Visual perceptual assimilation of non-native speech.

Inna Brodsky

ORCID 0000-0003-2809-2667

Submitted in partial fulfilment of the requirements of the degree of Master of Psychology/Doctor of Philosophy.
April, 2017.

School of Behavioural Science. The University of Melbourne.
ABSTRACT

Studies of auditory speech perception show that native language exerts a powerful influence on the perception of a second language. This influence is manifested in the assimilation of non-native sounds into native speech categories. However, speech is a multimodal experience, incorporating visual aspects of speech as well as sounds. The novel question addressed here is whether these aspects are assimilated naturally in the audio-visual (AV) perception of speech. The main hypothesis is that the visual aspects of non-native speech are assimilated into native categories, leading to the prediction that visual aspects of non-native speech that are dissimilar to the native language will be identified more readily than visual aspects of non-native speech that are similar to the native language. In several experiments, English-speaking participants were presented with native and non-native vowels and syllables that were experimentally validated and ranked according to their similarity to the native (English) language. Surprisingly, the results showed no evidence of visual perceptual assimilation for similar or dissimilar visual features of non-native speech. The study also investigated perception of visual aspects of non-native speech at the lexical level of cognate words and sentences compared to visual aspects of the native language. The results showed that participants were unable to discriminate native and non-native languages based on visual-only input. When native and non-native sentences were temporally reversed, participants were unable to discriminate veridical from temporally-reversed sentences in a visual-only mode condition. Furthermore, in cross-modal matching experiments, participants were unable to translate a visual image into the corresponding auditory signal, unless a written cue was provided or an auditory signal was present. Moreover, slowing the presentation of stimuli during AV speech perception did not facilitate processing of the visual aspects of non-native speech. Overall, the current study found a remarkably limited influence from the visual-only aspects of speech on non-native perception when compared to auditory-only input but did show for the first time that participants were more likely to falsely accept non-native stimuli as native in the multi-modal AV condition when compared to the auditory-only and visual-only conditions. The results therefore extend prior reports about the influence of visual aspects of speech on perception of sounds in a non-native language. These findings are discussed in terms of models of non-native speech perception and of applicability for teaching non-native languages to second language learners.
Visual perceptual assimilation of non-native speech

Declaration.

This is to certify that

(i) the thesis comprises only my original work towards the PhD,
(ii) due acknowledgement has been made in the text to all other material used,
(iii) the thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Signed: ____________________________________________
Date: __5/06/2018______________________________
Acknowledgements.

I am extremely grateful to Dr Meredith McKague for her enduring help, support, and guidance. Without her encouragement this thesis would not have been completed. I’m also very grateful to Associate Professor Brendan Weekees for taking on my supervision mid-way through and being instrumental in shaping this thesis. I also thank Professor Chris Davis for his expert advice during the early stages of the journey.

I acknowledge my overseas family, a circle of close friends, and work colleagues, too many to name, who believed in me despite adverse circumstance and encouraged me to finish this project. I am particularly grateful to my husband Mark for all his enduring support and help. I thank my daughter Eliana for providing me with a new impetus to finish the things I had started. I feel blessed thinking that you may apply your ever so critical thinking to this thesis one day.

Finally, I thank my son Daniel whose arrival into this world has had such a profound impact on all spheres of my life, including this thesis. Your disability has made me question everything I know about language, speech, and communication.

In loving memory of my polyglot grandparents- Dr Iliya and Dr Sima Elkin.
# Table of Contents

Chapter 1: Introduction & Literature Review ............................................................. 1
  1.1 Infant speech perception ....................................................................................... 1
    1.1.1 Infant auditory speech ................................................................................... 1
    1.1.2 Infant AV speech .......................................................................................... 3
    1.1.3 Infant AV Speech in a second language (L2) ................................................. 3
  1.2 Adult speech perception ....................................................................................... 5
    1.2.1 Adult auditory speech perception ................................................................. 5
    1.2.2 The Critical Period Hypothesis .................................................................... 6
    1.2.3 Phonotactics .................................................................................................. 9
    1.2.4 Lip-Reading in Adults .................................................................................. 10
    1.2.5 Speech-reading in L2 ................................................................................... 12
    1.2.6 Adult AV speech .......................................................................................... 15
    1.2.7 Cross-language McGurk effect ....................................................................... 17
  1.3 AV benefit effect .................................................................................................. 20
    1.3.1 Language-specific AV benefit ....................................................................... 23
    1.3.2 Speech synchrony .......................................................................................... 27
    1.3.3 Rate-dependent speech perception ............................................................... 31
    1.3.4 Cross-modal repetition priming ..................................................................... 37
    1.3.5 Orthographic primes ..................................................................................... 40
    1.3.6 Cognate effects ............................................................................................. 41
  1.4 Theories of Speech Perception ............................................................................ 43
    1.4.1 Native Language Magnet (NLM) Model ......................................................... 44
    1.4.2 The Perceptual Assimilation Model (PAM) ..................................................... 49
    1.4.3 The Speech Learning Model ......................................................................... 52
    1.4.4 Motor Theory ............................................................................................... 53
    1.4.5 The Fuzzy Logical Model of Perception (FLMP) ............................................ 55
    1.4.6 Analysis-by-synthesis model ......................................................................... 57
    1.4.7 Visual Place Articulatory Manner (VPAM) Theory ........................................ 58
    1.4.8 The Reverse Hierarchy Theory (RHT) ............................................................ 60
  1.5 Summary of research questions ............................................................................ 62

Chapter 2: Background to language, articulation, and Russian Language ............... 64
  2.1 Levels of language processing ............................................................................. 64
    2.1.1 Phonological level ......................................................................................... 64
    2.1.2 Morphological Level .................................................................................... 64
Chapter 4: Perception of the visual aspects of vowels and syllables in non-native speech

