100] / max (VO, AO)]. It should be noted that MSI is highly dependent on the unisensory response used for its calculation and the smaller the unisensory response, the greater the enhancement obtained (see Stein et al., 2009).

The Contrast index, originally used by Motter (1994), has also been used for the calculation of the magnitude of the multisensory gain (e.g., Alvarado, Stanford, Vaughan, & Stein, 2007a; Alvarado et al., 2009; Perrault, Vaughan, Stein, & Wallace, 2005; Stein et al., 2009). This index reflects the difference in the response magnitude between multisensory conditions and the most effective unisensory conditions, divided by their sum, it takes values in the range between -1 and 1, and calculated by Contrast = [AV - max (VO, AO)] / [AV + max (VO, AO)]. The advantage of this index is that it can be defined even in the lack of unisensory responses, however, even great differences in the gain obtained can appear to be similar across conditions due to the compression of values in the range between -1 and 1 (Stein et al., 2009).

The Absolute Difference index reflects the raw differences in the response magnitude between multisensory conditions and the most effective unisensory conditions. This is the only index that does not reflect gain as a relative enhancement from unisensory to multisensory conditions, but expresses this magnitude in absolute values (Holmes, 2007; Nahanni, 2014). It is calculated by Absolute = AV- max (VO, AO). A variation of this index has also been used by calculating the raw differences in the response magnitude between AV and auditory conditions (AV-AO), which value

indicates the magnitude of visual enhancement (e.g., Callan, et al., 2003; Grant & Braida, 1991; Grant & Seitz, 1998; Ma et al., 2009; Ross et al., 2007; Sumby & Pollack, 1954). Similarly, the Absolute Difference between AV and AO conditions has been calculated as a percentage (Absolute Difference in % index) via the formula (AV-AO) x 100. This index reflects multisensory enhancement due to the addition of information provided by the visual modality, or visual enhancement (Nahanni, 2014), and has been employed by a number of researchers in order to quantify the contribution of the visual channel in speech perception (e.g., Callan et al., 2003; Grant & Walden, 1996; Ross et al., 2007; Sumby & Pollack, 1954). We also adopted it to directly compare our findings with those obtained by Ross and colleagues (2007).

The Normalized Enhancement index, calculated by NE = [AV – max (AO, VO)] / [1 – max (AO,VO)], has been employed for the calculation of the multisensory gain in a number of studies (e.g., Desai, Stickney, & Zeng 2008; Grant & Seitz, 1998; Grant, Walden, & Seitz, 1998; Grant & Walden, 1996; Nahanni, 2014; Rabinowitz, Eddington, Delhorne, & Cuneo, 1992; Ross et al. 2007; Sommers, Tye Murray, & Spehar, 2005; Tye Murray, Sommers, Spehar, Myerson, & Hale, 2010) and was originally used by Sumby and Pollack (1954) in order to specifically assess the magnitude of visual enhancement, in which case the term max (AO, VO) is substituted with the term AO. The advantage of this index is that the magnitude of gain is normalized by the maximum possible enhancement that can be obtained from unisensory to multisensory conditions. Initially, we calculated this index as well, but no further analyses are presented due to the fact that, for more than half experimental conditions, the denominator was equal to zero making the value of the fraction undefined. This is a potential drawback of this index, as it arises when at least one of the sensory streams is highly reliable, resulting in a ceiling effect for accuracy scores. This issue has also been reported by Nahanni (2014).

The data obtained from these indices were analyzed using a repeated-measures ANOVA with the factors of Syllable (3 levels: /ba/, /fa/, /tha/), Visual Noise (4 levels: 1, 2, 3, 4), and Auditory Noise (4 levels: 1, 2, 3, 4). This analysis revealed a main effect of Syllable on the magnitude of gain as calculated with the MSI index [$F(1.074, 21.472) = 5.523$, $p = 0.026$, $\eta^2_p = 0.216$], the Contrast index [$F(1.351, 27.013) = 9.070$, $p = 0.003$, $\eta^2_p = 0.312$], the Absolute Difference index [$F(2, 40) = 6.936$, $p = 0.003$, $\eta^2_p = 0.258$], and the Absolute Difference in % index [$F(1.655, 33.095) = 55.264$, $p < 0.001$, $\eta^2_p = 0.734$].

In each case, the gain for the syllable /ba/ was lower (M = -12.031, -0.098, -0.120, and 15.496 for MSI, Contrast, Absolute Difference, and Absolute Difference in %, respectively) than for the syllables /fa/ and /tha/ (M = -2.120 and 14.266 for MSI, -0.021 and 0.019 for Contrast, -0.026 and 0.032 for Absolute, and 22.044 and 53.056 for Absolute Difference in %, respectively). The effect of Visual Noise was not significant for the magnitude of gain as calculated with the MSI index [$F(1.639, 32.774) = 1.015$, $p = 0.360$, $\eta^2_p = 0.048$] or with the Absolute Difference index [$F(1.472, 29.446) = 1.077$, $p = 0.366$, $\eta^2_p = 0.051$], but significance was obtained for the Contrast index [$F(1.511, 30.223) = 4.006$, $p = 0.039$, $\eta^2_p = 0.167$] and the Absolute Difference in % index [$F(1.907, 38.136) = 165.424$, $p < 0.001$, $\eta^2_p = 0.892$]. For the Contrast index, the gain was higher at the 1st level of visual noise (M = -0.007) as compared to the 3rd (M = -0.042), while for the Absolute Difference in % index, the gain was higher for the 1st level of visual noise (M = 40.159) as compared to the 2nd, 3rd, and 4th level (M = 38.492, 30.754, and 11.389, respectively) and for the 2nd as compared to the 3rd and 4th, and for the 3rd as compared to the 4th. There was a significant main effect of Auditory Noise on the levels of gain as calculated with the MSI index [$F(1.733, 34.652) = 14.610$, $p < 0.001$, $\eta^2_p = 0.422$], the Contrast index [$F(1.393, 27.864) = 30.121$, $p < 0.001$, $\eta^2_p = 0.601$], the Absolute Difference index [$F(1.193, 23.854) = 10.015$, $p = 0.003$, $\eta^2_p = 0.334$], and with the Absolute Difference in % index [$F(2.566, 51.324) = 15.382$, $p < 0.001$, $\eta^2_p = 0.435$]. For the MSI, Contrast, and Absolute Difference indices, we found that the gain was lower at the 4th level of auditory noise (M = -12.633, -0.127, and -0.144, respectively) than at the 1st, 2nd, and 3rd level of auditory noise (M = 5.096, 3.478, and 4.213 for MSI, -0.002, -0.001, and -0.003 for Contrast, and 0.001, -0.001, and -0.008 for Absolute, respectively). The gain as calculated with the Absolute Difference in % index was significantly higher at the 3rd and 4th level of auditory noise (M = 34.431 and 40.741, respectively) as compared to the gain at the 1st and 2nd level of auditory noise (M = 22.804 and 22.817, respectively).

The three-way interaction between Syllable, Visual Noise, and Auditory Noise was not significant for the gain as calculated with the MSI index [$F(3.770, 75.406) = 0.960$, $p = 0.431$, $\eta^2_p = 0.046$] or the Absolute index [$F(1.227, 24.537) = 1.751$, $p = 0.199$, $\eta^2_p = 0.081$], but reached significance for the Contrast index [$F(3.292, 65.834) = 3.637$, $p < 0.001$, $\eta^2_p = 0.154$] and the Absolute Gain in % index [$F(6.750, 135.003) = 17.101$, $p < 0.001$, $\eta^2_p = 0.461$]. For the Contrast index (see Appendix Table 1), we found that when

the syllable /ba/ was presented at the 1st, 2nd, 3rd, and 4th level of visual noise, gain was lower for the 4th level of auditory noise (M = -0.104, -0.291, -0.506, and 0.522, respectively) as compared to the gain obtained for the 1st, 2nd, and 3rd level of auditory noise (M = -0.003, -0.010, and -0.010 for the 1st level of visual noise, -0.007, -0.008, and -0.036 for the 2nd level of visual noise, and -0.005, 0.005, and -0.037 for the 3rd level of visual noise, respectively; see Figure 4).

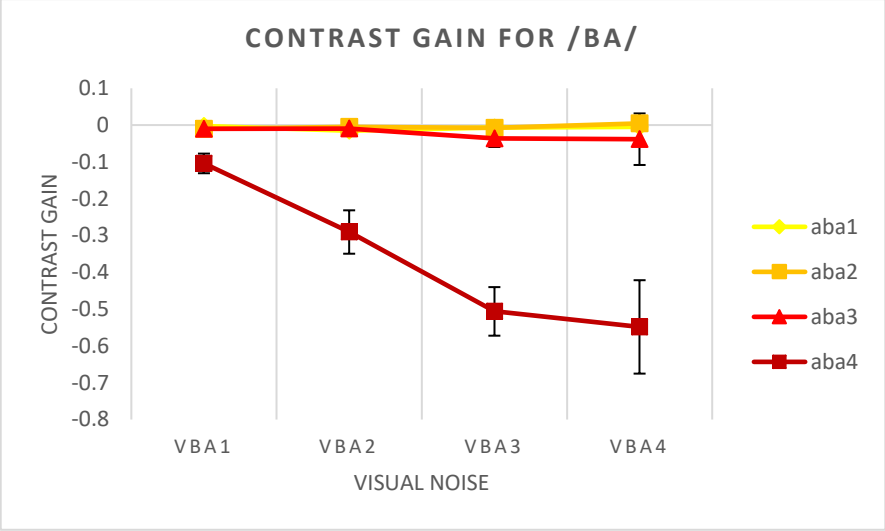

Figure 4. Mean levels of gain for the syllable /ba/ as calculated with the Contrast index, averaged across participants and plotted as a function of visual noise for every level of auditory noise. Error bars represent the standard error of the mean.

The Absolute Difference in % index (see Appendix Table 2) indicated that for the syllable /ba/, at the 1st, 2nd, and 3rd level of visual noise, the gain at the 1st and 2nd level of auditory noise was lower (M = -0.476 and 1.111 for the 1st level of visual noise, -2.857 and 1.111 for the 2nd level of visual noise, and -1.270 and 0.635 for the 3rd level of visual noise, respectively) than the gain obtained at the 3rd and 4th level of auditory noise (M = 40.794 and 73.492 for the 1st level of visual noise, 39.048 and 53.016 for the 2nd level of visual noise, and 27.460 and 23.968 for the 3rd level of visual noise, respectively; see Figures 5A, 6A). For the syllable /fa/, at the 1st, 2nd, and 3rd level of visual noise, we found that the gain was significantly higher at the 4th level of auditory noise (M = 57.619, 56.190 and 45.397, for the 1st, 2nd and 3rd level of visual noise, respectively) as compared to the 1st, 2nd, and 3rd level of auditory noise (M = 11.587, 12.381, and 35.556 for the 1st level of visual noise, 12.063, 13.651 and 34.762 for the 2nd level of visual

noise, and 12.063, 8.889, and 26.032 for the 3rd level of visual noise, respectively). For the 4th level of visual noise, the gain was lower at the 2nd level of auditory noise (M = -6.349) as compared to the 4th (M = 23.968; see Figures 5B, 6B). For the syllable /tha/, at the 1st, 2nd, and 3rd level of visual noise, we found that the gain at the 1st, 2nd, and 3rd level of auditory noise was significantly higher (M = 72.381, 68.254, and 60.635 for the 1st level of visual noise, 74.286, 70.476, and 61.111 for the 2nd level of visual noise, and 64.286, 66.190, and 53.968 for the 3rd level of visual noise, respectively) than at the 4th level of auditory noise (M = 48.571, 49.048, and 41.429 for the 1st, 2nd, and 3rd level of visual noise, respectively). For the 2nd level of visual noise, gain was also significantly higher at the 1st level of auditory noise as compared to the 3rd. For the 4th level of visual noise, gain was higher at the 2nd (M = 36.826) as compared to the 4th level of auditory noise (M = 16.825; see Figures 5C, 6C).

