The University of Birmingham

MSc Psychology

Project Title

Effect of Speech Rate and Overlapping on Multimodal Language Processing: Evidence from Eye Movements

Student Name

Alper Kumcu

Supervisor name

Robin Thompson

September 2014

Word count

5749

ABSTRACT

Previous research had shown that processing non-verbal visual stimuli such as pictures and objects is guided by spoken language. However, little is known about how verbal auditory and verbal visual stimuli are processed simultaneously. An experiment using eye-tracking methodology was conducted in an attempt to address this question along with the effect of processing difficulty and overlapping between auditory and visual input. Speech rate was manipulated as regular and fast to construct two different processing difficulty conditions. Sentences in the visual input were grouped into four types as to their overlapping degree with the speech and embedded in a relatively natural and contextual presentation. It was found that dwell time percentage increased with the overlapping degree but it was not modulated by speech rate. Results indicated that there is a consistent and dynamic interaction between visual and auditory language.

1. Introduction

Simultaneous reading and listening is a frequent part of every day life. When watching a movie with subtitles or during a presentation with slides at a lecture, we are surrounded with constant, contextual and rapid auditory and visual language, which will be referred to as "multimodal language", henceforth. Information coming from both streams is expected to interact on a basis of linguistic and processing variables. However, we know surprisingly little about the nature of this dynamic interaction and concerning factors, despite the vast amount of effort devoted to multimodal perception and processing. Further, laboratory research generally treats language as isolated units (e.g. single words or sentences per trial), although multimodal language processing takes place in complex natural settings. The aim of the present study is to examine multimodal language processing wrapped in contextual verbal input and also factors that might have an impact on the task.

1.1. Auditory and visual speech

Studies from audio-visual (AV) speech, which focus on the role of non-linguistic visual input during spoken language processing, have well established that AV language processing is different from audio-only language processing (Tuomainen, Andersen, Tiippana, & Sams, 2005). In one of the earliest studies in this context, McGurk and MacDonald (1976) showed that being subjected to conflicting lip movements and phonemes leads to an auditory illusion, known as McGurk effect. McGurk effect pointed out that language is multimodal in nature. It also showed that auditory and visual inputs are automatically and constantly integrated. Following this line of research, a great deal of evidence has supported the assumption that non-linguistic visual cues enhance the intelligibility of spoken language processing (Wagner, Malisz, & Kopp, 2014 for gestures; Kim & Davis, 2004 for face; Arnold & Hill, 2001 for lip reading and Ma, Zhou, Ross, Foxe, & Parra, 2009 for lip reading in noisy environment).

1.2. Processing difficulty in reading and eye movements

Findings from AV speech research pinpoint the fact that multimodality is an important constituent of language processing. However, AV speech studies gauge language processing mostly at phonemic level. Further, the target inquiry of this field is spoken language. In contrast, reading is a different behaviour and the language processing is inextricably associated to vision in reading (Ferreira & Tanenhaus, 2007). In line with our research direction, eye-tracking paradigm was selected as the main approach on several accounts: Eye-tracking has been regarded as the golden standard of online visual processing for more than two decades. Eye movements are part of the natural reading processing and eye-tracking provides researchers a relatively natural, on-line and moment-to-moment processing environment (Rayner & Pollatsek, 2006). Reading research with eye-tracking paradigm has clearly shown that people do not read in a continuously smooth fashion; but rather, they make short and rapid jumps (also called saccades) interrupted with stops (also called fixations) as they acquire information from the text. Fixations are associated with processing difficulty in reading. That is, as the text becomes more difficult, fixation durations increase. Word length (Kliegl, Grabner, Rolfs, & Engbert, 2004), word frequency (Inhoff & Rayner, 1986; Rayner & Duffy, 1986) and age of acquisition (Juhasz & Rayner, 2006) are among the well-established lexical properties that modulate text difficulty and therefore,

fixation measures. Intuitively, the link between cognition and fixation is very straightforward in reading. Reading comprehension involves visual attention, and visual attention requires fixation (Boland, 2004). This linkage between reading and mental processing is known as eye-mind hypothesis (Just & Carpenter, 1980). Eye-mind hypothesis simply assumes that if readers are looking at a particular word, they are mentally processing it. Therefore, increase in fixation measures are thought to manifest incremental cognitive effort. Dwell time (total duration of fixations) is a very common fixation measure and it was selected as the main response variable in this study to make inferences as to cognitive effort.

Reading, however, is not limited to recognizing individual words. There are higher-order processes of comprehension covering syntactic, semantic and pragmatic units (Kuperman & Van Dyke, 2011). Syntactic complexity consistently increases processing difficulty. It was found that embedded structures (longer sentences with subordinate clauses) are more demanding than non-embedded structures but no such evidence was found for passive vs. active sentences (Huestegge & Bocianski, 2010). Syntactic processing is also governed by lexical variables to a certain extent (Deutsch, Bentin, & Katz, 1995). So-called high-level factors, such as reader's goals, real-word knowledge and information outside the perceptual span may also have a role in processing and comprehension but eye movement data have its limitations to be able to reveal such factors (Clifton, Staub, & Rayner, 2007). Visual input in this study was constructed by taking the abovementioned lexical and syntactic factors into account.

1.3. Eye movements in auditory and visual processing

Literature is abundant in research on eye movements when people are looking at a scene with relevant objects as they listen to a speech. This paradigm is typically referred to as the "visual world" (see Andersson, Ferreira, & Henderson, 2011 for definition and Huettig, Rommers, & Meyer, 2011 for an exhaustive review). In his pioneering study, Cooper (1974) showed that participants launch fixations and saccades to pictures referred in a story they listen to. It was later observed that fixations to objects are considerably time-locked to the unfolding utterance (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Findings from visual world studies clearly demonstrated that eye movements are mediated with spoken language. However, there is little evidence with regard to spoken language driven eye movements on written text. In such a study (Yang, Chang, Chien, Chien, & Tseng, 2013), two groups of participants, who differed in background knowledge on the subject, were instructed to attend a presentation and a lecture simultaneously. It was concluded that prior knowledge has a role in how students read and listen in an authentic communication setting. Here, it is important to note that reading while listening is a highly demanding activity and limited by the availability of both general and task-specific processing capacity (Road, Medical, & Limits, 1977). Separate lines of evidence on the interface between eye movements and audio-visual processing have come from gesture community as well: For instance, It was demonstrated that fixation patterns are systematically restrained by spoken language when processing co-speech gestures. Incongruent words and gestures affect each other, suggesting that non-verbal visual input processing is automatic and occurs in parallel to processing of speech (Gullberg, 2003). These data hold the assumption that language and vision interact immediately and eye movements are robust indicators of this interaction. However, cognitive and linguistic factors constraining multimodal language processing and cross-modal interaction still remain as critical questions.