2.1.3 Syntactic level ................................................................. 64
2.1.4 Semantic Level ............................................................... 65
2.1.5 Pragmatic Level ............................................................... 65
2.2 Parameters of Sounds .......................................................... 65
2.3 Speech Organs ................................................................. 66
2.4 Russian Language ............................................................... 67
  2.4.1 Background ................................................................. 67
  2.4.2 Phonetic System of the Russian Language ......................... 67
  2.4.3 Comparative studies of English and Russian languages ....... 71
Chapter 3: Rationale for the current study ........................................... 74
  3.1 Visual perceptual assimilation ............................................ 74
    3.1.1 Native vs Non-native effect ....................................... 74
    3.1.2 AV benefit effect ..................................................... 74
    3.1.3 Mental representation of non-native speech ..................... 75
    3.1.4 Level of Input ....................................................... 75
    3.1.5 Temporal aspects of L2 processing ................................. 76
  3.2 Mapping of experiments to hypotheses .................................. 76
    3.2.1 Visual perceptual assimilation .................................. 76
    3.2.2 Native vs Non-native effect ..................................... 76
    3.2.3 AV benefit effect ..................................................... 77
    3.2.4 Level of Input ....................................................... 77
    3.2.5 Temporal aspects of L2 processing ................................. 77
Chapter 4: Perception of the visual aspects of vowels and syllables in non-native speech.................................................. 78
  4.1 Experiment 1: Perception of Vowels and Syllables ..................... 79
    4.1.1 Method ................................................................. 79
    4.1.2 Results ............................................................... 82
    4.1.3 Discussion ............................................................ 86
  4.2 Experiment 2: Validation study ........................................... 89
Experiment 2 Hypotheses: ......................................................... 89
  4.2.1 Method ................................................................. 92
  4.2.2 Results ............................................................... 93
  4.3 Experiment 3: Same-Different Paradigm ................................ 97
    4.3.1 Method ................................................................. 98
    4.3.2 Results ............................................................... 99
    4.3.3 Discussion ............................................................ 101
  4.4 Experiment 4: Item Rating ................................................. 103
### Table of Contents

**Visual perceptual assimilation of non-native speech**

4.4.1 Method ........................................................................................................ 103
4.4.2 Results ......................................................................................................... 104
4.4.3 Discussion ................................................................................................. 105

4.5 Experiment 5: Match-Mismatch ..................................................................... 106
4.5.1 Method ........................................................................................................ 107
4.5.2 Results ......................................................................................................... 108
4.5.3 Discussion ................................................................................................. 111

4.6 Summary of Chapter 4 .................................................................................. 113

Chapter 5: Perception of words and sentences .................................................. 115

5.1 Experiment 6: Language Identification of Words and Sentences ................. 116
5.1.1 Method ........................................................................................................ 117
5.1.2 Results ......................................................................................................... 119
5.1.3 Discussion ................................................................................................. 122

5.2 Experiment 7: Matching ................................................................................. 125
5.2.1 Method ........................................................................................................ 126
5.2.2 Results ......................................................................................................... 127
5.2.3 Discussion ................................................................................................. 134

5.3 Experiment 8: Sentence identification .......................................................... 135
5.3.1 Method ........................................................................................................ 136
5.3.2 Results ......................................................................................................... 137
5.3.3 Discussion ................................................................................................. 140

Chapter 6: General Discussion ............................................................................ 142

6.1 Summary of Experiments ............................................................................. 143
6.2 Visual only speech ......................................................................................... 147
6.3 AV speech ....................................................................................................... 161
6.4 Implications of the study .............................................................................. 164
6.5 Future direction ............................................................................................. 166
6.6 Conclusion ...................................................................................................... 171

References ............................................................................................................ 172

Appendix A: Instructions ..................................................................................... 189
Appendix B: Stimuli with IPA transcription ....................................................... 191
Figures

Figure 1. Interaction Between Modality and Similarity in the Identification Task (English versus Non-English). ................................................................. 95
Tables

Table 1. Mean Percentage Error Rates for Experiment 1 as a Function of Language, Modality, and Similarity for Audio-Visual (AV), Audio-Only (AO), and Video Conditions

Table 2. Mean Response Time for Experiment 1 as a Function of Language, Modality, and Similarity for Audio-Visual (AV), Audio-Only (AO), and Video Conditions

Table 3. The Best and the Worst Mean Percent Error for the Russian Items in the Audio Only, Video Only, and Audio Visual (AV) Conditions

Table 4. The Best and the Worst Mean Percent Error for the English Items in the Audio Only, Video Only, and Audio-Visual (AV) Conditions

Table 5. The Best and the Worst Mean Percent Error in the Audio Only, Video Only, and Audio Visual (AV) Conditions

Table 6. Mean Percentage Error Rates for Each Condition: Audio Only (AO) and Video

Table 7. Mean Response Time in ms for Each Condition: Audio Only (AO) and Video

Table 8. Mean Percentage Error Rates for Same-Different Experiment

Table 9. Mean Response Times in ms in Same-Different Experiment for Audio Only (AO) and Audio-Visual (AV) Conditions

Table 10. Average Ranking with standard deviation in parenthesis for 3 types of stimuli in the AO and AV conditions

Table 11. Mean Percentage Error Rates in Match-Mismatch Experiment

Table 12. Response Rates for Match-Mismatch Experiment

Table 13. Mean Percentage of Error Rates for Visual Only (VO) With and Without Written Primes Conditions in Experiment 6

Table 14. Mean Percentage of Error Rates for Audio Visual (AV) Conditions in Experiment 6

Table 15. Mean Response Time for Visual Only With and Without Written Primes Conditions in Experiment 6

Table 16. Mean Response Times for Audio-Visual (AV) Conditions in Experiment 6

Table 17. Error Rates for Sound-Video Condition in Experiment 7

Table 18. Error Rates for Video-Sound Condition in Experiment 7

Table 19. Response Times for Sound-Video Condition in Experiment 7

Table 20. Response Times for Video-Sound Condition in Experiment 7

Table 21. Mean Error Rates for Video and Audio-Visual (AV) Conditions in Experiment 8

Table 22. Mean Response Times for Video and Audio-Visual (AV) Conditions in Experiment 8
Chapter 1: Introduction & Literature Review

The goal of the experiments reported in this thesis is to test the hypothesis that integration of visual and audio-visual (AV) aspects of non-native speech is constrained by linguistic similarity, i.e. is language dependent. Research into development of AV speech is nascent and based on foundations in the auditory speech domain only. Within this domain, studies are limited to infant speech development in the non-native language. Although developmental findings provide a foundation for the study of adult non-native speech perception in the audio-visual (AV) modality, there are significant gaps in our understanding. The aim here is to address these research gaps with an exhaustive study of the influence of visual aspects of speech on perception of non-native sounds in adult speakers. The non-native language used in the study is Russian, which has not been used so far in any similar study. This language is relatively unknown in Australia as it is not one of the top 5 frequently spoken languages in the country. In order to motivate specific predictions for the current study it is necessary to understand non-native speech perception in developing speakers. Therefore, the thesis begins with a brief overview of this literature.