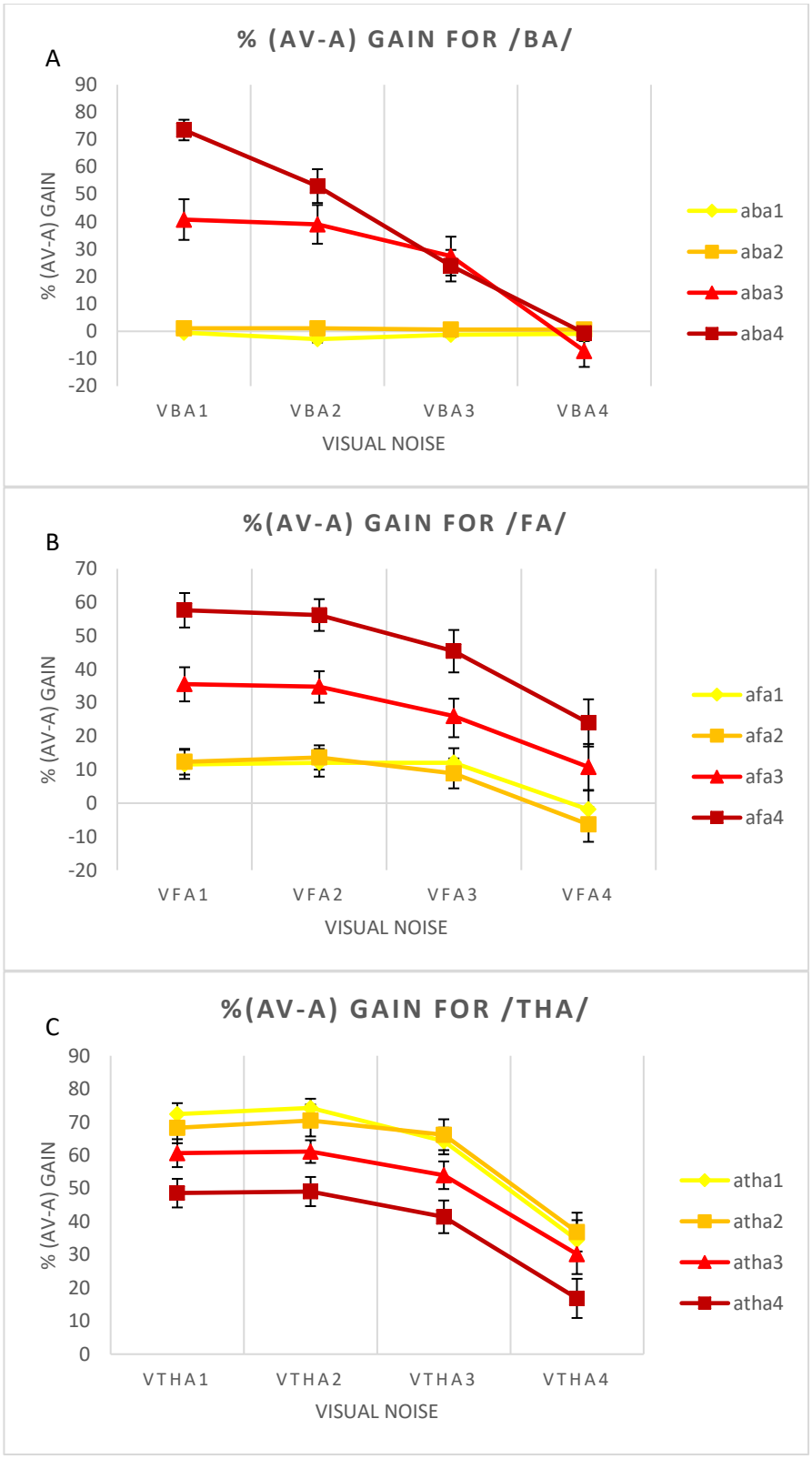

Figure 5. Mean levels of gain, as calculated with the % (AV-A) formula, averaged across participants and plotted as a function of visual noise at every level of auditory noise for the syllables A) /ba/, B) /fa/, and C) /tha/. Error bars represent the standard error of the mean.

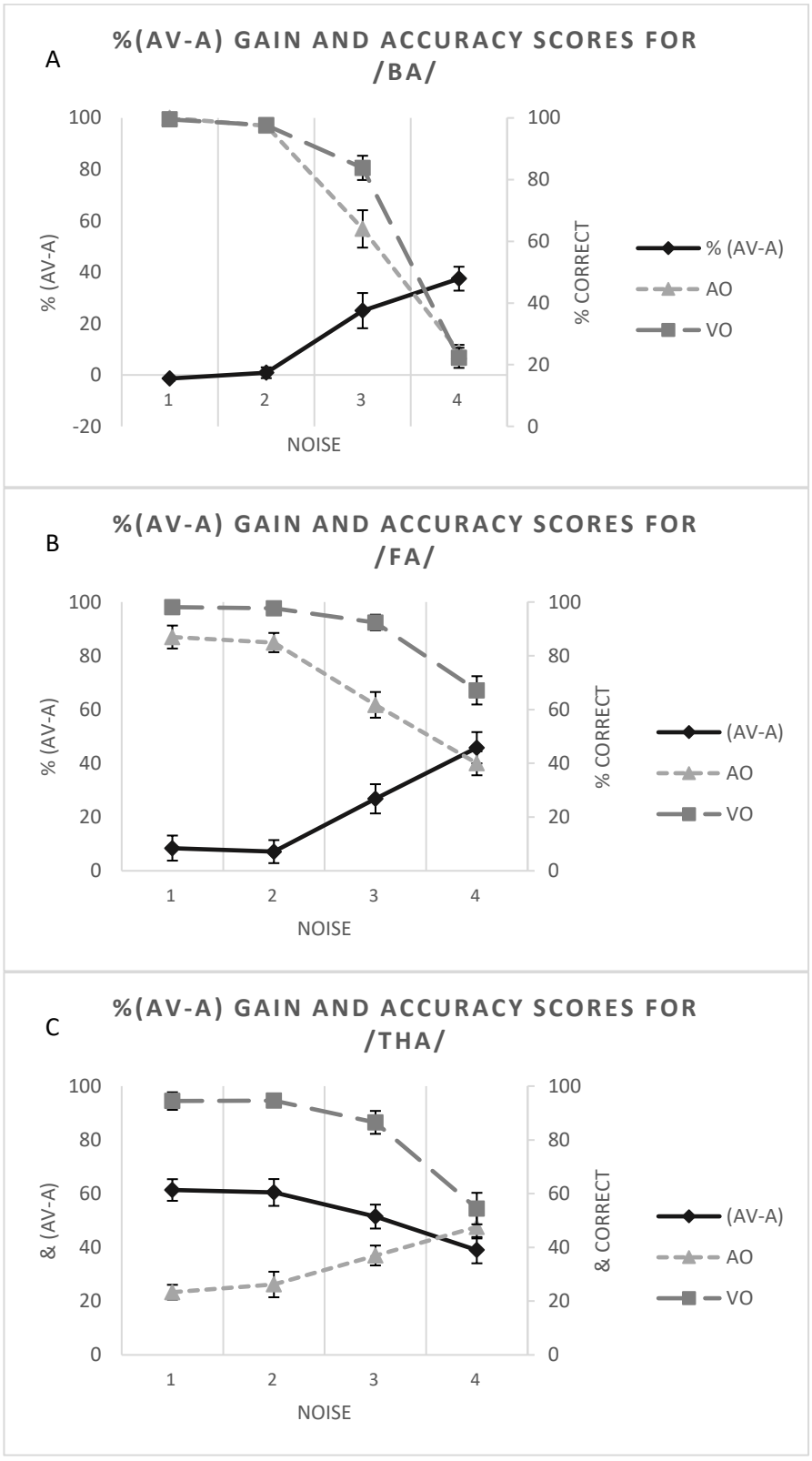

Figure 6. Absolute Gain in %, averaged across participants, is represented with the solid black line as a function of auditory noise. The percentage of correct responses for

unimodal conditions, averaged across participants, is depicted with dashed lines. Accuracy in AO conditions as a function of auditory noise is depicted with the light grey line, accuracy in VO conditions as a function of visual noise is depicted with the dark grey line for the syllables A) /ba/, B) /fa/, and C) /tha/. Error bars represent the standard error of the mean.

Finally, this analysis revealed a significant interaction between Syllable and Auditory Noise on the levels of gain, as calculated with the MSI index [$F(2.898, 57.956) = 11.824$, $p < 0.001$, $\eta^2_p = 0.372$], the Contrast index [$F(1.920, 38.408) = 36.848$, $p < 0.001$, $\eta^2_p = 0.648$], the Absolute Difference index [$F(1.165, 23.299) = 8.639$, $p = 0.005$, $\eta^2_p = 0.302$], and the Absolute Difference in % index [$F(6, 120) = 24.919$, $p < 0.001$, $\eta^2_p = 0.555$]. For the Contrast and Absolute Difference in % indices, this interaction is reflected on the three-way interactions reported above. As for the MSI and Absolute Difference indices (as shown in Appendix Tables 3 and 4, respectively), we found that for the syllable /ba/, the gain obtained for the 4th level of auditory noise was significantly lower (M = -41.991 and -0.409 for MSI and Absolute Difference, respectively) as compared to the 1st, 2nd, and 3rd level of auditory noise (M = -1.389, -0.635, and -4.109 for MSI and -0.014, -0.009, and -0.048 for Absolute Difference, respectively; see Figures 7 & 8).

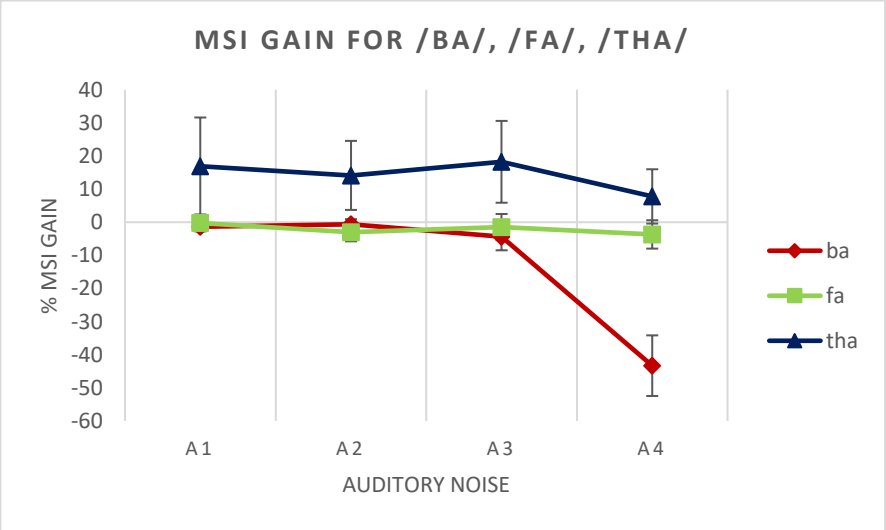

Figure 7. Mean levels of gain for the syllables /ba/, /fa/ and /tha/ as calculated with the multisensory integration (MSI) index, averaged across participants and plotted as a function of auditory noise. Error bars represent the standard error of the mean.