1.4. Effects of speech rate and overlapping

Speech rate manipulation in language processing studies with eye-tracking paradigm is rare. It was reported that dwell time allocated to non-verbal visual information along with spoken utterances is sensitive to speech rate (Speed & Vigliocco, 2013). More direct evidence was reported by Andersson et al. (2011). In their visual world study with relatively natural material, speech rate was manipulated as either slow or fast (by decreasing and increasing the original speech rate by 20%). It was found that high-speed rate has a negative effect on the ability to match the referential expressions with their visual counterparts as reflected with dwell time percentage. In a reaction time study; Gibson, Eberhard, and Bryant (2005) probed into the role of speech rate based on the findings of Spivey, Tyler, Eberhard and Tanenhaus (2001). They resulted that the benefit of concurrent visual context when comprehending spoken utterances diminishes as the speech rate increases. Evidence so far has suggested that faster speaking rate increases the processing demand and cognitive effort, which is mirrored through increase in fixation measures.

We also have very limited knowledge about how overlapping between auditory and visual language affect processing. Visual world studies presented evidence that phonetic (Allopenna, Magnuson, & Tanenhaus, 1998), visual (Huettig & Altmann, 2004) and semantic (Huettig & Altmann, 2005) similarity between spoken utterances and visual objects modulate fixations. Huettig and McQueen (2007) tapped into printed words in a visual world study and concluded that fixations are sensitive to phonological overlapping between spoken utterances and printed words as reading provides much more direct access to phonological knowledge. There is no such evidence with regard to overlapping between larger and contextual language segments, to our knowledge.

1.5. Research questions

A seemingly simple question then arises based on the previous research: Does spoken language interact with concurrent written language in a consistent way as it does with non-verbal stimuli? And to what extent does processing difficulty affect this interaction? We reasoned that if there exists a consistent and tight coordination between auditory and visual language, it should be manifested in dwell time and dwell time should be correlated with the overlapping degree. In other words, highly related items should accrue more fixations. We also assumed that increasing speech rate would pose a cognitively demanding condition and that it should be mirrored in fixations on the basis of eye-mind hypothesis. If this is the case, we should also observe increase in dwell time under fast processing condition.

2. Method

2.1. Participants

The experiment was carried out with thirty-nine undergraduate students at the University of Birmingham. 7 participants had to be excluded due to poor calibration accuracy and erroneous stimuli. As a consequence, all subsequent analyses are based on a sample of 32 participants (10 males; M_{age} = 20.3, *SD* = 1.8). All participants reported monolingual native speakers of British English as determined with Language History Questionnaire (version 1.0) (see Li, Sepanski, & Zhao, 2006). They also reported to have normal or corrected-to-normal vision, no speech or hearing difficulties and no history of any neurological disorder. They received either £6 or course credit for participation. All

participants were fully informed about the details of the experimental procedure and gave written consent. Ethical approval for the study was obtained from the Ethics Board of the School of Psychology at the University of Birmingham. Post-experiment debriefing revealed that all participants were naïve to the purpose of the experiment. Participants were randomly and evenly assigned to two experimental condition groups as regular speech rate group (R-SRG) and fast speech rate group (F-SRG) (n = 16, 5 and 4 males, respectively). Groups were matched on age, auditory and visual digit span (AVDS), visuospatial memory span (VMS), self-reported prior knowledge on the subject matter of the stimuli and average fixation duration (AFD) in a self-paced reading (see Table 1).

Table 1

Means (standard deviations) of characteristic measures for participants.

Variable	R-SRG	F-SRG	M_{diff}	$t(30)$	p	d
Age	20.1 (1.6)	20.5 (2.1)	-.31	-.46	.648	0.21
AVDS	7.1 (1.2)	7.1 (1.3)	0	.00	1	0
VMS	5.3 (1.1)	5.4 (0.9)	-.14	-.39	.705	0.11
Knowledge	2.3 (0.7)	1.8 (0.5)	.43	2.03	.052	-0.82
AFD	200.9 (22.2)	211.6(24.5)	-10.6	-1.28	.209	0.43

2.2. Materials

The experimental paradigm consisted of a baseline and a main test. The reading material used in both tests were Microsoft® PowerPoint presentations about life and works of Vincent Van Gogh and Leonardo da Vinci, respectively. Presentations were adapted from a number of different encyclopaedic entries and introductory texts. They did not refer to any time-specific context to enable replication. 20 participants who did not take part in the eye-tracking study rated the presentations on a 5-point scale for their naturalness and difficulty (where a score of 5 was very natural and very difficult). The mean naturalness score was 3.6 (*SD* = 0.8) and the mean difficulty score was 2.1 (*SD* = 0.6) for baseline presentation. The mean naturalness score was 3.5 (*SD* = 0.9) and the mean difficulty score was 3.1 (*SD* = 0.8) for main test presentation.

Concurrent with the reading material for the main test, a speech was digitally pre-recorded (44,100 Hz, 16-bit stereo) in a sound attenuated room by a native female speaker of British English and without a local accent. The record was manipulated with PRAAT software (version 5.3.83, http://praat.org) by increasing the original tempo (181 words/min) by 2% for fast speech (185 words/min) and decreasing the original tempo by 17% for regular speech (150 words/min). Pitch was not affected by the tempo manipulation. As a result, regular speech fell into "average" (125-160 words/min) and fast speech fell into "faster than normal" range (185 words/min) for lectures (Tauroza & Allison, 1990). Difference between the durations of the recordings was reliable [*paired* $t(11) = 10.48$, $p < .001$, $d = 6.32$]. Duration of regular speech was 286,487 ms (approximately 4.7 min) and duration of fast speech was 233,353 ms (approximately 3.8 min).

Baseline presentation consisted of four slides, 10 sentences and 185 words. One slide was added to the beginning and to the end of the presentation to familiarise participants with the testing conditions and to construct a relatively natural input. These slides were not included in the analysis. Slide titles were not analysed, either. Main test presentation consisted of six slides, 18 sentences and 226 words. There were three bulleted sentences on each slide. Two slides were added to the beginning and to the end of the presentation for

the same reason. 18 sentences were equally grouped into four sentence types according to their degree of overlapping with the corresponding sentences in the speech (see Table 2). Exact mapping sentences (EM) in the presentation and the speech were identical. Sentences in the filler words (FW) condition were manipulated by adding gap-filling expressions to the corresponding sentences in the speech. In contrast, sentences in the extra information (EI) condition were manipulated such that sentences in the speech contained more information. Lastly, sentences in the semantic mapping (SM) condition overlapped in meaning but lacked lexical similarity. That is, these sentences were paraphrases of each other. Overall, the degree of overlapping descended in the order of EM > FW > EI > SM. In addition to the mapping items, there were also six extra sentences (ES) in the presentation (one on each slide), which were not available in the speech and 24 non-mapping sentences in the speech (four sentences for each slide), which were not available in the presentation. Extra and non-mapping items were added to the stimuli to provide naturalness and to serve as control (see Appendix for a full list of sentences).