1.1 Infant speech perception

1.1.1 Infant auditory speech

Our understanding of adult speech perception is based on the foundation created by research into infant speech development. Research in infant speech perception spans over 50 years and there have been evolutions in our understanding of speech development during this time (see Soto-Faraco, Calabresi, Navarra, Werker, & Lewkowicz, 2012). We know that the linguistic environment shapes speech perception continuously and rapidly from birth (possibly prenatally) and it is also widely assumed that the development of speech perception follows an invariant trajectory across cultures (Cutler, 2012).

Early accounts of speech perception assumed that infants enter their linguistic world with a hardwired and automatic speech perception device that is independent of culture i.e. language-independent view. and differentiate between self-directed and non-self-directed speech (Cooper & Aslin, 1990; Pegg, Werker, & McLeod, 1992). Furthermore, infants are capable of discriminating speech contrasts, such as nasal contrasts, the place of articulation in stops (Eimas & Miller, 1980; Moffitt, 1971); voicing in stop consonant pairs (Archer, Zamuner, Engel, Fais, & Curtin, 2016), stop/glide distinction (Oller, Eilers, Burns, & Urbano, 1993) and
oral/nasal distinctions (Eimas & Miller, 1980). Such findings lent support to this universal view of speech development that has dominated the field for more than half a century.

However, the universal view has been challenged by reports of a differential time course for auditory speech perception of language specific features during infancy. For example, Lasky and colleagues (1975) found that by the age of 6-months, infants exposed to an English-Spanish speaking environment could discriminate contrasts in labial stops in English but not Spanish. Similarly, Werker and colleagues (1983) report that perception of non-native contrasts is time limited from the age of 6 months, diminishing to zero by 12 months. Indeed, Polka and Werker (1994) found that 4-month-old infants could discriminate non-native German vowels better than 6-month olds, with a complete decline in perception by 12 months. A similar pattern of atrophy has been reported for differentiation of non-speech contrasts such as Zulu clicks within 1 year (Best, McRoberts, & Sithole, 1988).

Comparative studies of auditory speech perception converge on the view that the dominant linguistic environment shapes infant perception of speech. Studies report dominant language characteristics shape the auditory perception of speech segments (consonants and vowels) and supra-segmental features of language (Hirsh-Pasek et al., 1987). For example, Jusczyk and colleagues (1986) reported that infants at birth are sensitive to all prosodic markers of clausal units in a narrative but focus on the dominant prosodic structures in their linguistic environment by 6 months. Several studies (Hirsh-Pasek et al., 1987; Jusczyk et al., 1992) reported that infants become increasingly attuned to fine-grained segmental properties of words between 6 and 9 months of age. For instance, while 6-month old American infants showed no preference for listening to Dutch or English words, by 9-months they demonstrate a preference for English words. The opposite pattern was reported in Dutch infants (Jusczyk & Cutler, 1993). Jusczyk and colleagues (1994) found that 9 month-old infants are more sensitive to the distributional properties of phonotactic ordering in their dominant language (combinations of consonant-vowel-consonants (CVC) sequences with high or low probability of occurrence in English) compared to 6 month old infants. Similar developmental shifts are observed in the sensitivity to variation in stress patterns for the dominant language (Jusczyk & Cutler, 1993; Jusczyk & Luce, 1994). Mattock and Burnham (2006) investigated changes in lexical tone sensitivity with infants between 6 and 9 months of age. They found that infants exposed to Chinese (a tonal language) perform equally well at 6 and 9 months on speech and non-speech (music) tone discrimination tasks. In contrast, discrimination of unfamiliar lexical tones declined for infants raised in an English language dominant environment, whereas discrimination for musical tones remained constant.
In sum, the evidence that infant speech perception is fine-tuned by the dominant features of the language environment during the first year of life is mostly uncontested. However, given that the visual aspects of language also influence speech perception, one outstanding question is whether the development of audio-visual (AV) speech perception follows a similar pattern. A review of AV speech perception studies in infants therefore follows.

1.1.2 Infant AV speech

A large body of research shows that infants use the visual aspects of speech and integrate AV attributes in speech perception very early in their language development. For instance, young infants dislike a mismatch between the face and voice of mother (Aronson & Rosenbloom, 1971). Bahrick (Bahrick, 1992) found that infants could detect temporal synchronies when objects are presented in an AV mode and can detect whether they are composed of large or small elements. Infants fixate longer when lip movements and voice are matched than when they are mismatched (Burnham & Dodd, 2004). Lewkowicz and Lickliter (1994) conducted experiments on discrimination of changes in multimodal displays, under audio-only (AO), video-only (VO), and audio-visual (AV) modalities. The results were that infants could discriminate AV changes earlier than they could discriminate changes to audio or visual information alone. Burnham and Dodd reported that by 1-month of age infants can match the face and voice of their mother and by about 3 months can recognise their mother in the visual modality alone. Furthermore, newborns are able to match lip movements to auditory input (Aldridge, Braga, Walton, & Bower, 1999) and to match vowels with lip movements by 2 months of age (Patterson & Werker, 2003). Furthermore, infants can match vowels in non-speech at around 4 months (Walton & Bower, 1993) and match emotional sounds with familiar voices by about 7 months of age (Walker-Andrews & Lennon, 1991).

1.1.3 Infant AV Speech in a second language (L2)

Weikum and colleagues (2007) showed silent video clips of 3 English-French bilingual women reciting sentences in either English (native) or French (non-native) to 4 and 6-month-old infants. She found that although both native and non-native listeners could discriminate the visual aspects of speech before 8-months of age, this ability diminished with age. However, non-native listeners showed less atrophy of visible speech perception. Weikum and colleagues argued that there is a perceptual narrowing in the processing of visual aspects of speech for native speakers but not necessarily for infants who are exposed to non-native languages. In a follow-up study, Sebastian-Galles and colleagues (2012) tested an alternative
hypothesis that non-native listening leads to better perceptual (including visual) attention as a result of bilingual experience. They predicted that bilingual infants would show better discrimination of visual aspects of speech in native languages and non-native languages. They presented bilingual Spanish-Catalan and monolingual Spanish or Catalan infants with the same silent videos used by Weikum and colleagues. Results showed that bilingual infants noticed differences between visible language cues in non-native English and non-native French but monolingual infants did not. The researchers interpreted the results as evidence for an “advantage” in the visual attention system in bilinguals compared to monolingual infants that delays natural perceptual narrowing of visible speech perception.