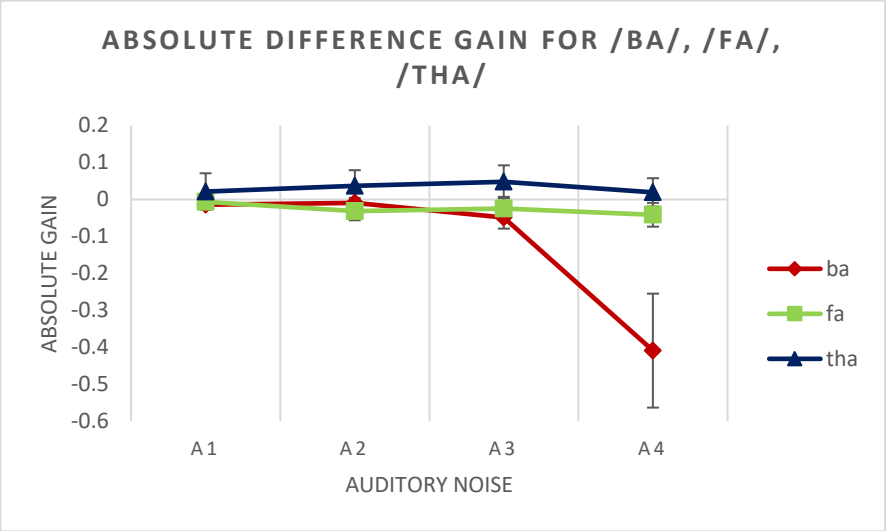

Figure 8. Mean levels of gain for the syllables /ba/, /fa/ and /tha/ as calculated with the Absolute index, averaged across participants and plotted as a function of auditory noise. Error bars represent the standard error of the mean.

There was no significant interaction between Syllable and Visual Noise on the gain levels for the MSI index [$F(1.833, 36.655) = 2.058$, $p = 0.146$, $\eta^2_p = 0.093$], the Contrast index [$F(2.467, 49.334) = 1.615$, $p = 0.204$, $\eta^2_p = 0.075$], the Absolute Difference index [$F(2.074, 41.476) = 1.694$, $p = 0.128$, $\eta^2_p = 0.078$], or the Absolute Difference in % index [$F(1.978, 39.566) = 2.906$, $p = 0.067$, $\eta^2_p = 0.127$]. There was no significant interaction between Visual and Auditory Noise on the gain levels for the MSI index [$F(3.037, 60.730) = 1.644$, $p = 0.188$, $\eta^2_p = 0.076$] or the Absolute Difference index [$F(1.267, 25.334) = 1.662$, $p = 0.212$, $\eta^2_p = 0.077$]. However, this interaction reached significance when the gain was calculated with the Contrast index [$F(2.833, 56.670) = 4.865$, $p = 0.005$, $\eta^2_p = 0.196$] and the Absolute Difference in % index [$F(3.637, 72.739) = 23.276$, $p < 0.001$, $\eta^2_p = 0.538$]. These interactions are reflected in the results of the three-way interaction between Syllable, Visual Noise, and Auditory Noise.

The findings from Experiment 1 suggest that the magnitude of gain depends on the noise combinations between the visual and auditory streams, on the presented syllable, and on the index used for its calculation. The main effect of the syllable on the magnitude of gain suggests that gain is modulated by the physical characteristics (visual

and auditory) of the presented speech stimulus. Visual noise manipulations had an effect on the levels of gain, only as calculated by the Contrast and Absolute Difference in % indices (whereby gain was maximized when the visual stream was of the lowest noise), while the effect of auditory noise manipulation on gain was evident for every calculated index. According to the MSI, Absolute Difference, and Contrast indices, gain levels were minimized when the auditory noise was maximized, which is against the predictions of inverse effectiveness. The tendency of gain to be maximized when the two sensory streams were of low noise suggests that the two components were most effectively integrated in these conditions. However, gain was maximized at the two highest levels of auditory noise as calculated with the Absolute Difference in % index (which reflects the levels of visual enhancement), showing that the contribution of the visual signal was more pronounced when the auditory stream was degraded the most.

2.2 Experiment 2

In the past, a few behavioral studies have been conducted to investigate for possible interactions between the rules of multisensory perception. The interrelationship between inverse effectiveness and the temporal rule with the spatial rule has been investigated in limited behavioral studies (e.g., Macaluso et al., 2004; Nidiffer et al., 2016; Royal et al., 2009), whereby, interactions between these factors (i.e., between space and time and between space and effectiveness) were demonstrated. Stevenson and colleagues (2012) used simple auditory (white-noise bursts) and visual (white circles) stimuli and manipulated both their spatial locations and the stimulus onset asynchronies (SOAs) between their presentations. They asked participants to perform a spatial localization task (participants had to indicate in which out of 4 possible locations the stimulus had been presented in each trial) and a simultaneity judgment task (SJ; participants had to report if the auditory and visual components of a sentence were in or out of synchrony in each trial). The results showed that participants responded faster when AV stimuli were presented centrally and at smaller SOAs than when their presentation was peripheral and at larger SOAs. Furthermore, Stevenson et al. reported an interaction between the factors of location and timing, since the effect of SOA was larger in peripheral as compared to central locations. As for the SJ task, it was found that when AV stimuli were presented at large SOAs, participants reported them as synchronous mostly in the periphery as compared to central locations. Together,

these findings highlighted that multisensory integration depends not only on the temporal or spatial coincidence of the unisensory components, but also, on the interaction between these two factors.

The interrelationship of the spatial rule and the inverse rule has been examined by Nidiffer and colleagues (2016). Specifically, they used flashes of light and auditory noise bursts, presented at four possible spatial locations and at two levels of intensity and asked participants to complete a localization task. The stimuli could either be presented alone or in combination. For the latter case, the stimuli were always spatially coincident and of matching intensity. Their main findings were that in the VO conditions, participants responded slower and less accurately as eccentricity increased, while for the AO conditions, participants had the worst performance at the 2 intermediate locations (not at maximum or minimum eccentricity). For the AV conditions, it was found that participants responded faster and more accurately for AV combinations of high intensity. The authors supported that their findings were in accordance with the predictions of the principle of inverse effectiveness, since performance was enhanced when AV stimuli were presented at locations where the worst performance had been recorded in unisensory conditions. The values of multisensory gain increased as eccentricity increased, with this effect being mostly evident when the AV stimuli were presented at low intensities, thus highlighting the interaction between stimulus location and effectiveness in multisensory integration.

Behavioural research in synchrony perception has mostly focused on the characteristics of the temporal window of integration (TWI; the temporal interval during which any discrepancy between the streams of information is not perceived; Vatakis, 2013; Vatakis, Maragos, Rodomagoulakis, & Spence, 2012), such as its width, variability, and asymmetrical nature. Importantly, the interdependency between these characteristics and the intelligibility of speech stimuli in relation to cross-modal integration has also been examined (e.g., Conrey & Pisoni, 2006; Grant & Greenberg, 2001). For instance, in a study by Grant and Greenberg (2001), participants were presented with sentences, composed of auditory streams with an intelligibility rate (i.e., percentage of correctly identified words) between 9 and 31%, and visual streams with an intelligibility rate between 1 and 22%. It was found that the synchronous presentation of the auditory and visual streams resulted in enhanced performance, with

intelligibility rates in AV conditions being raised to 63%. Subsequently, the authors presented the auditory and visual components of the sentences from their first experiment at various levels of asynchrony, in order to characterize the asymmetric nature of the TWI. They found that when the visual stream led the auditory by up to 160-200 ms, intelligibility scores were similar to the ones obtained in synchrony. However, when the auditory stream led the visual, even by 40 ms, intelligibility scores were lower than in synchronous conditions. Their findings support the existence of an asymmetrical window of integration for speech stimuli, but also highlight the interdependency between timing and stimulus effectiveness. The relationship between intelligibility scores and synchrony perception was also investigated by Conrey and Pisoni (2006), who examined whether the variability of intelligibility scores in VO, AO, and AV conditions was related to the variability in synchrony detection scores, within each participant separately. They found that participants with higher intelligibility scores (as measured in AO and AV conditions) were less tolerant to asynchronies between the speech stimulus components, exhibiting consequently, narrower TWIs. However, no relationship between intelligibility scores obtained in VO conditions and synchrony detection scores was obtained. While the findings reported here suggest that there is a relationship between the levels of intelligibility and synchrony perception for speech stimuli, they do not allow us to draw any conclusions about the nature of the relationship between inverse effectiveness and the temporal rule. That is mainly due to the fact that the examination of inverse effectiveness requires varying levels of stimulus effectiveness, so that the magnitude of gain obtained for highly effective stimulus combinations can be compared with the one obtained for weakly effective stimulus combinations. Thus, the relationship between the temporal rule and inverse effectiveness remains to be investigated due to the lack of effectiveness manipulation of the sensory streams in the studies reported above.

Here, we attempted, for the first time, to examine if these two fundamental rules of multisensory integration interact with each other, in order to further investigate the processes that underlie multisensory integration and to specify their nature. More specifically, we focused on investigating whether and in what way, stimulus effectiveness can alter the width of the TWI, thus we examined whether the integration of two events can accept discrepancies of prolonged or shorter temporal asymmetries, depending on how effective the two sensory components are. The parallel modulation

of both factors commonly occurs in the physical world, therefore it is crucial to examine and determine whether and how multisensory integration would be enhanced or depressed, depending on the parallel modulation of both the factors of timing and effectiveness. The results from Experiment 1 did not suggest the existence of a consistent pattern following the principle of inverse effectiveness. Therefore, previously reported limitations about the extension of inverse effectiveness from a neuronal to a behavioral level (see Holmes, 2007) are further supported by our findings. However, it remains an open question whether a high gain stimulus pair will have a different temporal widow width as compared to a low gain stimulus pair.

In the present study, we selected the stimuli that led to the highest and lowest levels of multisensory gain from Experiment 1 and presented them at various SOAs, while participants were asked to perform a temporal order judgment (TOJ) task. The main purpose of Experiment 2 was to investigate the relationship between the temporal rule and the rule of inverse effectiveness by examining whether the width of the TWI would vary as a function of multisensory gain level. Given that high levels of gain indicate a robust integration between the two components (i.e., thus, they are expected to be perceived as unified events), we hypothesized that in these conditions participants would be more tolerant to asynchronies between the visual and auditory speech components (i.e., it would be harder for participants to detect differences between the onsets of each component for high gain stimuli), while in low gain conditions they would be more sensitive in detecting asynchronies between the onsets of the two sensory streams. Low gain indicates that the two components were as effective in isolation as they were in combination. That is, whether equally high or equally low levels of accuracy were obtained for both unisensory and multisensory conditions. In the first case, if high levels of accuracy were obtained for both unisensory and multisensory conditions, redundant -or even complementary- information would be provided when participants were exposed to both components. This would result in increased sensitivity when participants are judging which sensory stream preceded the other, thus, participants were expected to exhibit a narrower TWI in these conditions. In the second case, if low levels of accuracy were obtained for both unisensory and multisensory conditions, the lack of performance enhancement for multisensory conditions would suggest that the two components were not effectively integrated. Therefore, it would be easier for participants to detect differences between the onset of

the two sensory components, suggesting again that the TWI in low gain conditions would be narrower than the one obtained in high gain conditions. Thus, the width of the TWI was expected to be larger in high, as opposed to low gain conditions.