Table 2

Sample experimental sentences in the auditory and visual stimuli.

Sentence type	Speech	Presentation
Exact mapping (EM)	There is much speculation over who the woman is and why she has such a mysterious smile.	There is much speculation over who the woman is and why she has such a mysterious smile.
Filler words (FW)	If truth be told, Leonardo can well be regarded as an inventor who was ahead of his time.	Leonardo can well be regarded as an inventor who is ahead of his time.
Extra information (EI)	During his time in Milan, Leonardo Da Vinci worked on The Last Supper, another notable work of his, which stands out among others in terms of its artistic features.	Da Vinci painted The Last Supper during his time in Milan.
Semantic mapping (SM)	Instead, he took the startling approach of actually observing nature and asking questions about it.	Leonardo's attitude towards science was an observational one based on his queries as to the environment.
Extra sentences (ES)	-	Little is known about Leonardo's early life.
Non-mapping sentences (NM)	His mother moved away shortly after, leaving Leonardo Da Vinci's father to raise him.	-

5 sentence types were balanced on word frequency [$F(4, 13) = 0.81$, $p > .05$, $d = 2.63$], age of acquisition [$F(4, 13) = 1.50$, $p > .05$, $d = 2.28$] and word length (as calculated by word count/character count) [$F(4, 13) = 2.35$, $p > .05$] (see Table 3). Word frequency was expressed by the logarithmic values of occurrences per million in the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995), which is based on approximately 5.4 million words. Age of acquisition (AOA) values were calculated using ratings from Kuperman, Stadthagen-Gonzalez and Brysbaert (2012), which is based on 30,121 words. Sentence length (as calculated by sentence count/word count) was also controlled between four sentence types

excluding extra sentences[1] [$F(3, 8) = 2.79$, $p > .05$, $d = 3.95$]. Sentences were not systematically controlled for syntactic structure but there was no difference in dwell time percentage between simple and complex ($p = .746$) or between active and passive sentences ($p = .660$).

Table 3

Means (standard deviations) of experimental items in the visual input.

Sentence type	Log frequency	Age of acquisition	Word length	Sentence length
EM	2.6 (0.6)	6.3 (1.2)	4.9 (0.7)	17.3 (1.5)
FW	2.9 (0.3)	6.3 (0.7)	4.8 (1.0)	14.6 (0.5)
EI	3.1 (0.4)	5.0 (1.0)	3.1 (0.4)	12.6 (2.3)
SM	5.3 (0.3)	5.9 (0.5)	2.9 (0.2)	13.6 (3.0)
ES	4.7 (0.4)	5.3 (0.8)	3.0 (0.1)	8.1 (1.1)

Low-level visual features were also controlled to provide a familiar, coherent and easy to read material. Presentations were displayed double-spaced in Times New Roman, anti-aliased font; size 40 pt. for titles and 22 pt. for body (see Beymer & Russell, 2008). The number of items in each condition was the same for each participant. In sum, 32 participants read 12 experimental items, grouped into 4 types and under 2 conditions.

2.3. Apparatus

Eye movements were measured using an infrared, desktop eye-tracker, SR EyeLink® 1000, (http://sr-research.com), tracking at 1000 Hz with a range of 32° horizontal and 25° vertical. Software provided by the SR EyeLink® system was used to identify saccades using the thresholds for motion (0.15°), velocity (30°/s) and acceleration (8000°/s2). Approximately 3 characters equalled 1° of visual angle. X and Y coordinates of the participant's point of gaze for the right eye were estimated every millisecond. The host PC that was used to present the visual input was networked to a second PC controlling the eye-tracker. Visual input was presented on a 19" TFT monitor with a resolution of 1280 × 1024 pixels and a refresh rate of 60 Hz. Reading was binocular but the eye-tracker monitored movements of the left eye. Auditory input was presented through Sony MDRNC7B noise-cancelling headphones.

2.4. Procedure

Participants were tested individually. (1) Language History Questionnaire (LHQ) was administered to all participants before experimental tasks. Participants were subjected to (2) auditory and visual digit span test and (3) Corsi Block-tapping test to measure visuospatial memory span. Tests were selected among the test battery of and applied with PEBL (Psychology Experiment Building Language, version 0.13, test battery version 0.7, http://pebl.org). Afterwards, participants were seated 52 cm away from the monitor and their eyes were levelled at the centre of the screen. A chinrest and forehead rest minimized head movements. After they received experimental instructions on the screen, participants were calibrated/validated with a standard nine-point grid for the left eye. They were recalibrated as necessary. Drift correction check was performed between each trial. The average calibration accuracy (the difference between true and measured gaze direction)

[1] Extra sentences served as control and were intentionally shorter. They were controlled within for word length, frequency and sentence length ($ps > .05$).

was 0.3° (*SD* = 0.1°) and the average of maximum calibration error was 0.8° (*SD* = 0.3°) for all sessions. As the values were below the threshold of 1.0° for all trials, they were considered as acceptable. First, participants were asked to read the (4) baseline presentation silently and in a self-paced manner. Then, they were presented the (5) main test presentation and the speech simultaneously at a loudness level to approximately 60dB. They were instructed to attend both stimuli with full concentration for comprehension. A correct response rate of 70% in the (6) comprehension test indicated that participants were engaged fully in the tasks. A typical session lasted approximately 45 minutes. The same experimental steps were followed for all participants in the same order (see Figure 1).

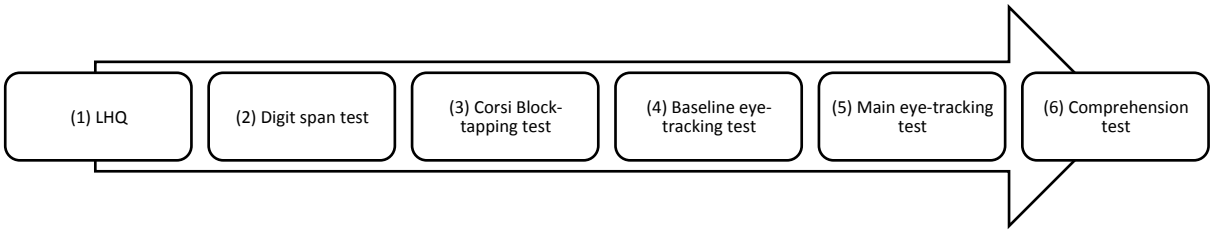

Figure 1. Tests and procedural steps of the experiment.