Pons and colleagues (2009) later adapted the paradigm of Weikum et al. (2007) to the perception of audio-visual (AV) cues. They compared the identification of visually presented audible phonemes between 6- and 11-month-old native Spanish and English monolingual infants with side-by-side silent videos of different language syllables and measured looking preferences. Infants were then familiarised with one syllable auditorily and retested for looking preference. One critical contrast /ba/-/va/ was included as it occurs in English but not in Spanish. They compared performance with adult monolingual English and Spanish speakers. Results showed that the ability of Spanish-learning infants to match phonemes presented in the visual auditory and auditory modalities had diminished by 11 months of age, falling to the equivalent level of Spanish adults. In contrast, English infants (and adults) retained the ability to identify the auditory and visual aspects of /ba/ and /va/, presumably because they are distinguished as separate contrasts in their native visible speech. There is therefore evidence for selective perceptual tuning of native and non-native audio-visual aspects of speech in the adult literature. It is not clear however if non-native audio-visual speech contrasts can be perceived by adults in the visual domain only when applied to a given language. This answer to this question would be a strong test of the domain specific visible identification of non-native speech. Such an experiment has not yet been conducted.

There is evidence of a developmental trajectory in the integration of audio and visible aspects of speech. McGurk and MacDonald (1976) found that concurrent presentation of a syllable /ba/ with conflicting visual information /ga/ leads to the false perception of the syllable as a fused sound i.e. /da/. This is the classic McGurk effect. They also reported that the influence of visible aspects on fused speech perception increased between 3 and 8 years. Burnham and colleagues (2004) using a habituation paradigm reported that the McGurk effect was present by the age of 4 months and possibly as early as 2-months (L. Rosenblum, Schmuckler, & Johnson, 1997). More recent studies show that by 4-5 months infants are capable of matching
Visual perceptual assimilation of non-native speech

They are able to use temporal synchrony cues to facilitate perception by 6 months of age (Kopp, 2014).

In sum, the development of AV perception follows the trajectory similar to auditory speech perception. Although the potential for a language environment to shape the auditory perceptual system is unlimited at birth, there is a shift in preference to familiar phonetic contrasts in the first year of life and sensitivity to non-familiar sounds diminishes thereafter. Visible aspects of speech seem to supplement the pattern set by the auditory domain. However, few studies have considered the possibility of differential effects from native and non-native contrasts on infant and adult AV speech perception. That question motivated the experimental work reported here; that is, do AV cues constrain the perception of non-native speech in naïve adult participants? The literature relevant to this question will be reviewed below.

1.2 Adult speech perception

1.2.1 Adult auditory speech perception

Evidence from studies of auditory speech perception highlights challenges in the acquisition of L2 phonetic features for naïve adult participants. One example is the well-documented difficulty that native Japanese speakers have producing and perceiving the English distinction [r]-[l], presumably because this is absent in the Japanese language. Furthermore, Sheldon and Strange (1982) report that Japanese speakers have difficulties perceiving a difference between [r] and [l] even when they can correctly produce it English. Several studies (Aoyama, Flege, Guion, Akahane-Yamada, & Yamada, 2004; Lively, Pisoni, Yamada, Tohkura, & Yamada, 1994; Underbakke, Polka, Gottfried, & Strange, 1988) report that Japanese listeners weigh acoustic parameters differently than native English speakers. Their studies showed that Japanese listeners perceive the temporal distinction between [r] and [l] even if they have difficulty differentiating spectral differences in F3 and F2 (spectral peaks of the sound spectrum) and utilise F2 cues more often, identifying stimuli as neither [r] or [l] but as [w].

Constraints on non-native speech perception are also found in other languages. For instance, Beddor and Strange (1982) reported that experience-related differences can be detected in the perception of oral-nasal contrasts by native Hindi and English speakers. Gottfried and Beddor (1984) also reported that English speaking adults had difficulties discriminating French vowels during the initial stages of learning French. Furthermore, Tees and Werker (1984)
reported that English-speaking subjects fail to discriminate a velar-uvular place contrast in voiceless ejective stops [k-q] of the Hindi language.

Some studies suggest that cross-language constraints on non-native speech perception arise at the rhythmic level. Specifically, results from cross-linguistic experimental studies (Altmann & Young, 1993; Dupoux & Green, 1997; Mehler et al., 1993a; Pallier, Sebastian-Gallés, Dupoux, & Mehler, 1998; Sebastián-Gallés, Dupoux, Costa, & Mehler, 2000) show that the intelligibility of sentences in time-compressed speech when a language belongs to the same rhythmic group as the native language is better than when a language comes from a different rhythmic group. Cross-language disadvantages in speech perception were also noted with common stress patterns, wherein listeners who use a fixed-stress language were unable to perceive contrastive stress in the non-native language (Dupoux & Green, 1997; Dupoux, Peperkamp, & Sebastián-Gallés, 2001).

One factor to emerge from the field of auditory speech perception and production in L2 is the effect of age of acquisition of L2. Therefore we now turn to a brief review of this issue in the context of the Critical Period Hypothesis.

1.2.2 The Critical Period Hypothesis

The Critical Period Hypothesis (CPH) was proposed by Eric H. Lenneberg (1967) in his seminal work *Biological Foundations of Language*.

He argued that humans require raw material (speech) to trigger language development. He argued that this linguistic input during a critical period for language acquisition is responsible for normal language development. His hypothesis was based on the cases of children who were deprived of any linguistic input from birth to post-puberty. The most famous case is that of “Genie” who was deprived of linguistic input and punished by any attempted verbal communication. In contrast, a young boy raised in isolation, was able to make rapid progress. Lenneberg argued that beyond a critical point at the onset of puberty, normal acquisition of language is compromised. He also used the cases of aphasia to support his theory, arguing that children suffering from aphasia are more likely to recover than adults. Furthermore, CPH argues that accents commonly observed in L2 adult learners could be accounted for by physiological factors limiting L2 attainment. Scovel (Scovel, 1988) elaborated on the CPH by proposing that the first 15 years of life are not critical but sensitive for language development, especially for phonology. He proposed that beyond puberty, the acquisition
of native-like language is attainable. In contrast, the acquisition of non-accented native speech is difficult due to biological factors, such as brain plasticity.