2.2.1 Participants

Forty-three new participants (34 females), aged between 18-30 years (Mean age = 21.8 years) took part in this experiment. The duration of the experiment was approximately 45 minutes.

2.2.2 Apparatus and Stimuli

The apparatus remained the same as in Experiment 1 and the stimuli for this experiment were selected based on the data analysis from the first 15 participants in Experiment 1. We selected the stimuli for which the highest and lowest levels of multisensory gain were obtained from the MSI, Contract, and Absolute Difference indices. In total, six stimuli were selected, two for each of the three syllables.

All stimuli in this experiment were AV video clips of 3090 ms in duration and each video file consisted of 84 frames (frame size = 720 x 576 pixel, depth = 24 bits). In order to introduce the asynchronies between the two sensory streams of the stimuli, a still image (depicting the avatar with its mouth closed; see Figure 9) and background auditory noise were extracted from the last video frame of each stimulus and implemented on the beginning and ending of each video clip's components. The presentations of the still image and the background auditory noise on the beginning and ending of each video clip were of different durations. This difference was equivalent to the SOA tested in each condition (the SOA values are reported below). This procedure was made so that the visual and auditory streams would always start and end at the same time and so that they would be of equal durations, in order to avoid cuing participants about the nature of the AV delay. A white noise auditory stimulus was delivered in the background at 30 dB during the entire experimental procedure so as to ensure continuous AV stimulation during video asynchrony exposure. Participants responded using a standard keyboard, with their left hand's index or middle finger providing "auditory first" responses and their right hand's index or middle finger providing "vision first" responses.

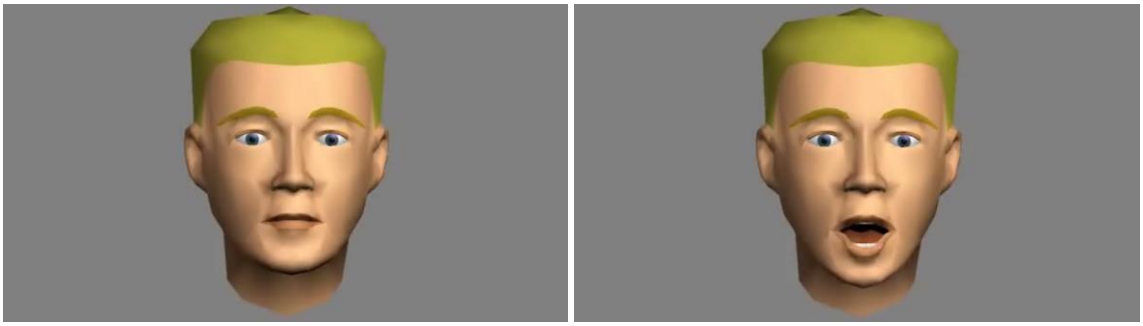

Figure 9. Images depicting the avatar with its mouth (A) open and (B) closed.

2.2.3 Design

The intervening SOAs between the presentation of the auditory and visual components of the stimuli were 11 in total (±300, 266, 200, 133, 67, and 0 ms), with negative SOAs indicating that the presentation of the auditory component preceded the presentation of the visual component. The experiment was divided into two blocks and trials were randomized within each block. A practice block was performed prior to the beginning of the main experimental procedure. The practice block included two randomized loops consisting of the stimuli with the higher asynchrony levels (i.e., 6 video clips with a 300 ms visual lead and 6 video clips with a 300 ms auditory lead). After completing the practice block, participants completed 2 blocks, each consisting of 330 experimental trials (i.e., 11 levels of asynchrony for each of the 6 AV stimuli presented, with 5 repetitions for each condition within each block).

2.2.4 Procedure

Each participant received verbal instructions prior to the beginning of the experiment. Specifically, participants were informed that three different syllables will be presented to them and that in each trial, they had to report which stimulus stream was presented first, the auditory or the visual, independent of the syllable uttered. Participants were asked to remain as focused as possible on the task. They were also informed that sometimes the task would be difficult in which case they should make an informed guess as to which stimulus stream was presented first. They were encouraged to avoid guessing and to try to maintain high levels of accuracy. They were instructed to respond by pressing the key “a” when the auditory stream occurred first and the key “o” when the visual stream occurred first. The experiment started as soon as the participant

pressed the "Enter" key. The task was self-paced and participants were allowed to take breaks between the experimental blocks.

2.2.5 Results and Discussion

Participants' responses were recorded as a proportion of "vision first" responses and were subsequently converted to their equivalent z-score values under the assumption of a cumulative normal distribution (cf. Finney, 1964). Figure 10 shows the proportion of "vision first" responses, averaged across participants, as a function of SOA, for all six experimental conditions. Slope and intercept values were derived by first calculating the best-fitting straight lines for every participant, under each experimental condition. We then calculated the just noticeable difference (JND=0.675/slope; since ± 0.675 represents the 75% and 25% point on the cumulative normal distribution) and point of subjective simultaneity (PSS; PSS = - intercept/slope) values (see Coren, Ward, & Enns, 2004). The JND is a standardized measure that represents participants' accuracy when judging the temporal order between the two sensory events, while the PSS reflects the amount of time for one stimulus modality to precede the other stimulus modality, required for a participant to respond as if the two stimuli are perceptually synchronous to them (i.e., for participants to provide "audition first" and "vision first" responses with the same frequency; Vatakis, Maragos, Rodomagoulakis, & Spence, 2012). Bonferroni-corrected t-tests were used for every post hoc comparison reported in this section (where $p < .05$ prior to correction).

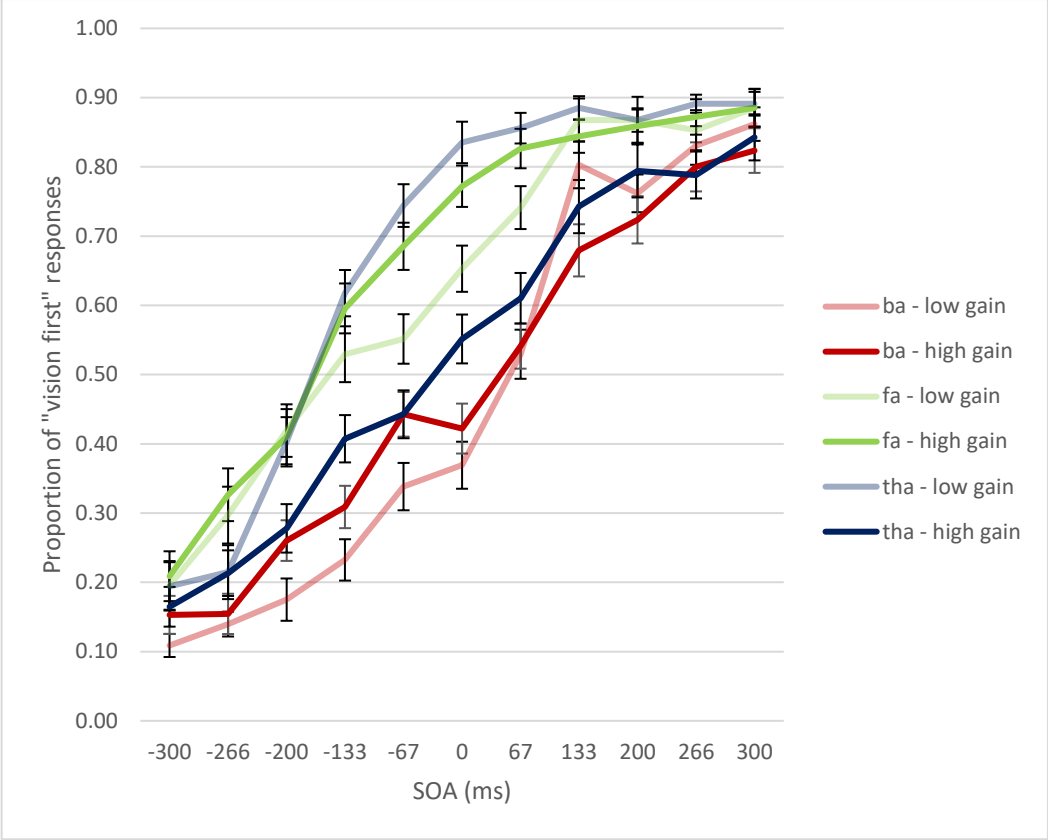

Figure 10. Proportion of "vision first" responses, averaged across participants, plotted as a function of the stimulus onset asynchrony (SOA) between the auditory and visual streams for the 6 tested experimental conditions (syllable ba-low gain, syllable ba-high gain, syllable fa-low gain, syllable fa-high gain, syllable tha-low gain, syllable tha-high gain). Red, green, and blue lines represent the syllables /ba/, /fa/, and /tha/, respectively. Bold colours are used to indicate high gain conditions. The error bars represent the standard error of the mean.

The JND and PSS data were both analysed with the use of repeated measures ANOVA with the factors Syllable (3 levels: /ba/, /fa/, and /tha/) and Gain (2 levels: low, high). We removed data obtained from 26 participants due to their large PSS and/or JND values, which indicated inability to complete the task (cf. Spence et al., 2001, for similar exclusion criteria). Analysis of the JND data revealed a main effect of Syllable [$F(2, 32) = 6.470$, $p = 0.004$, $\eta^2_p = 0.288$], with the discrimination performance for /ba/ being significantly better (M = 175.45 ms) than that for /tha/ and /fa/ (M = 207.42 and 210.24 ms, respectively). There was also a main effect of Gain [$F(1, 16) = 6.275$, $p = 0.023$, $\eta^2_p = 0.282$], but no interaction between these two factors was obtained [$F(2, 32)$

= 1.882, p = 0.169, η^2_p = 0.105]. The main effect of gain showed that discrimination accuracy was poorer at high gain conditions (M = 213.126 ms) as compared to low gain conditions (M = 189.292 ms; see Figure 11).