2.5. Data analysis

Eye movement data were analysed off-line with the EyeLink® Data Viewer software (version 1.11.900). Fixations <40 ms were deleted (<1%). Dwell time percentages, which were equal to 0, were considered as outliers and excluded from the analysis (<2%). No drift correction procedures were applied. The summarized and cleaned data were then transferred to IBM® SPSS Statistics for OSX (version 22.0, Armonk, NY: IBM Corp.) for statistical analyses. 4 frequently reported eye movement measures were analysed, as also discussed and reported in Yang et al. (2013): (1) total dwell time (TDT), (2) dwell time percentage (DTP), (3) average fixation duration (AFD) and (4) fixation count (FC). Further, two temporal measures were used to control the difference in task durations between regular and fast speech conditions: (1) full trial duration (FTD) and (2) interest period duration (IPD) (see Table 4). 18 "look-zones", i.e., areas of interests (AOIs), encompassing each sentence in the presentation (e.g. EM1, FW2, EI3, SM1, ES4 etc.), were defined for item-level analysis. Mean AOI area was 48.290 pixels (*SD* = 13.942) for overlapping sentences. FTDs were trimmed into smaller time periods as IPDs (e.g. EM1, FW2, FW2, EI3, SM1 etc.), so that corresponding visual and auditory items matched (e.g. EM1_AOI - EM1_IP etc.).

Table 4

Definitions for the eye-movement measures used in the experiment.

Eye movement measures	Definition
Total dwell time (TDT)	Sum of durations of all fixations in a trial[a] or AOI
Dwell time percentage (DTP)	Total fixation duration divided by FTD or IPD
Average fixation duration (AFD)	Average duration of all fixations in a trial or AOI
Fixation count (FC)	Number of fixations in a trial or AOI
Temporal measures	**Definition**
Full trial duration (FTD)	Total duration of a trial
Interest period duration (IPD)	Duration of sentence in the speech

Note: [a] Trial in the experiment corresponded to slide in the presentation.

3. Results

3.1. Effect of speech rate

Independent-samples *t*-test was run to determine if there were any significant differences in eye movement measures between regular and fast speech rate groups. The data were normally distributed and there was homogeneity of variances for all variables, as assessed by Shapiro-Wilk test and Levene's test, respectively (*p*s > .05). Analysis on raw data revealed a significant difference in TDT and FC (see Table 5). This is an expected result and only implies that the difference between task durations was reliable. Therefore, DTP was referred to control duration and hence, compare two speech rate conditions. The analysis showed that visual input was viewed slightly longer when speech was slower but the difference was not statistically significant. There was no significant difference in comprehension score between groups.

Table 5

Means (standard errors) of eye movement measures and comprehension score between regular and fast speech rate groups, mean group differences, *t*-scores and significance levels of group comparisons and effect sizes.

Variable	Regular speech	Fast speech	M_{diff}	*t*(30)	*P*	*d*
TDT (ms)	236,034 (29,934)	184,684 (4,801)	51,350	9.08	< .001	3.31
DTP (%)	82 (1)	79 (2)	3	1.37	.182	0.49
AFD (ms)	261.8 (10.2)	255.8 (7.6)	5.9	0.47	.645	0.17
FC	153.7 (16.3)	122.7 (17.7)	31	5.14	< .001	1.87
Comprehension score	9.7 (2.3)	9.8 (1.8)	-.12	-0.16	.868	0.04

3.2. Effect of overlapping degree

Both FTD and IPD were used to calculate DTP. As can be expected, TDT/IPD [$F(3, 90) = 11.09$, $\eta_p^2 = .27$] was more sensitive to the differences between sentence types than TDT/FTD [$F(3, 90) = 6.26$, $\eta_p^2 = .17$). Therefore, only DTP based on IPD (TDT/IPD) was reported in the subsequent item-level analyses.

One-way analysis of variance (ANOVA) was run to determine if there were any significant differences in DTP across sentence types without taking the effect of speech rate into account. The data were normally distributed as assessed by Shapiro-Wilk test ($p > .05$). Homogeneity of variances was violated, as assessed by Levene's test ($p = .012$). DTP was statistically significantly different between different sentence types, Welch's $F(3, 66.419) = 11.76$, $p < .001$, $d = 1.49$. Games-Howell post-hoc analysis revealed that the mean decrease in DTP from EM ($M = 0.56$, $SD = 0.14$) to EI ($M = 0.38$, $SD = 0.10$) was statistically significant ($M_{diff} = 0.17$, 95% CI [0.09, 0.26], $p < .001$), as well as the mean decrease from EM to SM ($M = 0.42$, $SD = 0.20$) ($M_{diff} = 0.14$, 95% CI [0.02, 0.26], $p = .009$) and also from FW ($M = 0.50$, $SD = 0.20$) to EI ($M_{diff} = 0.11$, 95% CI [0.09, 0.22], $p = .030$) (see Figure 2).

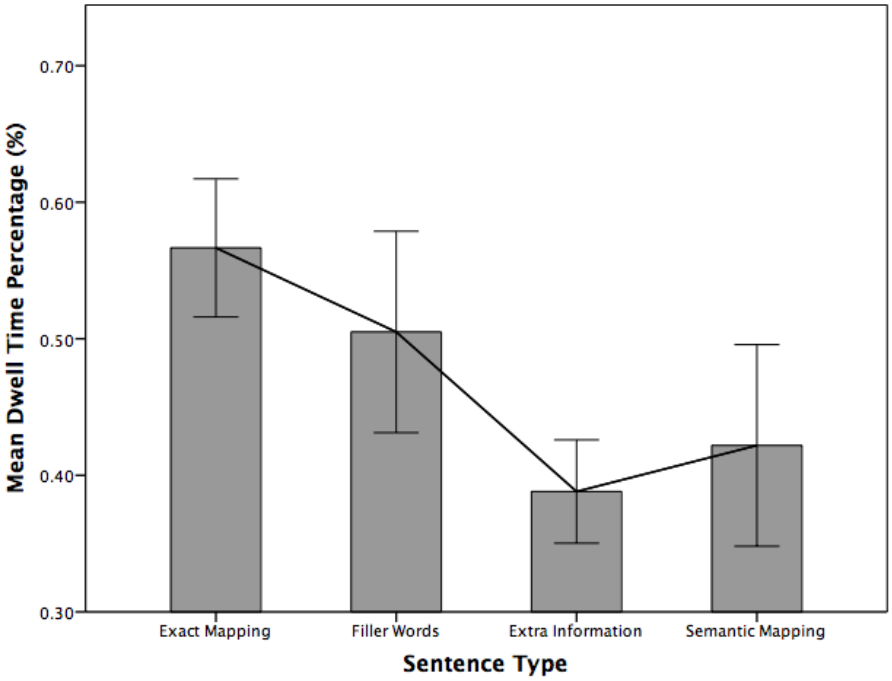

Figure 2. Bar chart showing mean dwell time percentages as per sentence type with error bars showing standard error of means. 95% CI.