The CPH has been generating debate for many decades, especially in defining the critical period boundaries. Some argue that the critical period ends around the middle to late teenage years (Birdsong, 2006) whilst Pinker (1994, cited in Slabakova (2013) defines the period of maximum sensitivity as beginning at birth and ending as early as at the age of 6. Furthermore, just being exposed to a language is not sufficient, according to Johnson and Newport (1989) who argued with their “exercise theory” that exercising language abilities during the critical period ensures that they remain intact during the life span.

There is conflicting evidence for CPH in relation to L2 acquisition. Some researchers argue that is near impossible for L2 speakers to attain the same level of proficiency as native speakers (e.g. (Dekeyser, Alfi-Shabtay, & Ravid, 2010), (Doughty & Long, 2003),(Long, 2005). Indeed, the age of L2 acquisition emerges as one of the biggest predictor of L2 attainment. For instance, the study of DeKeyser and colleagues (2010) tested 75 Russian learners of English and 64 Russian learners of Hebrew. They used a grammatical judgment task to test their L2 attainment. A verbal aptitude test in participants’ native Russian language was used to predict their performance. The results showed that the aptitude test’s results only predicted performance in those who acquired L2 between the age of 18 to 40 but not in the early age acquisition group. However, no native controls performing the same tests were included in the study. Another large-scale study conducted in Sweden (Abrahamsson & Hyltenstam, 2009) tested a sample of 195 native Spanish speakers with advanced proficiency in Swedish. The results revealed that only 5% of late bilinguals who acquired Swedish after adolescence were perceived as native speakers. In contrast, 62% of bilinguals who acquired Swedish before the age of 12 were perceived by native judges to be at the level of native speakers. The authors interpreted their results as supporting the CPH which argues that native-like language attainment is not possible after the critical period.

On the other side, many studies argue that the differences between native and L2 speakers which are taken to support the CPH are quantitative but the language acquisition mechanism is the same (Singleton & Ryan, 2004; Slabakova, 2013). For instance, when 70 post-puberty L2 learners and 67 adult heritage speakers were compared in different areas of Spanish morphology and syntax, the results revealed similar errors for the two groups, as well as some advantages for the early child learners (Montrul & Slabakova, 2003). The study argued that continuing high qualitative linguistic input is just as important as the age of L2
acquisition and that native-like performance is possible for some aspects of grammar. Montrul and Slabakova showed that as many as 30% of L2 speakers performed within the range of native speakers on some aspects of phrasal semantics. They argued that it is important to identify the areas of native like acquisition as well as the areas of failure in our understanding of L2 process.

Another issue to discuss in relation to the CPH is what is known as the Bilingual Turn issue (Ortega, 2009). Similarly to other researchers, such as Cook (2008) and Singleton (2004), Ortega says that it is inappropriate to compare bilinguals with monolinguals as control, especially within the CPH field. As stated by Slabakova (Slabakova, 2013) “the input of monolinguals and bilinguals is too varied for direct comparisons to be justified”.

In addition, the direct comparison between monolinguals and bilinguals is undermined by several studies which show that bilingual experience changes neural pathways and whilst one language is activated the other one is inhibited (Bialystok, 2009). Bialystok argues that this creates an additional burden of attentional control which is unique to bilinguals which may lead to underestimating their abilities.

Whilst the argument that bilinguals should not be compared to native speakers continues, some are proposing that they should be compared to early bilinguals (Sorace, 2011). In contrast, some argue that late, very proficient bilinguals should be used as controls in addition to native-speaker controls (Slabakova, 2013).

In the field of auditory speech perception specifically, studies of bilingual children who acquired their second language in their first years of life (early) show differential patterns of phonetic sensitivity when compared to children and adults who acquire a second language later. For example, studies comparing the perception of native language (L1) and contrasts in L2 find a perceptual advantage for L1 contrasts that increases with age even for speakers who are balanced in exposure to both languages (Bosch, Costa, & Sebastian-Galles, 2000; Pallier, Bosch, & Sebastian-Gallés, 1997; Pallier, Colomé, & Sebastián-Gallés, 2001; Sebastian-Galles & Soto-Faraco, 1999). Furthermore, Williams (1974) showed that Spanish-speaking children (ages 8-10 years and 14-16 years) learning English in the US exhibited a shift in their perceptual boundaries for Spanish between voiced and voiceless labial stops that depended on age of arrival in the US. Streeter and Landauer (1976) reported similar results with Kikuyu children in Africa who were learning English labial stop contrasts. Similarly, Tees and Werker (1984) found that English speakers could not discriminate Hindi phonemes based on place of articulation after 1 year of training (requiring 5 years of extensive training).
In sum, the age of L2 acquisition is an important factor to take into consideration when interpreting research and conducting studies in the field of L2 acquisition. In order to control for this important factor, the current study attempted to use a speaker who is a balanced bilingual. Specifically, the speaker who recorded stimuli had acquired his L2 (English) before puberty. In addition, the aforementioned review highlights the issue of making generalisations when interpreting results comparing bilinguals who acquired their L2 in different age groups, as well as research which compared bilinguals with monolingual controls without acknowledging biases inherited in this comparison.

In addition to the age of L2 acquisition, phonotactics also constrain speech acquisition in L2. Therefore we now briefly discuss this topic.

1.2.3 Phonotactics

Phonotactics refers to the restrictions in sounds within a language that are measured by the permissible combinations of phonemes. Polivanov (1931) first noticed that native Japanese speakers mispronounced common Russian words and hypothesised that this was due to misperception of non-native phonemes rather than their inability to pronounce non-native phonemes properly. He reasoned that this occurs when the position of the final consonant in a Russian word does not occur (or is not legal) in Japanese. Many experiments have demonstrated effects of phonotactic filtering (Lentz & Kager, 2015; Solé, 1989). In a study by Massaro and Cohen (Massaro, Cohen, & Gesi, 1993), participants perceived ambiguous sounds as phonotactically legal in the native language. Later studies showed that phonotactic influences are language-specific (Hallé, Segui, Frauenfelder, & Meunier, 1998). For example, Dupoux and colleagues (1998) constructed Japanese phoneme combinations that are not found in French and French phonemes that are absent in Japanese. They reported a double dissociation for phonetic clusters that do not exist in one language but do in the other. Such studies show how language-specific phonotactics impact non-native speech perception.