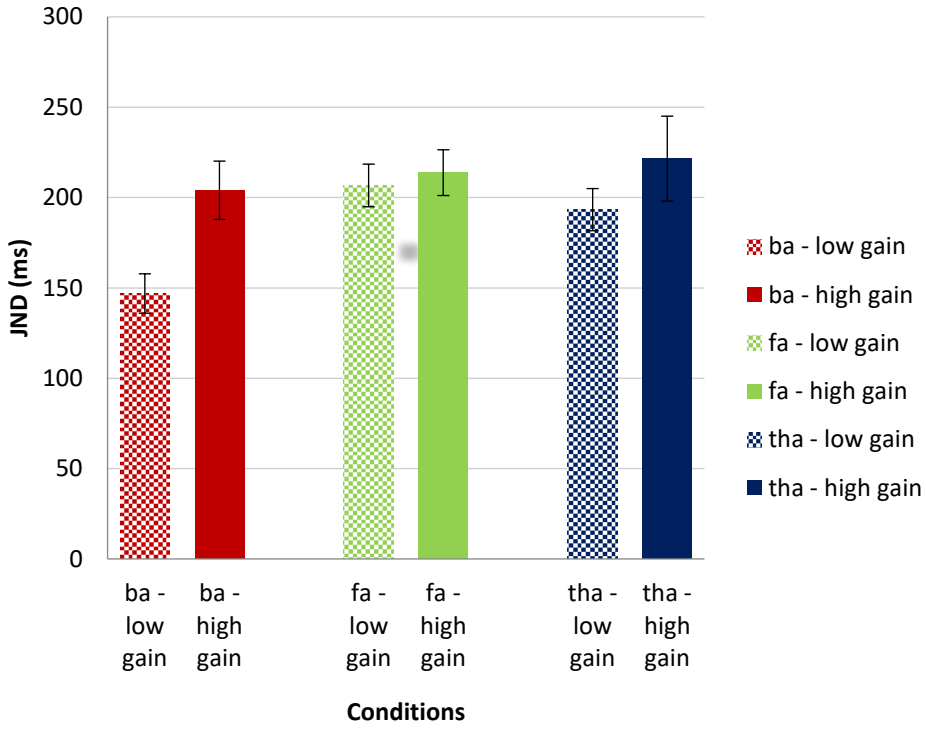

Figure 11. The Just Noticeable Difference (JND) values, averaged across participants for every experimental condition. Low gain stimuli are represented by pattern-filled bars, high gain stimuli are symbolized with bold, solid bars. Red bars represent data for /ba/, green bars represent the data obtained for /fa/, blue bars represent the data for /tha/. The error bars represent the standard error of the mean.

Analysis of the PSS data revealed a main effect of Syllable [F(2, 32) = 52.132, p < 0.001, η^2_p = 0.765], with larger visual leads obtained for /fa/ (M = 150.664 ms) than for the visual leads of /tha/ (M = 98.910 ms) and the auditory leads for /ba/ (M = -21.151 ms). Also, /tha/ required a visual lead as compared to the auditory lead for /ba/. There was not a significant main effect of Gain [F(1, 16) = 1.910, p = 0.186, η^2_p = 0.107], but the interaction between these two factors was significant [F(1.374, 21.985) = 16.386, p < 0.001 η^2_p = 0.506]. For the syllable /fa/, there was a larger visual lead obtained for high gain conditions (M = 172.582 ms) than for low gain conditions (M = 128.746 ms). For the syllable /tha/, the opposite pattern was observed with the visual lead being larger

for low gain conditions (M = 163.956 ms) as compared to the visual lead for high gain conditions (M = 33.864 ms; see Figure 12). No differences were obtained for /ba/.

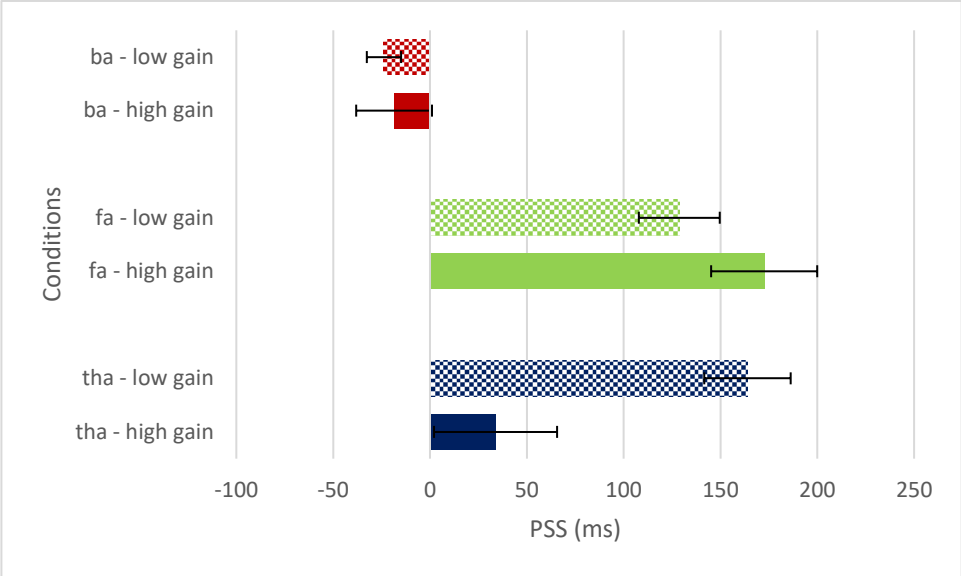

Figure 12. The Point of Subjective Simultaneity (PSS) values, averaged across participants, from the data obtained in Experiment 2. Negative values on the x-axis represent "audition first" conditions, positive values on the x-axis represent "vision first" conditions. Low gain stimuli are represented by pattern-filled bars, high gain stimuli are symbolized with bold, solid bars. Red bars represent data for /ba/, green bars represent the data obtained for /fa/, blue bars represent the data for /tha/. Error bars represent the standard errors of the means.

We, subsequently, calculated the TWIs for every experimental condition based on participants' JND and PSS scores. The amount of auditory lead (indicated in this case by negative values) is calculated by subtracting the JND from the PSS value, while the amount of visual lead (indicated in this case by positive values) is calculated by adding those two values. The total duration of the TWI is then calculated as the sum of the absolute positive and negative values, mentioned above (see Figure 13). The TWI for the syllable /ba/ appears to be extended in high (TWI = 408.07 ms) as compared to low gain conditions (TWI = 293.75 ms). The same was evident for the syllables /fa/ (TWI = 427.57 ms in high as compared to 413.39 ms in low gain conditions) and /tha/ (TWI = 443.10 and 386.59 ms for high and low gain conditions, respectively). This finding suggests that participants were more tolerant to asynchronies between the auditory

and visual components for high gain stimuli as compared to low gain stimuli, whereby, participants exhibited higher sensitivity in detecting asynchronies between the two sensory components.

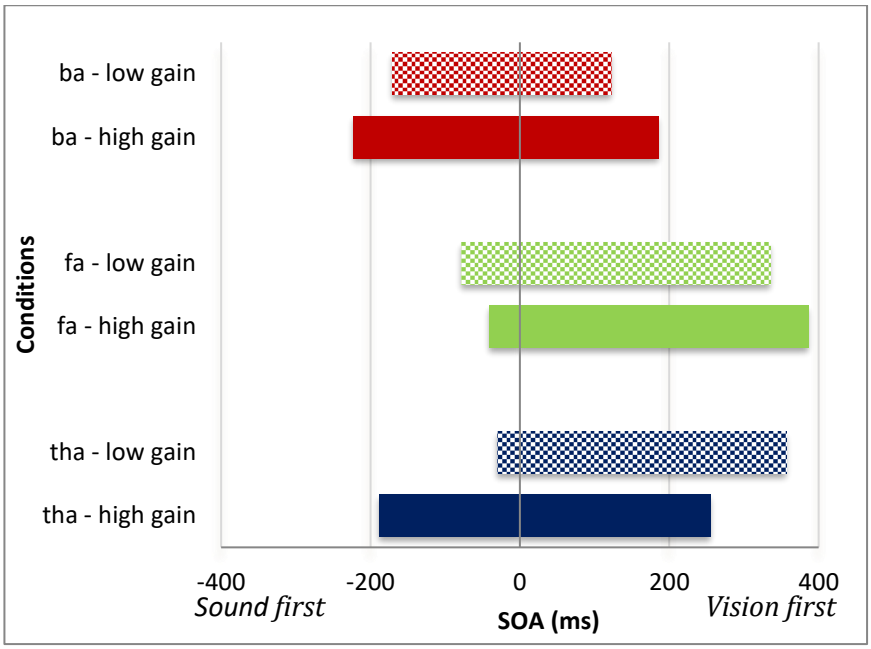

Figure 13. The Temporal Window of Integration (TWI) across experimental conditions, averaged across participants. Low gain stimuli are depicted with pattern filling, high gain stimuli are filled with colour. Negative values indicate that the auditory stream preceded the visual. The syllable /ba/ is represented with red bars, /fa/ with green bars and /tha/ with blue bars.

3. General Discussion

In the present study, we first examined whether the principle of inverse effectiveness would be demonstrated behaviorally with the use of AV speech stimuli. We utilized three simple consonant-vowel (CV) syllables, presented visually, aurally, or audiovisually. Naturalistic degradation was implemented in both the visual and auditory streams of the speech event. The indices used for the calculation of multisensory gain showed that the magnitude of gain depended on the visual and auditory noise levels that were combined, on the presented syllable (i.e., the specific characteristics of a speech event), and on the index used for its calculation (Experiment 1). After specifying the AV combinations that led to the highest and lowest levels of multisensory gain, we examined whether the width of the TWI would be modulated as a

function of gain. We did this in order to investigate if multisensory perception is governed by interactions between the temporal rule and inverse effectiveness (Experiment 2). We found that participants were more tolerant to asynchronies between the two streams when high gain stimuli were presented, as compared to when low gain stimuli were presented, therefore our findings support that the width of the TWI for AV speech events can be modulated as a function of multisensory gain. This interaction between timing and effectiveness was demonstrated for the first time in the present study with the use of AV speech stimuli.

The findings from Experiment 1 did not altogether support the behavioral extension of the principle of inverse effectiveness. In particular, we noted that gain was minimized rather than maximized at high levels of noise as indicated by three of the indices employed for the calculation of gain. Specifically, gain was minimized when the auditory stream was maximally degraded as was indicated by the MSI and Absolute Difference indices, while visual noise did not affect the levels of multisensory gain. Gain as calculated by the Contrast index was also minimized when the auditory stream was maximally degraded, and this effect was evident across every level of visual noise. Similar patterns have been reported by Ross et al. (2007) and Nahanni (2014), but with the use of different metrics for the calculation of gain. More specifically, Ross et al. reported that the magnitude of gain was maximized at the highest SNRs (i.e., when the auditory stream was of lowest noise) as indicated by the Normalized Enhancement index and Nahanni found that gain peaked at the highest SNRs, when gain was calculated by the Visual Enhancement and Normalized Enhancement indices. The maximization of gain at intermediate levels of noise observed in Ma et al.'s (2009) and Ross et al.'s studies with the use of the Absolute Difference (in %) index, was not obtained by any calculated index in our study. Thus, neither the principle of inverse effectiveness, nor the phenomenon of stochastic resonance can fully explain the pattern of our results. Based on our findings, we support that there is an evident dependency of the magnitude of gain on the physical characteristics of the presented stimulus. This is supported by the main effect of the syllable on the magnitude of gain, obtained by each calculated index in our study. Therefore, in order to fully characterize the processes that underlie multisensory integration, it is necessary to take into account the physical differences between the presented stimuli, given that the physical characteristics of a stimulus define its reliability. This is in line with the "optimal integration hypothesis" in

multisensory perception, whereby each sensory signal has an estimated weight determined by its reliability, and the final multisensory percept is estimated by the summation of the weighted signals and dominated by the most reliable sensory signal (Ernst & Banks, 2002; Ernst & Bülthoff, 2004).