3.3. Interaction of speech rate and overlapping degree

A 2 (speech rate: regular and fast) X 4 (sentence type: EM, FW, EI, SM) mixed ANOVA, treating speech rate as between-subjects and sentence type as within-subjects factor was run. The data were normally distributed and there was homogeneity of variances as assessed by Shapiro-Wilk test and Levene's test, respectively (*p*s > .05). The assumption of sphericity was not violated, as assessed by Mauchly's test of sphericity, $\chi^2(2) = 2.64$, $p = .756$. Speech rate X sentence type interaction was not significant, $F(3, 90) = 1.15$, $p = .334$, $\eta_p^2 = .30$.

We assumed that sentence order bias could affect DTP, which would obfuscate the effect of sentence type manipulation. A 2 (speech rate: regular and fast) X 3 (sentence order: first, second and third) mixed ANOVA, treating speech rate as between-subjects and sentence order as within-subjects factor was run to reveal whether there is such an effect. The data were normally distributed and there was homogeneity of variances as assessed by Shapiro-Wilk test and Levene's test, respectively (*p*s > .05). The assumption of sphericity was not violated, as assessed by Mauchly's test of sphericity, $\chi^2(2) = .61$, $p = .738$. Sentence order did have a significant main effect on DTP, $F(2, 60) = 3.94$, $p = .025$, $\eta_p^2 = .68$. Post hoc analysis with a Bonferroni adjustment revealed that DTP significantly increased from sentences in the second order ($M = 0.40$, $SD = 0.11$), to sentences in the third order ($M = 0.50$, $SD = 0.17$), ($M_{diff} = -0.10$, 95% CI [-0.18, -0.01], $p = .023$). However, speech rate X sentence order interaction was not significant, $F(2, 60) = 1.52$, $p = .226$, $\eta_p^2 = .31$, showing that the effect was not different across groups.

Data were further analysed with linear mixed-effects models (LMEMs) to control both by-item and by-participant variation. LMEMs are advocated due to their sensitivity and flexibility as a powerful alternative for typical participants (*F1*) and items (*F2*) testing procedure with ANOVA (Baayen, Davidson, & Bates, 2008). The model used in the analysis

had items and participants as random effects, speech rate and sentence type as fixed effects and DTP as the response variable. Random intercepts (baseline measure) and random slopes (differences between conditions) were entered into the model fit by maximum likelihood[2] (cf. Frisson, Koole, Hughes, Olson, & Wheeldon, 2014). Maximum iterations were changed from default (100) to 150 to match the values those produced in R using the *lme4* package. Correlation matrix and boxplots were used to choose covariance structure and constant covariance was assumed between observation points (Log *L* = 91.772). Results are summarised in Table 6 and in Figure 3. *F* tests of fixed effects again revealed a significant effect of sentence type, $F(3, 352) = 9.34$, $p < .001$; but not a significant main effect of speech rate, $F(1, 32.000) = 0.13$, $p = .722$. Speech rate X sentence type interaction was not significant, $F(3, 352) = 0.78$, $p = .506$. The relationship between regular and fast speech rate with regard to DTP was not significant either (β = 0.05, *SE*= 0.06, $t = .88$, $p = .381$). Our model showed that the effect of overlapping degree on DTP varied, depending on the sentence type. The coefficient for DTP was significantly different in comparing EM with SM, (β = 0.22, SE = .053, $t = 4.10$, $p < .001$) and marginally different in comparing FW with SM (β = 0.10, SE = .053, $t = 1.92$, $p = 0.55$). Post hoc analysis with a Bonferroni adjustment confirmed the significant relationship between EM and EI ($p < .001$), EM and SM ($p < .001$), which were also found with mixed ANOVA. The model also revealed a significant decrease in DTP from EM to FW ($p = .003$).

Table 6

Results summary of coefficient estimates (β), standard errors *SE* (β), associated *t*-score and significance level for all predictors in the analysis.

Predictor	Estimate (β)	*SE* (β)	*t*	*p*
Intercept	0.31	0.43	7.37	< .001
Sentence type				
EM	0.22	0.05	4.10	< .001
FW	0.10	0.05	1.92	.055
EI	0.08	0.05	1.54	.126
SM	[a]			
Speech rate				
Regular	0.05	0.06	.88	.381
Fast	[a]			

Note: [a] Reference condition. SPSS automatically chooses the last category as the reference for categorical variables.

[2] SPSS syntax for the model was as follows:

```
MIXED DTP BY Item Type Rate
  /CRITERIA=CIN(95) MXITER(150) MXSTEP(10) SCORING(1) SINGULAR(0.000000000001)
HCONVERGE(0, ABSOLUTE) LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
  /FIXED=Type Rate | SSTYPE(3)
  /METHOD=ML
  /PRINT=COVB  SOLUTION
  /RANDOM=INTERCEPT Item | SUBJECT(ID) COVTYPE(ID).
```