Do naïve participants also perceive some visual aspects of a non-native language as ‘illegal’ or unusual? Studies of visual speech (speech-reading or silent lip-reading) perception address this question.
1.2.4 Lip-Reading in Adults

Research on silent speech reading is well established but relatively new to the field of speech processing. Traditionally, research on speech-reading has been dominated by studies focusing on visemes. These are the basic units of visual speech perception assumed by some researchers to be equivalent to phonemes in auditory speech (Files, Tjan, Jiang, & Bernstein, 2015). The study of visemes dates back to studies of lip-reading in hearing-impairment (Erber, 1982). A viseme may be associated with several phonemes and several phonemes may correspond to a single viseme, e.g. /k, g, η/ (viseme: /k/). Walden and colleagues (Walden, Prosek, Montgomery, Scherr, & Jones, 1977) showed that hearing participants can distinguish categories of consonantal visemes during lip-reading and discrimination improves with training. Montgomery and Jackson (1983) report that hearing participants can lip-read vowels presented in CVC form and can identify viseme categories when lip-reading. Thus, visible categories in speech perception are not contested.

Lachs and colleagues (2000) showed that the errors produced by speech-readers are organised according to the perceptual confusability of individual visemes. Using data from the Hoosier Audiovisual Multimodal Database, viseme groups were derived from sets compiled by Bernstein and Auer (1996). Results showed perceptual ‘distances’ between speech segments, whereby segments that are confused are perceptually ‘closer’ than those that are not. Lachs assumed that perceptual space for visemes could be represented on a continuum of confusability, so that high values represent segment spaces that are highly different from each other while low values represent spaces where segments are similar to each other. Lachs concluded that visemes do not represent natural categories but more likely an abstraction derived from language experience.

The notion of a viseme as a linguistic class has been questioned. Files and colleagues (2015) asked participants to distinguish pairs of consonant-vowel spoken nonsense syllables that were designed to differ in perceptual distance. Stimuli were presented as natural (non-modified visual format) or synthesized video clips. Natural stimuli were discriminated better than synthetic speech and natural within-viseme stimulus pairs were discriminated significantly above chance (except for /k/-/h/). The researchers argued that visemes have internal perceptual structure and there is more phonetic information in visual speech then would be predicted if visemes were modular perceptual categories. Importantly, they reasoned that viseme categories typically used in research studies can lead to misleading inferences because phonemes in visemes can be discriminated and perceivers are sensitive to sub-visemic phonemic features. Therefore, the question of viseme categories is controversial.
Given individual variability in the perception of visible speech, Massaro, Cohen and Gesi (1993) looked at the effects of long-term training in speech reading. A learning trial consisted of a feature, for example, closed lips at onset occurring in a test item followed by feedback (e.g. syllable /ba/). Participants were taught to speech-read 22 initial consonants in three different vowel contexts, using discrimination and identification lessons. They were tested throughout their training on recognition of syllables, words, and sentences. Input was presented visually, auditorily, and bimodally at a normal rate or at three times normal rate. The results revealed that subjects improved in their speech reading ability across all three item types. Some researchers argue further that a focus on visemes classification does not reflect the complexity of speech. For instance, Bernstein and Liebental (2014) argue that any feature of speech can be perceived visually. They reviewed the empirical evidence on lip reading at the level of visemes, features, phonemes, spoken words, and prosody ( multisyllabic words and sentences ) and identified differences in lip reading abilities. Individual differences were observed for people with normal hearing and hearing impairment. For instance, when deaf lip readers, who rely on visual speech more than normal-hearing individuals are asked to lip read words in sentences, performance varies from zero to over 85% correct. Variability is also observed with phoneme identification in non-sense syllables. At the group level, people who have congenital hearing loss are better than normal-hearing individuals at lip reading but some hearing participants are capable of good lip reading.

Another methodological issue to emerge from studies of speech reading is speaker variability in terms of intelligibility. This is true for auditory and visual speech. Sheffert and colleagues (1998) investigated intelligibility of word tokens from the Hoosier Audiovisual Multimodal Database consisting of ten talkers uttering 300 isolated words. Half of the words were “easy” (high frequency), while the other half were “hard” (low frequency). Results showed that audio-visual intelligibility of words was near ceiling (above 90% accuracy) and easy words were identified more often than hard words. There were no differences across talkers in the audio-visual condition. However, differences emerged in the audio only (AO) and video only (VO) conditions. Compared to the audio-visual condition, the intelligibility of words in the VO modality was poor, with 90% of words identified with less than 40% accuracy. Interestingly, there was still a main effect of lexical category, with “easy” words identified more often than “hard” words.

Speaker variability issues are further highlighted in the studies which report that perceivers’ ability to extract visual information from speech varies across speakers (Bond & Moore, 1994; Ferguson, 2004; Gagné, Masterson, Munhall, Bilida, & Querengesser, 1994; Yehia, Rubin, & Vatikiotis-Bateson, 1998). However, adaptation effects are evident wherein prior
exposure to a speaker facilitates subsequent recognition of speech from that same speaker (Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994). Other studies also demonstrated that auditory and visual speech were recognised better when there is no change of speaker from trial to trial compared to when speakers vary across trials. More specifically, Rosenblum and colleagues (2008; 2007) showed that exposure to one speaker’s visual speech (sentences) speeds up subsequent auditory recognition of the same sentences. However, Van Zande and colleagues (2014) found no facilitation from hearing same speakers to the subsequent recognition of visual speech at the level of isolated words. In their studies, repeated words were identified better than new words, regardless of speaker identity. They concluded that speaker-specific characteristics may not transfer across modalities at the level of word perception. Rosenblum and colleagues found that when participants were familiarised with a speaker who was reading sentences silently, they were better at transcribing sentences presented in the auditory mode. The researchers interpreted the results as evidence that participants benefit from speaker-specific properties extracted from the lip reading condition and are able to utilise them in perception of auditory speech.

The aforementioned studies highlight both positive and potential limitations of using a single speaker but the limitations do not preclude inferences about potential underlying mechanisms involved in non-native visual speech perception. We will now narrow our review of the literature to the field of L2 visual speech processing which has a relative paucity of studies looking into the visual aspects of L2 in comparison to studies of L1 visual speech.