Gain reflects the (relative or absolute) enhancement from unisensory to multisensory conditions. Therefore, one possible account for the maximization of gain levels when both streams were least degraded (as indicated by the MSI, Contrast, and Absolute Difference indices), would be that participants performed poorly for unisensory conditions of low noise, while the concurrent presentation of both streams led to performance enhancement. However, this was not the case, since participants' performance for VO and AO conditions at low levels of noise was highly accurate. The sole exception to this finding was the accuracy obtained for the syllable /tha/ at the two lowest levels of auditory noise. Participants' performance declined for these conditions (and, generally, for AO presentations of the syllable /tha/, with the lowest accuracy being paradoxically obtained when the auditory stream was of low noise). For /tha/, we did observe that there were great levels of performance enhancement for AV as compared to AO conditions. However, three of the indices in our study (MSI, Contrast, and Absolute Difference) were calculated based on the most effective unisensory response (i.e., in this case, the VO). Given that participants were highly accurate for VO presentations of /tha/ at the two lowest levels of noise, the magnitude of gain in these conditions reflects the difference between AV and VO conditions. Hence, the maximization of gain at the lowest visual and auditory noise levels cannot be attributed to poor performance for unisensory conditions of low noise. Thus, participants benefited more from AV stimulus presentations when both streams were highly reliable. It should be emphasized, however, that this finding is counterintuitive and contradicts the existence of an inverse relationship between multisensory gain and the quality of the unisensory streams of a speech event, when this relationship is examined behaviorally.

Another possible explanation for the pattern of gain that was observed with the MSI, Absolute Difference, and Contrast indices (i.e., minimization of gain at highest noise levels) could be that participants were more accurate in identifying the three syllables under conditions of high noise for the most effective unisensory stimulus presentations (i.e., AO or VO) than for AV stimulus combinations of high noise (given that the most

effective unisensory response is taken into account for the calculation of these indices). This pattern could indicate that the synchronous presentation of the two sensory streams when they were of high noise might have led to a decline (instead of enhancement, as predicted by inverse effectiveness) in participants' performance. In this case, the deterioration of AV performance could be attributed to high levels of uncertainty by the concurrent presentation of the visual and auditory streams of a speech event. In these cases, performance for AV conditions would be poorly (if any) enhanced, as compared to the performance for the most informative unisensory condition. For the syllables /ba/ and /fa/, participants' performance was more resistant to visual than to auditory noise, and therefore, the VO scores were employed for the calculation of gain by the MSI, Contrast and Absolute Difference indices. For AV combinations of high noise performance was better for VO than for AV presentations, resulting in the low magnitude of gain obtained in these conditions. This effect was not observed for /tha/, whereby performance was more accurate for AV combinations of high noise than for the respective unisensory stimulus presentations. Interestingly, when examining the magnitude of gain for each of these syllables separately, /tha/ was the only syllable for which minimization of gain at the highest noise levels was not observed. According to Binnie, Montgomery, and Jackson (1974), the ability to identify speech stimuli is modulated depending on the reliability of the auditory and visual streams of speech stimuli presented in noise. Audition provides poor information about the place of articulation of consonants when they are presented in noise (Binnie et al., 1974; Miller & Nicely, 1955), while people can reliably identify syllables when they are presented visually based on their place of articulation (Binnie et al., 1974). On the contrary, audition provides more reliable information as for the manner of articulation as compared to vision (Binnie et al., 1974; Vatakis et al., 2014). We suggest that for conditions of high noise, participants mostly relied on the place of articulation for the identification of the syllables /ba/ and /fa/, which are highly visible given that their articulation requires highly salient lip movements (/ba/ is a bilabial and /fa/ is a labiodental consonant as for their place of articulation) and lips are the most visible articulator (Flahire & Hodson, 2014), while identification of /tha/ (which is dental and, therefore, its articulatory movements are not as salient as the ones made for /ba/ or /fa/) was enhanced by the concurrent presentation of the auditory speech stream. Therefore, in order to correctly identify the presented syllable, participants relied

mostly on the most informative sensory stream, determined by each syllable's articulatory characteristics (place and manner of articulation), while the parallel presentation of the least informative stream led to a decline in participants' performance.

Gain as calculated by the Absolute Difference (in %) index peaked when the visual stream was of low noise and the auditory stream was of high noise (with the exception of /tha/, whereby combinations of low auditory and visual noise led to maximization of gain). This finding is in line with the complementarity of vision and audition in speech perception as suggested by Massaro (2004, Chap.10), given that contribution of the visual modality reached its peak as the auditory stream became less informative. Similar gain patterns were reported by previous studies (e.g., Erber, 1969, 1971; O' Neil, 1954; Sumby & Pollack, 1954). These findings, while considered to be in accordance with the principle of inverse effectiveness given the maximization of gain at the highest levels of noise, lacked the manipulation of the visual speech stream. It should be underlined that previous neurophysiological research has suggested the maximization of multisensory enhancement when both streams are equivalently and maximally degraded (see Perrault, Vaughan, Stein, & Wallace, 2005; Stanford, Quessy, & Stein, 2005; Stanford & Stein, 2007). Therefore, the findings from studies that have neglected the effectiveness manipulation of one sensory stream should be interpreted with caution when it comes to the extension of the principle of inverse effectiveness to a behavioral level. In our study, the reliability of the visual modality in speech perception is highlighted by participants' higher accuracy scores for VO as compared to AO presentations in our study. Moreover, participants' performance for VO conditions was resistant to noise, with accuracy scores declining only at the peak of visual noise, while for the auditory channel the effect of noise was more pronounced and had a clear detrimental effect on participants' performance. While the high levels of accuracy in VO conditions support that the presented syllables were of high visual saliency, we do not suggest that the visual modality is generally more informative than the auditory for the identification of speech stimuli. This effect should be attributed to the fact that, in order to avoid any potential ceiling effects, the auditory stream in our study was moderately degraded even at the lowest levels of noise. These findings demonstrate that the visual stream is essential in speech perception and the manipulation of its effectiveness is suggested for future studies in this field.

Holmes (2007) has pointed out that different indices used for the calculation of multisensory gain can result in different patterns even when the same dataset is being used, an effect that was supported by our findings. In our study, we obtained different gain patterns across the combinations of visual and auditory noise, depending on both the calculated index and the presented syllable. More specifically, for the syllable /ba/, the Contrast index suggested that gain was minimized at the highest auditory noise levels, independent of visual noise. Similarly, the MSI and Absolute Difference indices suggested that gain was minimized when the auditory stream was of the highest noise, while the Absolute Difference (in %) peaked at maximum auditory and minimum visual noise levels. For the syllable /fa/, no differences in the magnitude of gain were obtained with the use of the Contrast, MSI and Absolute Difference indices, while the Absolute Difference (in %) peaked at maximum auditory and minimum visual noise levels, as was observed for the syllable /ba/. Finally, for the syllable /tha/, there were no gain differences indicated by the Contrast, MSI and Absolute Difference indices, while the Absolute Difference (in %) was maximized when both streams were of low noise. These findings indicate that the most major difference between the patterns of gain for each syllable was observed amongst the MSI, Absolute Difference and Contrast indices with the Absolute Difference (in %) index. This effect should be attributed to the fact that for the three first indices, the most reliable unisensory response was accounted for the final calculation, while for the latter index the response for auditory conditions was accounted across every experimental condition, in order to isolate the contribution of the visual signal for the identification of each AV syllable. Therefore, the different gain pattern obtained by the Absolute Difference (in %) index, in relation to the patterns obtained by the MSI, Absolute Difference and Contrast indices, can be attributed to the different term that was employed for its calculation.

The interaction between effectiveness and timing was demonstrated for the first time in our study. While a number of factors that modulate the width of the TWI have been specified based on previous studies (see Vatakis, 2013; Vatakis & Spence 2006), the levels of effectiveness of the auditory and visual speech streams are for the first time shown to affect its width. More specifically, the width of the TWI was significantly extended for high gain, as opposed to low gain conditions as indicated by participants' JND and PSS values. This finding suggests that in high gain conditions, participants exhibited increased tolerance to asynchronies between the sensory components of the

speech event, as compared to their increased sensitivity in asynchrony detection for low gain stimuli. Thus, if the auditory and visual streams of an AV speech event are robustly integrated, their integration will not be disrupted over larger temporal windows. This finding provides novel insight in regard to the processes that govern multisensory perception and the behavioral outcomes of such processes. Multisensory events that occur in the physical world are characterized by parallel modulations of both the factors of timing and effectiveness. Therefore, the modulation of multisensory gain based on their interaction would be essential for the integration between distinct sensory events that originate from the same underlying multisensory event and occur in ecologically valid environments.

In Grant and Greenberg's (2001) study, when unisensory streams of low intelligibility were synchronously presented, participants' performance was highly enhanced. Intelligibility rates were similar as those obtained in synchrony when high gain stimuli were presented at various SOAs, as long as the visual stream led the auditory by up to 160-200 ms. Their findings highlighted that the integration between the auditory and visual components was robust for high gain stimuli, as participants' performance was similar to that obtained when the two streams were presented in synchrony. Our findings are in line with their findings, since high gain stimuli led in participants' increased tolerance to asynchronies between the two sensory streams. However, Grant and Greenberg did not investigate if the TWI would be modulated for low gain stimuli as well. Furthermore, they only used stimuli of low intelligibility rates (that led to high levels of gain when they were combined), and they did not compare the width of the TWI for both high and low gain stimuli, which was attempted in Experiment 2. In our study, high gain coincided with high accuracy scores for the respective AV stimuli (except for the syllable /tha/, for which higher accuracy was obtained for the low gain AV combination), and we found that the TWI expanded for high gain stimuli. Therefore, our findings are not in line with Conrey and Pisoni's (2006) findings, whereby participants with higher intelligibility scores in AO and AV conditions exhibited increased sensitivity in detecting asynchronies between the two sensory streams of an AV speech event. However, it should be underlined that we did not examine participants' sensitivity in detecting AV asynchronies as a function of their accuracy scores (which would be analogous to intelligibility scores tested in Conrey and

Pisoni's study). Instead, we examined their TOJ performance as a function of gain and, thus, the two studies are not directly comparable.

Our JND analysis showed that participants' discrimination accuracy depended on the presented syllable, with participants finding it harder to correctly identify which stimulus stream preceded the other when the syllables /fa/ and /tha/ were presented, as compared to when the syllable /ba/ was presented. The lack of interaction between the syllable and the gain reflected that the effect of gain on participants' JND values was similarly pronounced for every presented syllable. We suggest that the effect of syllable on participants' JNDs can be attributed to the different characteristics between the three syllables that were presented in Experiment 2. Vatakis et al. (2012) have demonstrated that the physical characteristics of speech stimuli (i.e., articulatory movements made during speech production) can modulate the temporal perception of AV speech events. In their study, participants performed a TOJ task, whereby the auditory and visual components of speech events (i.e., syllables consisting of one consonant and one vowel) were presented at various SOAs and participants were asked to perform a TOJ task. Vatakis and her colleagues found that participants' sensitivity was modulated as a function of the visual saliency of each syllable, with higher sensitivity being reported for syllables of high visual saliency. In our experiment, participants' JNDs indicated an increased sensitivity in detecting asynchronies between the two sensory streams when the syllable /ba/ was presented, as compared to when /fa/ and /tha/ were presented. Although it is beyond the scope of our study to analyze participants' PSS values, it should be noted that the PSS values obtained in our study are in line with those reported by Vatakis and her colleagues. The present study indicated that significantly higher visual leads were required for the two streams to be perceived as synchronous (i.e., for the PSS to be achieved) for the syllable /fa/ as compared to the auditory lead required for /ba/ and the visual lead required for /tha/. Vatakis et al. found that higher visual leads were required for labiodental as compared to dental fricative consonants, as was also suggested by our findings (higher visual leads for /fa/, as compared to /tha/). Furthermore, they found that only when bilabial stop consonants were presented, a small auditory (instead of visual) lead was required for the PSS to be achieved, as was indicated by our findings for the syllable /ba/. This finding highlights the increased levels of visual saliency of the syllable /ba/ (and, thus,

increased reliability on the visual signal due the richness of articulatory information), which can in turn account for the smaller TWI obtained for this syllable in our study.