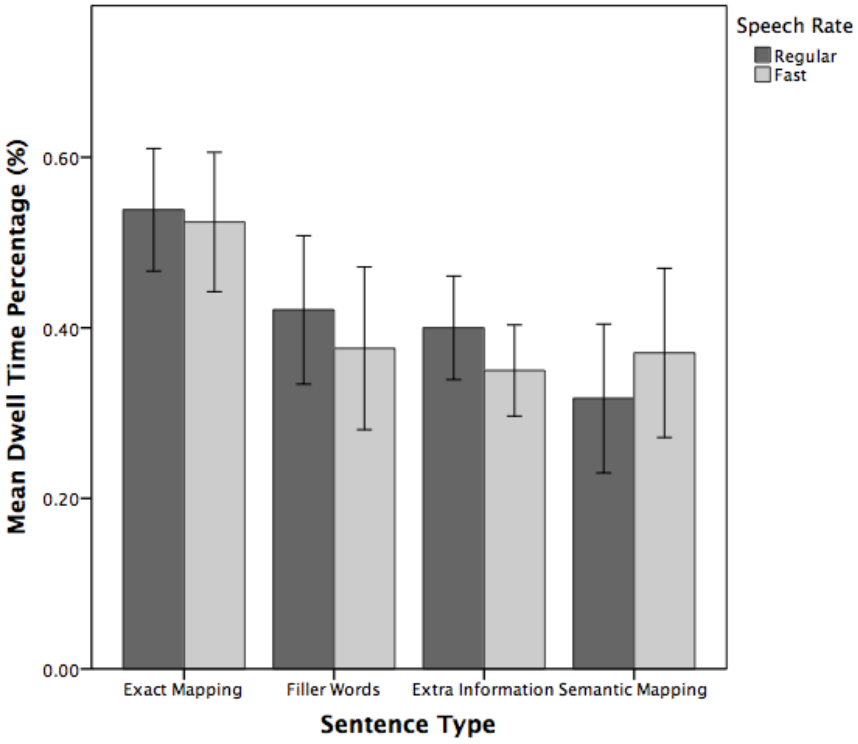

Figure 3. Bar chart showing dwell time percentages across sentence types and speech rate conditions with error bars showing standard error of means. 95% CI.

3.4. Individual differences

Although the primary interest of this study lies in the effect of speech rate and overlapping degree, a Pearson's product-moment correlation was run to assess the relationship between individual difference measures, i.e., auditory and visual digit span, visuospatial memory span, comprehension score and DTP. The data showed no violation of normality, linearity or homoscedasticity as assessed by Shapiro-Wilk test, scatterplots and Levene's test, respectively (*p*s > .05). There was a moderate positive correlation between auditory and visual digit span and visuospatial memory span, as highly expected, *r(*30) = .37, *p* = .005, *d* = .79 and a moderate positive correlation between visuospatial memory span and comprehension score, *r*(30) = .37, *p* = .038, *d* = .79 with visuospatial memory span explaining 13% of the variation in comprehension score. A scatterplot summarises the relationship between (see Figure 4).

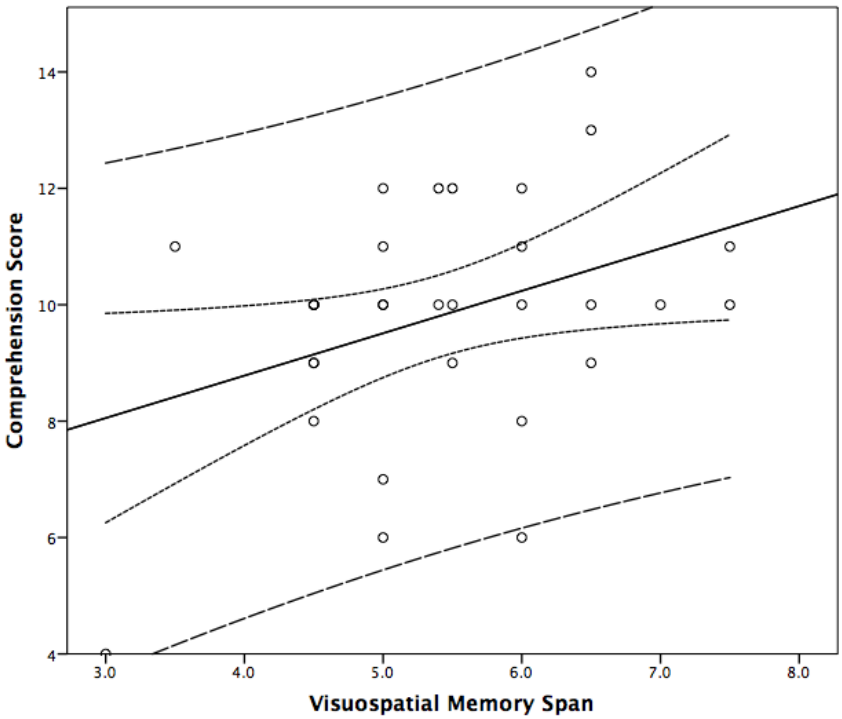

Figure 4. Scatterplot showing the correlation between visuospatial memory span and comprehension score with individual and mean fit lines. CI = 95%

4. Discussion

The goal of the present study was to explore how increase in speech rate as a proxy of processing difficulty and varied overlapping conditions between auditory and visual input would be reflected in eye movements during multimodal language processing. DTP was used to scale total dwell time and as the main indice of cognitive effort within the framework of eye-mind hypothesis (Just & Carpenter, 1980).

Speech rate was manipulated to increase processing difficulty in a systematic and quantitative manner. Prolonged dwell time in correlation with the increase in speech rate was expected, accordingly. However, there was no evidence that readers are systematically affected by the increase in the delivery rate of auditory input. Although not significantly, DTP was higher when the speech was slower. This finding contrasts with the previous research (e.g. Andersson et al., 2011) and the direction of our hypothesis. However, the difference was regarded as a variation rather than a significant effect. One possibility is that the speech rate range in this study was not wide enough to reveal differences in scaled eye movements. There is need for more focused studies with distinctive and varied temporal manipulations (e.g. slower than normal: 100 words/min and faster than normal: 185 words/min speech rates) (Tauroza & Allison, 1990) in order to shed light on the complex interplay between speech rate, processing difficulty and eye movements in multimodal language processing.

Overlapping degree manipulation was introduced to test interaction between auditory and visual language. Item-level analysis revealed a salient relationship between overlapping degree and DTP. Although not reported, DTPs were also significantly lower on extra sentences (i.e., sentences which were available on the presentation but not in the speech)

as compared to mapping sentence types. It suggests the consistency of the coordination between overlapping items. Results indicate that eye movements directed by verbal auditory input have a relatively similar pattern to word-object matching in the visual world. That is, eye movements were sensitive to the similarities between sentences in the presentation and the speech. Further, comparison between full trial duration (i.e., total duration of a slide) and interest period duration (i.e., duration of sentence in the speech) revealed that language processing is also closely time-locked with the verbal information coming from auditory channel (cf. Allopenna et al., 1998). Results also showed that visual and auditory language can be integrated even at the syntactic level and when items are embedded in a larger linguistic context.