1.2.5 Speech-reading in L2

The ability to discriminate spoken languages is a fundamental capacity of speech perception. Does the ability to discriminate languages also occur for lip-reading? Soto-Faraco and colleagues (2007) asked whether a visual cue alone allows discrimination between languages. They reasoned that language discrimination should be difficult if languages have a similar visemic repertoire, rhythm, and overlapping lexicon, for example, comparing Spanish and Catalan, two Romance languages that are similar in phonology, rhythm, and lexicon. Soto-Faraco and colleagues used natural speech to include non-visual segmental, supra-segmental and lexical information that could be derived from a visual signal. A single female, who was judged to be a balanced Spanish-Catalan simultaneous bilingual, recorded all stimuli, which comprised of Spanish and Catalan sentences extracted from poetry texts. The participants included Spanish-Catalan bilinguals, who were divided into Catalan-dominant bilinguals, Spanish-dominant bilinguals, and simultaneous bilinguals. They were presented with a video clip of a speaker silently pronouncing a sentence either in Catalan or in Spanish. After a one-
second inter-stimulus interval, a second video clip of a sentence was played and participants had to judge whether the two sentences were the same or not. Results showed that all bilingual groups were able to discriminate Spanish from Catalan sentences and no differences emerged between the three groups. Discrimination was easier when sentences were longer. In a follow-up experiment, Soto-Faraco and colleagues (2007) presented the Spanish-Catalan sentences to native speakers of English and Italian. The Italian language but not the English language has syllable-based stress and a similar rhythm to Spanish and Catalan. Despite the similarities between Romance languages, when sentences were presented to native Italian speakers, they could not distinguish Spanish from Catalan and were no more sensitive to this discrimination than native English-speakers. The researchers also reported that monolingual Spanish and Catalan speakers could do the visual discrimination task, albeit not as accurately and quickly as bilingual speakers. Bilingual participants were then asked to report verbally the language of the target sentence and recall words from the sentence. Participants could report the language of a target sentence significantly better than chance and performance improved with longer sentences. However, very few (3%) of words could be recalled from the sentences, showing that speech reading is a challenging task. If words were distinctive exemplars of one language (e.g., the in English) they would have been more easily identified. The researchers argued that the capacity to identify spoken languages visually is constrained by linguistic experience and familiarity with at least one of the languages is necessary.

Can these results be generalized to other languages? Arguably, Catalan and Spanish are two linguistically close languages and some monolingual speakers would have been exposed to the other language over their lifespan. Furthermore, there are methodological issues inherent in the Soto-Faraco and el (2007) study, including presentation of passages from poetry which might have added to the rhythmic aspect of speech as well as increasing stimulus familiarity, thus potentially facilitating lip-reading.

The issue of language familiarity as a confounding variable was also highlighted in the study by Ronquest and colleagues (2010). They asked whether bilingual speakers could identify English and Spanish from visual speech only. They replicated the data from Soto-Faraco et al. (2007), with English and Spanish sentences spoken by a female and a male bilingual speaker. However, they further asked participants to decide whether the target word was spoken in English or Spanish. They found that both monolingual and bilingual speakers of English and/or Spanish could identify sentences as English or Spanish from visual input. However, unlike Soto-Faraco et al., they found no significant differences between monolingual and bilingual speakers. For monolingual English speakers, they assessed how much lexical information was extracted from visual-only presentation by presenting English and Spanish
sentences in the visual-only mode. Results showed that participants were able to extract about a quarter of the content from the English stimuli but none from Spanish (their unfamiliar language). They reasoned that lexical information in the visual-only mode is scarce even if perception of rhythmic gestures in visual speech is available. Ronquest and colleagues (2010) also presented participants with sentences and single words played backward retaining the rhythmic but not lexical properties of each language. The results showed that participants could identify words and sentences in a target language presented in the visual modality from rhythmic information only. One methodological criticism of the Ronquest et al study is identification of a speaker’s language identity in their experimental design, namely directly asking participants whether stimuli are presented in English or Spanish. It is therefore likely that American English-speaking monolinguals would have been primed in their performance to tune in for certain features specific to Spanish, a language to which they are likely to be exposed to in America given that Spanish is the second most commonly spoken language in the United States (Lipski, 2008). It is an open question whether similar results would be obtained for languages that are unfamiliar and the current study addresses this question by using Russian.

Replication of results with different languages and participant populations was highlighted in a study of Weikum and colleagues (2013) who examined French as L2. They tested a sample of 120 adult English-speaking monolinguals and French bilinguals to investigate the effect of age of acquisition on perception of visual aspects of speech. Groups were further divided into sets of 60 participants who had acquired English as L1 before the age of 2 (40 English only speakers and 20 infant multilinguals). The second group comprised of 60 participants who learned English as L2 after the age of 2 (30 early multilinguals, 30 late multilinguals). The bilingual sample was heterogeneous, with a variety of 16 languages as a L2 across participants. No English-speaking participant was fluent in French. They were shown silent videos of bilingual French-English speakers pronouncing sentences either in English or French. Participants were presented with two consecutive randomly selected sentences and asked to indicate if sentences were from the same or a different language. The results showed that adults who had acquired English as a native language or as L2 before the age of 6 could differentiate between the videos of French and English visible speech significantly better than chance. In contrast, participants who had learned English as L2 after the age of 6 could not distinguish languages. Weikum et al. argued that L2 age of acquisition is crucial in detecting relevant visual aspects in speech perception. Furthermore, they argued late L2 acquisition is hindered by low sensitivity to non-native visual aspects of speech, just as in the auditory speech domain.
In sum, limited studies using lip-reading or visual presentation of speech in L2 suggest that it may be possible to differentiate native from non-native visual speech but the results are not conclusive across languages. Rhythmic differences between languages have been the focus of attention in the small number of studies. Research to date has not addressed the issue of L2 assimilation into native visual speech. This research gap serves as the primary motivation for the current study. Russian will be used as the non-native language and stimuli will be chosen according to similarity and dissimilarity to the native language (English). Familiarity with the Russian language in Australia is low as it is infrequently spoken and excluded from the most frequently spoken languages in Australia (Protopopov, Kosij, & Felner, 1995). Russian is thus a suitable domain to investigate assimilation of visual aspects of L2 speech in speakers of Australian English.

No review of ecologically valid speech is complete without discussion of audio-visual speech. Although speech is encountered exclusively in the auditory domain (without visual cues), for example, when speaking on the phone, speech-reading on its own rarely occurs during natural interaction. Audio-visual (AV) speech processing, and more specifically AV processing in L2, is the next focus of this literature review.