Overall, the findings from the two experiments conducted in the present study suggest that both the magnitude of multisensory gain, and the TWI depend on the effectiveness of the sensory streams that constitute an AV speech event. Our findings from Experiment 1 did not altogether support that the principle of inverse effectiveness extends to a behavioral level, given that maximization of multisensory gain was obtained at various combinations of visual and auditory noise levels, depending on the presented syllable and on the index used for its calculation. Thus, our findings suggest that both the indices, and the stimuli used in previous studies of inverse effectiveness might partially add up to the inconsistency that has been systematically reported. In Experiment 2, participants' performance indicated that high gain stimuli are effectively integrated over larger temporal windows, as compared to low gain stimuli, thus highlighting the interdependency between timing and effectiveness. Previous studies have pointed out the interaction between timing and space (e.g., Stevenson et al., 2012), space and effectiveness (e.g., Macaluso et al., 2004; Nidiffer et al., 2016; Royal et al., 2009). The interaction between timing and effectiveness in multisensory perception is supported, for the first time, in our study and these findings add up to the notion that the principles governing multisensory perception are altogether interdependent (Nidiffer et al., 2016). Future studies will have to be conducted to further characterize this interdependency and examine the extent to which these findings generalize, with the use of novel multisensory stimuli.

References

Alexanderson, S., & Beskow, J. (2014). Animated Lombard speech: Motion capture, facial animation and visual intelligibility of speech produced in adverse conditions. *Computer Speech & Language, 28*(2), 607-618. doi:10.1016/j.csl.2013.02.005

Alvarado, J. C., Stanford, T. R., Rowland, B. A., Vaughan, J. W., & Stein, B. E. (2009). Multisensory Integration in the Superior Colliculus Requires Synergy among Corticocollicular Inputs. *Journal of Neuroscience, 29*(20), 6580-6592. doi:10.1523/jneurosci.0525-09.2009

Alvarado, J. C., Stanford, T. R., Vaughan, J. W., & Stein, B. E. (2007). Cortex Mediates Multisensory But Not Unisensory Integration in Superior Colliculus. *Journal of Neuroscience,27*(47), 12775-12786. doi:10.1523/jneurosci.3524-07.2007

Baum, S., Wallace, M., Micheli, C., Thelen, A., & Colonius, H. (2016). Above the Mean: Examining Variability in Behavioral and Neural Responses to Multisensory Stimuli. *Multisensory Research, 29*(6-7), 663-678. doi:10.1163/22134808-00002536

Bell, A. H., Corneil, B. D., Munoz, D. P., & Meredith, M. A. (2003). Engagement of visual fixation suppresses sensory responsiveness and multisensory integration in the primate superior colliculus. *European Journal of Neuroscience, 18*(10), 2867-2873. doi:10.1111/j.1460-9568.2003.02976.x

Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication, 44*(1-4), 5-18. doi:10.1016/j.specom.2004.10.011

Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974). Auditory and Visual Contributions to the Perception of Consonants. *Journal of Speech Language and Hearing Research, 17*(4), 619. doi:10.1044/jshr.1704.619

Callan, D. E., Callan, A. M., Kroos, C., Vatikiotis-Bateson E. (2001). Multimodal contribution to speech perception revealed by independent component analysis: a single-sweep EEG case study. *Cognitive Brain Research, 10*(3), 349—353. doi:10.1016/S0926-6410(00)00054-9

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *NeuroReport, 14*(17), 2213-2218. doi:10.1097/00001756-200312020-00016

Carney, T. (2012). *The Vocoder*. Lecture presented in The University of Sydney, Syndey.

Conrey, B., & Pisoni, D. B. (2006). Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. *The Journal of the Acoustical Society of America, 119*(6), 4065-4073. doi:10.1121/1.2195091

Coren, S., Ward, L. M., & Enns, J. T. (2004). *Sensation & perception* (6th Ed.). Fort Worth: Harcourt Brace.

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

Diederich, A., & Colonius, H. (2004). Bimodal and trimodal multisensory enhancement: Effects of stimulus onset and intensity on reaction time. *Perception & Psychophysics, 66*(8), 1388-1404. doi:10.3758/bf03195006

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of speech and hearing research*, *12*(2), 423–5.

Erber, N. P. (1971). Auditory and Audiovisual Reception of Words in Low-Frequency Noise by Children with Normal Hearing and by Children with Impaired Hearing. *Journal of Speech Language and Hearing Research, 14*(3), 496. doi:10.1044/jshr.1403.496

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*(6870), 429-433. doi:10.1038/415429a

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences, 8*(4), 162-169. doi:10.1016/j.tics.2004.02.002

Ewertsen, H. W., & Nielsen, H. B. (1971). A Comparative Analysis of the Audiovisual, Auditive and Visual Perception of Speech. *Acta Oto-Laryngologica, 72*(1-6), 201-205. doi:10.3109/00016487109122473

Finney, D. J. (1964). *Probit analysis: Statistical treatment of the sigmoid response curve*. London: Cambridge University Press.

Flahire, L. K., & Hodson, B. W. (2014). Speech Sound Disorders: An Overview of Acquisition, Assessment, and Treatment. In *Language Development: Foundations, Processes and Clinical Application* (pp. 185-202). Burlington, MA: Jones and Bartlett Learning.

Frens, M. A., & Opstal, A. V. (1998). Visual-auditory interactions modulate saccade-related activity in monkey superior colliculus. *Brain Research Bulletin, 46*(3), 211-224. doi:10.1016/s0361-9230(98)00007-0

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory Integration of Dynamic Faces and Voices in Rhesus Monkey Auditory Cortex. *Journal of Neuroscience, 25*(20), 5004-5012. doi:10.1523/jneurosci.0799-05.2005

Grant, K. W., & Braida, L. D. (1991). Evaluating the articulation index for auditory–visual input. *The Journal of the Acoustical Society of America, 89*(6), 2952-2960. doi:10.1121/1.400733

Grant, K. W., & Greenberg, S. (2001). Speech intelligibility derived from asynchronous processing of auditory-visual information. In *Proceedings of the Workshop on Audio-Visual Speech Processing (AVSP-2001)*. Scheelsminde, Denmark.

Grant, K. W., & Seitz, P. F. (1998). Measures of auditory–visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America, 104*(4), 2438-2450. doi:10.1121/1.423751

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*(3 Pt 1), 1197–208. doi:10.1121/1.1288668

Grant, K. W., & Walden, B. E. (1996). Evaluating the articulation index for auditory–visual consonant recognition. *The Journal of the Acoustical Society of America, 100*(4), 2415-2424. doi:10.1121/1.417950

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

Holmes, N. (2007). The law of inverse effectiveness in neurons and behaviour: Multisensory integration versus normal variability. *Neuropsychologia*, *45*(14), 3340–3345. doi:10.1016/j.neuropsychologia.2007.05.025

Holmes, N. P., & Spence, C. (2005). Multisensory Integration: Space, Time and Superadditivity. *Current Biology, 15*(18). doi:10.1016/j.cub.2005.08.058

James, T. W., Stevenson, R. A., & Kim, S. (2012). Inverse effectiveness and BOLD fMRI. In *The new handbook of multisensory processes*. Cambridge, MA: The MIT Press.

Jiang, W., Wallace, M. T., Jiang, H., Vaughan, J. W., & Stein, B. E. (2001). Two Cortical Areas Mediate Multisensory Integration in Superior Colliculus Neurons. *Journal of Neurophysiology, 85*(2), 506-522. doi:10.1152/jn.2001.85.2.506

Lukas, S., Philipp, A. M., & Koch, I. (2010). Switching attention between modalities: further evidence for visual dominance. Psychological Research, 74, 255-267. doi: 10.1007/s00426-009-0246-y

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space. *PloS one*, *4*(3), e4638. doi:10.1371/journal.pone.0004638

Macaluso, E., George, N., Dolan, R., Spence, C., & Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: a PET study. *NeuroImage, 21*(2), 725-732. doi:10.1016/j.neuroimage.2003.09.049

Massaro, D. W. (2004). From Multisensory Integration to Talking Heads and Language Learning. In *The Handbook of Multisensory Processes*. Cambridge, Massachussets: The MIT Press.

Massaro, D. W., & Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 9*(5), 753-771. doi:10.1037//0096-1523.9.5.753

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748. doi:10.1038/264746a0

Meredith, M. A., Nemitz, J. W., & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, *7*(10), 3215–29.

Meredith, M. A., & Stein, B. E. (1983). Interactions among converging sensory inputs in the superior colliculus. *Science (New York, N.Y.)*, *221*(4608), 389–91.

Meredith, M. A., & Stein, B. E. (1986a). Visual, auditory and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of Neurophysiology, 56*, 640-662.

Meredith, M. A., & Stein, B. E. (1986b). Spatial factors determine the activity of multisensory neurons in cat superior colliculus. *Brain research*, *365*(2), 350–4.

Meredith, M. A., Stein, B. E. (1987). Multisensory integration in superior colliculus neurons is determined by modality-specific receptive field properties. *Soc Neurosci Abstr* 13:431

Miller, G. A., & Nicely, P. E. (1955). An Analysis of Perceptual Confusions Among Some English Consonants. *The Journal of the Acoustical Society of America, 27*(2), 338-352. doi:10.1121/1.1907526

Moss, F., Ward, L. M., & Sannita, W. G. (2004). Stochastic resonance and sensory information processing: a tutorial and review of application. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, *115*(2), 267–81. doi:10.1016/j.clinph.2003.09.014

Motter, B. C. (1994). Neural correlates of attentive selection for color or luminance in extrastriate area V4. *Journal of Neuroscience, 14*(4), 2178-2189.

Nahanni, C. (2014). *Sources and Correlates of Performance Enhancement in Audiovisual Speech Perception* (Unpublished master's thesis). Queen's University.