Taken together, DTP increased with the overlapping degree. Overlapping effects were most pronounced for sentences in EM condition. LMEMs analysis, along with *F*-tests revealed that DTP was considerably higher when processing identical sentences. Considering the similarity between visual world processing and multimodal language processing, higher DTP associated with EM sentences can also be interpreted in terms of visual, phonological and semantic congruency. FW condition fits into the pattern, further implying the tight audio-visual coordination. Sentences in the EI condition were attended for the shortest duration. Limited cognitive capacity and divided attention might be responsible for this behaviour (Road et al., 1977). That is, when information from auditory and visual channels mismatches, participants might have chosen to attend the speech by abandoning the presentation. Alternatively, they could be searching for mapping information on the slide in an analogy to visual search, which resulted in erratic gaze paths and lower DTP directed to the target area of interest. Considering the existence of comprehension test following simultaneous reading and listening task and participants' goal accordingly, allocating attention to the speech to access more information might be a plausible explanation. Following this line of thought, the question arises whether multimodal language processing represents a cognitive strategy or rather, a low-level, oculomotor routine. While there is evidence against the cognitive strategy approach (Rayner, & Pollatsek, 1989; Mishra, Olivers, & Huettig, 2013) and the participants were found to be naïve to the experimental conditions, there is still room for research to investigate whether and to what extent language-mediated eye movements are consciously controlled when natural and contextual verbal input is concerned.

A further interesting observation is the reverse trend of SM condition. It was expected that DTP would further diminish when there is a lexical mismatch. However, results showed that participants were more likely to detect semantic overlapping even in the absence of lexical, phonological and visual similarity. Conceptual representations activated by semantic associations might be responsible for directing eye movements to the sentences in this set. However, as DTP allocated to SM sentences was not significantly different than EI or FW sentences, we are cautious about interpreting the results. More research addressing the semantic link between auditory and visual language is clearly needed.

Where do readers look when auditory and visual information mismatch? Eye movements during NM sentences (i.e., sentences that were available in the speech but not in the presentation) were not systematically analysed. However, informal observations revealed gaze aversion; that is, participants tended to looked away from the presentation. Intriguingly, non-visual gaze patterns including gaze aversion are linked with working memory (Micic, Ehrlichman, & Chen, 2010). Processing sentences in EI condition can also be interpreted in that vein. Why and where people move their eyes when engaged in a

listening-only task or whether there is a causal link between eye movements and auditory processing is one of the least understood aspects of language processing and it is evident that more research is needed (Boland, 2004).

We did not find a significant interaction between speech rate and overlapping degree. This result is not surprising since there was no significant indication of increased DTP with faster speech. Along with that, SM condition represents the only case in which DTP under fast speech condition was higher than regular speech condition. Why semantic access and cross-modality integration were more efficient with increased processing demand is not clear. Again, the difference was not significant and might be explained with variation.

Lastly, there was a positive correlation between visuospatial memory span and comprehension score. There are studies showing a similar correlation between visuospatial memory and comprehension of paragraphs containing visuospatial language (e.g. De Beni, Pazzaglia, Gyselinck, & Meneghetti, 2005). Although the stimuli used in the experiment did not systematically involve visuospatial language, availability of presentation, a visually rich input, may account for this correlation.

Although every effort was made to control lexical and syntactic discrepancies between the sentence types and their order in the stimuli, there was a trade-off between naturalness and experimental control. Therefore, results should be interpreted cautiously. Fine-grained research is highly needed with more controlled stimuli to more accurately examine the effect of speech rate and the congruency between speech and text. Future studies can aim at integration of auditory and visual language at the morphological and lexical level and additional eye movement measures can be employed accordingly.

Processing multimodal language represents a highly critical skill considering the growing information load and rapid processing demands in the real world. In this respect, our findings may have broader implications within the general architecture of information processing. Here, we present a relatively novel research design that can be adapted to study other aspects of multimodal language. We expect the findings to converge with evidence from similar studies towards constructing a holistic processing model and to further clarify the interface between visual and auditory information processing mechanisms.

Appendix

Type	Speech	Presentation
EM1	Beyond basic reading, writing and mathematical skills; he did not receive much of a formal education.	Beyond basic reading, writing and mathematical skills, he did not receive much of a formal education.
EM2	Throughout his long life, he studied many topics such as anatomy, zoology, botany, geology, optics and aerodynamics among others.	Throughout his long life, he studied many topics such as anatomy, zoology, botany, geology, optics and aerodynamics among others.
EM3	There is much speculation over who the woman is and why she has such a mysterious smile.	There is much speculation over who the woman is and why she has such a mysterious smile.
FW1	As you all know very well; Leonardo Da Vinci is regarded as one of the great creative minds of the Italian Renaissance.	Da Vinci is regarded as one of the great creative minds of the Italian Renaissance.
FW2	As a matter of fact, the masterpiece began to deteriorate even during his lifetime and has undergone an extensive restoration.	The masterpiece began to deteriorate even during his lifetime and has undergone an extensive restoration.
FW3	If truth be told, Leonardo can well be regarded as an inventor who was ahead of his time.	Leonardo can well be regarded as an inventor who is ahead of his time.
EI1	On April 15, 1452, he was born near the village of Vinci about 25 miles west of Florence, Italy and the world was never the same.	He was born on April 15, 1452 in Vinci, Italy.
EI2	Leonardo da Vinci was the illegitimate son of Ser Piero da Vinci, a prominent notary of Florence, a public official who certifies legal documents, and a young peasant woman, Caterina.	Da Vinci was the son of a prominent notary and a young peasant woman.
EI3	During his time in Milan, Leonardo Da Vinci worked on The Last Supper, another notable work of his, which stands out among others in terms of its artistic features.	Da Vinci painted The Last Supper during his time in Milan.
SM1	Instead, he took the startling approach of actually observing nature and asking questions about it.	Leonardo's attitude towards science was an observational one based on his queries as to the environment.
SM2	Probably, the majority of the people on our planet, from an art professional to a high school student, can visualize the painting in seconds.	Mona Lisa can easily be recognized all around the world by people with different levels of knowledge.
SM3	He envisaged a New World shaped by the state of the art devices developed with the knowledge of mechanics.	Leonardo pictured a future created with modern devices to be invented with engineering.
ES1	-	Leonardo spent his first five years in his mother’s home.
ES2	-	Little is known about Leonardo's early life.
ES3	-	He also worked with physics and perspective.
ES4	-	Mona Lisa is on display in the Louvre Museum.
ES5	-	There are many references to The Last Supper.
ES6	-	Some of his drawings have never been recovered.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439. doi:10.1006/jmla.1997.2558.

Andersson, R., Ferreira, F., & Henderson, J. M. (2011). I see what you're saying: The integration of complex speech and scenes during language comprehension. *Acta Psychologica*, *137*(2), 208–16. doi:10.1016/j.actpsy.2011.01.007.

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, *92*(2), 339–355. doi:10.1348/000712601162220.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. doi:10.1016/j.jml.2007.12.005.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database.* [CD-ROM]. University of Pennsylvania, Philadelphia: Linguistic Data Consortium.

Beymer, D., & Russell, D. (2008). An eye tracking study of how font size and type influence online reading. In *People and Computers XXII Culture, Creativity, Interaction (Vol. 2)* (pp. 15–18). Liverpool.

Boland, J. (2004). Linking eye movements to sentence comprehension in reading and listening. In M. Carreiras & C. Clifton, Jr. (Eds.), *The on-line study of sentence comprehension* (pp. 51–76). New York: Psychology Press.

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In R. P.G. Van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye Movements: A Window on Mind and Brain* (pp. 341–371). Oxford: Elsevier.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*(1), 84-107. doi:10.1016/0010-0285(74)90005-X.

De Beni, R., Pazzaglia, F., Gyselinck, V., & Meneghetti, C. (2005). Visuospatial working memory and mental representation of spatial descriptions. *European Journal of Cognitive Psychology*, *17*(1), 77–95. doi:10.1080/09541440340000529

Deutsch, A., Bentin, S., & Katz, L. (1995). Lexical and semantic influences on syntactic processing. *Haskins Laboratories Status Report on Speech Research*. 221–234.

Ferreira, F., & Tanenhaus, M. K. (2007). Introduction to the special issue on language–vision interactions. *Journal of Memory and Language*, *57*(4), 455–459. doi:10.1016/j.jml.2007.08.002.

Frisson, S., Koole, H., Hughes, L., Olson, A., & Wheeldon, L. (2014). Competition between orthographically and phonologically similar words during sentence reading: Evidence from eye movements. *Journal of Memory and Language*, *73*, 148–173. doi:10.1016/j.jml.2014.03.004.

Gibson, B. S., Eberhard, K. M., & Bryant, T. a. (2005). Linguistically mediated visual search: The critical role of speech rate. *Psychonomic Bulletin & Review*, *12*(2), 276–81. doi: 10.3758/BF03196372.

Gullberg, M. (2003). Eye movements and gestures in human face-to-face interaction. In J. Hyönä, R. Radach, & H. Deubel (Eds), *The mind's eyes: Cognitive and applied aspects of eye movements* (pp. 685-703).

Huestegge, L., & Bocianski, D. (2010). Effects of syntactic context on eye movements during reading. *Advances in Cognitive Psychology, 6,* 79–87. doi:10.2478/v10053-008-0078-0.

Huettig, F., & Altmann, G. T. M. (2004). The online processing of ambiguous and unambiguous words in context: Evidence from head-mounted eye-tracking. In M. Carreiras, & C. Clifton (Eds.), *The on-line study of sentence comprehension: Eyetracking, ERP and beyond* (pp. 187-207). New York: Psychology Press.

Huettig, F., & Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition*, *96*(1), B23–32. doi:10.1016/j.cognition.2004.10.003.

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language*, *57*(4), 460–482. doi:10.1016/j.jml.2007.02.001.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: a review and critical evaluation. *Acta Psychologica*, *137*(2), 151–171. doi:10.1016/j.actpsy.2010.11.003.

Inhoff, a W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics*, *40*(6), 431–439. doi: 10.3758/BF03208203.

Juhasz, B. J., & Rayner, K. (2006). The role of age of acquisition and word frequency in reading: Evidence from eye fixation durations. *Visual Cognition*, *13*(7-8), 846–863. doi:10.1080/13506280544000075.

Just, M., & Carpenter, P. (1980). A theory of reading: from eye fixations to comprehension. *Psychological Review, 87*(4), 329–354. doi:10.1037/0033-295X.87.4.329.

Kim, J., & Davis, C. (2004). Investigating the audio–visual speech detection advantage. *Speech Communication*, *44*(1-4), 19–30. doi:10.1016/j.specom.2004.09.008.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*(1-2), 262–284. doi:10.1080/09541440340000213.

Kuperman, V., & Van Dyke, J. a. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of Memory and Language*, *65*(1), 42–73. doi:10.1016/j.jml.2011.03.002.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30 thousand English words. *Behavior Research Methods*, *44*(4), 978–990. doi: 10.3758/s13428-012-0210-4.

Li, P., Sepanski, S., & Zhao, X. (2006). Language history questionnaire: A web-based interface for bilingual research. *Behavior Research Methods*, *38*(2), 202–10. doi:10.1017/S1366728913000606.

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.

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

Micic, D., Ehrlichman, H., & Chen, R. (2010). Why do we move our eyes while trying to remember? The relationship between non-visual gaze patterns and memory. *Brain and Cognition*, *74*(3), 210–224. doi:10.1016/j.bandc.2010.07.014.

Mishra, R. K., Olivers, C. N. L., & Huettig, F. (2013). Spoken language and the decision to move the eyes: To what extent are language-mediated eye movements automatic? In V. S. C. Pammi, & N. Srinivasan (Eds.), *Progress in Brain Research: Decision making: Neural and behavioural approaches* (pp. 135-149). New York: Elsevier.

Rayner, K. & Pollatsek, A. (2006). Eye Movement Control in Reading. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistics* (pp. 613-658). London: Elsevier.

Rayner, K., & Duffy, S. a. (1986). Lexical complexity and fixation times in reading: effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, *14*(3), 191–201. doi:10.3758/BF03197692.

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading*. Hillsdale, NJ: Erlbaum.

Road, S. P., Medical, B., & Limits, S. (1977). Reading while listening: A linear model of selective attention. *Journal of Verbal Learning and Verbal Behavior, 16*, 453–463.

Speed, L. J., & Vigliocco, G. (2013). Eye movements reveal the dynamic simulation of speed in language. *Cognitive Science*, *38*(2). 367-382. doi:10.1111/cogs.12096.

Spivey, M. J., Tyler, M. J., Eberhard, K. M., & Tanenhaus, M. K. (2001). Linguistically mediated visual search. *Psychological Science*, *12*(4), 282–286. doi:10.1111/1467-9280.00352.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*(5217), 1632–1634.

Tauroza, S., & Allison, D. (1990). Speech rates in British English. *Applied Linguistics*, *11*(1), 90–105. doi:10.1093/applin/11.1.90.

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio-visual speech perception is special. *Cognition*, *96*(1), B13–22. doi:10.1016/j.cognition.2004.10.004.

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, *57*, 209–232. doi:10.1016/j.specom.2013.09.008.

Yang, F., Chang, C., Chien, W., Chien, Y., & Tseng, Y. (2013). Tracking learners' visual attention during a multimedia presentation in a real classroom. *Computers & Education*, *62*, 208–220. doi:10.1016/j.compedu.2012.10.009.