Nath, A. R., & Beauchamp, M. S. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *31*(5), 1704–14. doi:10.1523/JNEU-ROSCI.4853-10.2011

Nidiffer, A. R., Stevenson, R. A., Fister, J. K., Barnett, Z. P., & Wallace, M. T. (2016). Interactions between space and effectiveness in human multisensory performance. *Neuropsychologia, 88*, 83-91. doi:10.1016/j.neuropsychologia.2016.01.031

O'Neill, J. J. (1954). Contributions Of The Visual Components Of Oral Symbols To Speech Comprehension. *Journal of Speech and Hearing Disorders, 19*(4), 429. doi:10.1044/jshd.1904.429

Perrault, T. J., Vaughan, J. W., Stein, B. E., & Wallace, M. T. (2005). Superior colliculus neurons use distinct operational modes in the integration of multisensory stimuli. *Journal of neurophysiology*, *93*(5), 2575–86. doi:10.1152/jn.00926.2004

Rabinowitz, W. M., Eddington, D. K., Delhorne, L. A., & Cuneo, P. A. (1992). Relations among different measures of speech reception in subjects using a cochlear implant. *The Journal of the Acoustical Society of America, 92*(4), 1869-1881. doi:10.1121/1.405252

Ross, L. A., Molholm, S., Blanco, D., Gomez-Ramirez, M., Saint-Amour, D., & Foxe, J. J. (2011). The development of multisensory speech perception continues into the late childhood years. *European Journal of Neuroscience, 33*(12), 2329-2337. doi:10.1111/j.1460-9568.2011.07685.x

Ross, L., Saint-Amour, D., Leavitt, V., Javitt, D., & Foxe, J. (2007). Do You See What I Am Saying? Exploring Visual Enhancement of Speech Comprehension in Noisy Environments. *Cerebral Cortex*, *17*(5), 1147–1153. doi:10.1093/cercor/bhl024

Rowland, B. A., & Stein, B. E. (2007). Multisensory integration produces an initial response enhancement. *Frontiers in Integrative Neuroscience, 1*. doi:10.3389/neuro.07.004.2007

Royal, D. W., Carriere, B. N., & Wallace, M. T. (2009). Spatiotemporal architecture of cortical receptive fields and its impact on multisensory interactions. *Experimental Brain Research,198*(2-3), 127-136. doi:10.1007/s00221-009-1772-y

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-Visual Speech Perception and Auditory-Visual Enhancement in Normal-Hearing Younger and Older Adults. *Ear and Hearing, 26*(3), 263-275. doi:10.1097/00003446-200506000-00003

Stanford, T. R., Quessy, S., & Stein, B. E. (2005). Evaluating the operations underlying multisensory integration in the cat superior colliculus. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *25*(28), 6499–508. doi:10.1523/JNEURO-SCI.5095-04.2005

Stanford, T. R., & Stein, B. E. (2007). Superadditivity in multisensory integration: putting the computation in context. *Neuroreport*, *18*(8), 787–92. doi:10.1097/WNR.0b013e3280c1e315

Stein, B. E., & Meredith, M. A. (1994). *The merging of the senses*. London: MIT Press.

Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience, 9*(4), 255-266. doi:10.1038/nrn2331

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

Stevenson, R. A., Bushmakin, M., Kim, S., Wallace, M. T., Puce, A., & James, T. W. (2012). Inverse Effectiveness and Multisensory Interactions in Visual Event-Related Potentials with Audiovisual Speech. *Brain Topography, 25*(3), 308-326. doi:10.1007/s10548-012-0220-7

Stevenson, R. A., Fister, J. K., Barnett, Z. P., Nidiffer, A. R., & Wallace, M. T. (2012). Interactions between the spatial and temporal stimulus factors that influence multisensory integration in human performance. Experimental Brain Research, 219(1), 121-137. doi:10.1007/s00221-012-3072-1

Stevenson, R. A., & James, T. W. (2009). Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. *NeuroImage, 44*(3), 1210-1223. doi:10.1016/j.neuroimage.2008.09.034

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America, 26*(2), 212-215. doi:10.1121/1.1907309

Synface. (n.d.). Retrieved Novemver 20, 2017, from https://www.speech.kth.se/synface/

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in cognitive sciences*, *14*(9), 400–10. doi:10.1016/j.tics.2010.06.008

Tsilionis, E., & Vatakis, A. (2014). Audiovisual Speech Integration in the Brain: Semantics and Temporal Synchrony. *Procedia - Social and Behavioral Sciences, 126*, 160-161. doi:10.1016/j.sbspro.2014.02.354

Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J., & Hale, S. (2010). Aging, Audiovisual Integration, and the Principle of Inverse Effectiveness. *Ear and Hearing,* 1. doi:10.1097/aud.0b013e3181ddf7ff

Vatakis, A. (2013). The Role of Stimulus Properties and Cognitive Processes in the Quality of the Multisensory Perception of Synchrony. In *Handbook of Experimental Phenomenology: Visual Perception of Shape, Space and Appearance* (pp. 243-263). UK: Wiley-Blackwell. doi:10.1002/9781118329016

Vatakis, A., Ghazanfar, A. A., & Spence, C. (2008). Facilitation of multisensory integration by the "unity effect" reveals that speech is special. *Journal of Vision, 8*(9), 14-14. doi:10.1167/8.9.14

Vatakis, A., Maragos, P., Rodomagoulakis, I., & Spence, C. (2012). Assessing the effect of physical differences in the articulation of consonants and vowels on audiovisual temporal perception. *Frontiers in Integrative Neuroscience, 6*. doi:10.3389/fnint.2012.00071

Vatakis, A., & Spence, C. (2006). Audiovisual synchrony perception for music, speech, and object actions. *Brain Research, 1111*(1), 134-142. doi:10.1016/j.brainres.2006.05.078

Wallace, M. T., Carriere, B. N., Perrault, T. J., Vaughan, J. W., & Stein, B. E. (2006). The Development of Cortical Multisensory Integration. *Journal of Neuroscience, 26*(46), 11844-11849. doi:10.1523/jneurosci.3295-06.2006Xu, J., Yu, L., Rowland, B. A., Stanford, T. R., & Stein, B. E. (2012). Incorporating Cross-Modal Statistics in the Development and Maintenance of Multisensory Integration. *Journal of Neuroscience, 32*(7), 2287-2298. doi:10.1523/jneurosci.4304-11.2012

Xu, J., Yu, L., Rowland, B. A., & Stein, B. E. (2017). The normal environment delays the development of multisensory integration. *Scientific Reports, 7*(1). doi:10.1038/s41598-017-05118-1

Xu, J., Yu, L., Stanford, T. R., Rowland, B. A., & Stein, B. E. (2014). What does a neuron learn from multisensory experience? *Journal of Neurophysiology, 113*(3), 883-889. doi:10.1152/jn.00284.2014

Yu, L., Rowland, B. A., & Stein, B. E. (2010). Initiating the Development of Multisensory Integration by Manipulating Sensory Experience. *Journal of Neuroscience, 30*(14), 4904-4913. doi:10.1523/jneurosci.5575-09.2010

Yu, L., Xu, J., Rowland, B. A., & Stein, B. E. (2013). Development of cortical influences on superior colliculus multisensory neurons: effects of dark-rearing. *European Journal of Neuroscience, 37*(10), 1594-1601. doi:10.1111/ejn.12182

Appendix

Magnitude of multisensory gain at every experimental condition as calculated by every index

Table 1.

Gain as calculated with the Contrast index for each syllable at the various combinations of visual and auditory noise levels

Visual Noise Levels	Auditory Noise Levels	Contrast					
		/ba/		/fa/		/tha/	
		Mean	SD	Mean	SD	Mean	SD
1	1	-.002	.011	.000	.023	.012	.131
	2	-.010	.021	-.004	.034	.004	.127
	3	-.010	.026	-.004	.030	.023	.114
	4	-.104	.122	-.003	.036	.016	.120
2	1	-.015	.030	.002	.026	.017	.060
	2	-.005	.028	.000	.023	.011	.067
	3	-.009	.045	-.007	.047	.020	.055
	4	-.290	.270	-.008	.039	.012	.068
3	1	-.006	.022	.016	.048	.022	.179
	2	-.007	.031	-.007	.076	.042	.129
	3	-.036	.105	-.025	.098	.033	.150
	4	-.506	.302	-.045	.109	.013	.102
4	1	-.005	.015	-.049	.166	-.019	.396
	2	.004	.055	-.077	.153	.037	.293
	3	-.036	.320	-.040	.201	.050	.311
	4	-.522	.581	-.088	.283	.001	.230

Table 2.

Gain as calculated with the Absolute Difference (in %) index for each syllable at the various combinations of visual and auditory noise levels

Visual Noise Levels	Auditory Noise Levels	Absolute Difference (%)					
		/ba/		/fa/		/tha/	
		Mean	SD	Mean	SD	Mean	SD
1	1	-.476	2.182	11.587	19.877	72.381	15.280
	2	1.11	9.447	12.381	17.611	68.254	21.229
	3	40.793	34.008	35.555	23.126	60.634	19.282
	4	73.492	17.207	57.619	23.597	48.571	19.764
2	1	-2.857	5.606	12.063	18.929	74.285	12.655
	2	1.111	9.447	13.650	16.630	70.476	22.017
	3	39.047	32.389	34.761	21.358	61.111	15.682
	4	53.015	28.282	56.190	21.737	49.047	20.169
3	1	-1.269	4.146	12.063	19.957	64.285	18.443
	2	.634	9.753	8.888	20.475	66.190	21.325
	3	27.460	32.470	26.031	23.678	53.968	19.079
	4	23.968	26.407	45.396	29.011	41.428	22.597
4	1	-.952	3.007	-1.904	26.761	34.444	27.413
	2	.634	9.753	-6.349	23.802	36.825	26.801
	3	-7.142	26.859	10.793	31.568	30.158	27.517
	4	-.634	13.359	23.968	32.311	16.825	27.089

Table 3.

Gain as calculated with the Multisensory Integration (MSI) index for each syllable at the various combinations of visual and auditory noise levels

Visual Noise Levels	Auditory Noise Levels	MSI (%)					
		/ba/		/fa/		/tha/	
		Mean	SD	Mean	SD	Mean	SD
1	1	-.476	2.182	.105	4.725	8.015	45.687
	2	-1.905	4.023	-.634	6.854	5.039	35.772
	3	-1.852	5.250	-.687	5.978	9.748	44.106
	4	-16.931	18.548	-.370	7.038	8.373	44.807
2	1	-2.857	5.606	.634	5.300	4.338	14.009
	2	-.847	5.620	.105	4.725	3.280	15.146
	3	-1.534	8.724	-1.005	9.005	4.801	12.975
	4	-38.894	31.270	-1.322	7.606	3.492	15.271
3	1	-1.270	4.146	3.879	11.236	16.018	70.093
	2	-1.376	5.846	-.525	14.253	13.399	35.416
	3	-4.902	22.197	-3.100	20.863	12.619	38.958
	4	-61.799	28.989	-6.521	22.202	4.933	22.321
4	1	-.952	3.007	-5.625	25.230	39.342	139.296
	2	1.587	12.879	-11.181	23.961	34.790	104.667
	3	-8.148	39.988	-1.118	36.690	45.733	130.375
	4	-50.340	88.989	-6.547	41.658	14.331	67.868

Table 4.

Gain as calculated with the Absolute Difference index for each syllable at the various combinations of visual and auditory noise levels

Visual Noise Levels	Auditory Noise Levels	Absolute Difference					
		/ba/		/fa/		/tha/	
		Mean	SD	Mean	SD	Mean	SD
1	1	-.004	.021	.000	.044	.012	.189
	2	-.019	.040	-.007	.065	.000	.173
	3	-.019	.051	-.007	.057	.031	.158
	4	-.169	.184	-.004	.066	.017	.171
2	1	-.028	.056	.004	.049	.030	.105
	2	-.009	.053	.000	.044	.020	.116
	3	-.017	.084	-.011	.084	.034	.096
	4	-.831	2.210	-.014	.072	.020	.120
3	1	-.012	.041	.028	.084	.019	.236
	2	-.014	.057	-.014	.127	.061	.175
	3	-.066	.169	-.046	.161	.047	.214
	4	-.507	.267	-.074	.168	.023	.159
4	1	-.009	.030	-.058	.209	.025	.370
	2	.006	.097	-.103	.217	.065	.313
	3	-.090	.252	-.031	.269	.077	.346
	4	-.127	.165	-.071	.282	.015	.250