1.2.6 Adult AV speech

Early studies of adult AV speech perception focused on how visual cues enhance perception of degraded auditory stimuli in controlled speech. In a classic study, Sumby and Pollack (1954) embedded words in noise, varying in speech-to-noise ratio from low to high. In one condition, participants heard words via headphones without seeing the face of the speaker and in the other they heard words while seeing the face. Participants had to choose a spoken target from a list of alternatives. Results showed that seeing a face speaking had more impact as less auditory information was provided. Similar results showed that seeing the face speaking helps in degraded listening conditions (Grant, Ardell, Kuhl, & Sparks, 1986; Summerfield, 1979), with syllables, words, and sentences (Dodd, Oerlemans, & Robinson, 1988).

Visible speech also has an effect on sentence processing (MacLeod & Summerfield, 1987). Intelligibility of sentences presented in a background of noise improved when the talker's face was visible. Early studies, such as that of Catford and Pisoni (1970), showed that seeing a person articulating non-native speech facilitated more accurate pronunciation compared to hearing sounds without seeing a talker's face. At the sentence level, the percentage of words
correctly shadowed was larger when the talker’s face was visible when participants had to shadow text spoken in a second language with a foreign accent (Dodd & Campbell, 1987).

Stimuli presented in noise show complementarity of the auditory and visual aspects of speech. Campbell (1987) presented participants with consonant-vowel syllables in three conditions A-only, V-only, and AV and asked them to match each with a known syllable. Syllables were embedded in various levels of noise. The results indicated complementarity between auditory and visual aspects of speech: in the V-only condition the most informative feature was coronal and the least informative was voicing, while in the A-only condition the most informative features were voicing and the least information was coronal. In the AV condition, both aspects were utilised as response accuracy increased significantly compared to A-only and V-only condition. Erber (1982) investigated the relative contribution of each modality to distinguishing the place of articulation. They contrasted three groups: hearing impaired, deaf, and normal controls. The results showed that deaf participants’ performance under A-only and AV condition was equivalent; thus they gained no additional information from bimodal presentation. In contrast, performance of hearing-impaired participants improved significantly under the bimodal condition because they could add the place of articulation dimension in the AV condition to voicing and nasality dimensions present in the A-only condition. Dodd (1987) presented participants with monosyllable words in white noise either in AV or A-only conditions. Results revealed that in the A-only condition the participants’ discrimination of front consonants was worse than for middle and back consonants, and that AV condition was the most informative for front consonants.

The seminal study by McGurk and McDonald (1976) demonstrated the impact of AV speech in typical hearing conditions as opposed to suboptimal conditions. They demonstrated this by dubbing visible articulation of /ba/ with the auditory sound of /ga/ which results in 95% of participants reporting a fused syllable /da/. The researchers concluded that both sources of information- visual and auditory, impact on perceptual recognition of place of articulation but not voicing and manner of articulation - which is determined by the auditory component. McGurk and McDonald proposed that manner of articulation is perceived auditorily while place of articulation is perceived by visual sense but when information is combined the visual information dominates perception. However, this claim was refuted by Summerfield (1979) who showed that visual information alone did not dominate speech perception but contributed additional information. McGurk (McGurk & MacDonald, 1976) extended the findings from the syllable to word and sentence levels. When the visual sentence "My gag kok me koo grive" was dubbed with the auditory sentence "My bab pop me poo brive", the participants reported hearing "My dad taught me to drive". The McGurk effect has been replicated.
numerous times under different conditions (Green, Kuhl, Meltzoff, & Stevens, 1991; P. K. Kuhl, Green, & Meltzoff, 1988; Munhall & Tohkura, 1998; Norrix & Green, 1996; Saldana & Rosenblum, 1994; Sekiyama, 1997; Walker, Bruce, & O'Malley, 1995); selective attention (Massaro, 1987) when audition lags behind visual presentation (Munhall, Gribble, Sacco, & Ward, 1996) and if speakers’ faces are rotated or inverted (Jordan & Bevan, 1997; Massaro et al., 1993).

The McGurk paradigm has been shown to provide a good platform for tapping into potential processes involved in AV integration. The latter was experimentally tested by Nahorna and colleagues (2012). Either a coherent or an incoherent AV context at various durations was presented prior to a congruent “ba” AV target or prior to an incongruent “McGurk” target combining an audio “ba” with a visual “ga”. Participants were asked to monitor “ba” or “da” stimuli. The researchers hypothesised that the McGurk effect would be reduced if the stimuli were preceded by an incoherent context. They manipulated the context by dubbing the auditory input on a random sequence of video sentences (large incoherence) or video syllables (small incoherence). The result showed a reduced McGurk effect in the incoherent contexts, even for short context durations of less than 4 seconds. These findings were interpreted as evidence against AV integration being an automatic process. Nahorna and colleagues suggested that AV integration is a two-stage process, beginning with the binding of audio and video information, followed by an AV fusion stage. The binding stage occurs early in the process of speech perception and allows extraction and grouping of appropriate attributes of the auditory and visual input. AV fusion would occur later and would depend on the results of the preliminary binding stage. They interpreted their results as evidence for unbinding, with participants selectively reducing weight on the visual input. The researchers argued that the binding system assesses the amount of coherence between the auditory and visual streams and modulates the decision process accordingly, supporting a two-stage process with AV binding preceding AV fusion. The idea of selectively varying or reducing the weight of the visual input is directly relevant to the topic of the current study, which investigates non-native audio-visual speech processing. Therefore, we now turn to a brief review of relevant studies of the cross-linguistic McGurk effect.

1.2.7 Cross-language McGurk effect

One of the most controversial findings in the study of cross-linguistic McGurk effects is that the size of the effect varies across languages. For instance, Sekiyama (Sekiyama, 1997) presented Japanese and American participants with the McGurk task comparing Japanese
Visual perceptual assimilation of non-native speech

speakers with English speakers. They found a stronger McGurk effect with non-native speech stimuli. Generally, Asian language speakers show a weaker McGurk effect than speakers of other languages (Sekiyama, 1997; Sueyoshi & Hardison, 2005) but differences are reported within a language family. For example, Werker and colleagues (Werker, Frost, & McGurk, 1992) presented the same visual representations of phonemes paired with the auditory phoneme /ba/ to French and English speakers. There was an inverse relationship between speaker proficiency in English and the size of the McGurk effect—the more proficient non-native speakers showed a weaker McGurk effect in their second language whilst novice speakers had the strongest McGurk effect. Werker et al. argued that visual speech perception is affected by language experience and listeners rely more heavily on visual input when auditory perception becomes challenging. Thus exposure to non-native stimuli seems to moderate the McGurk effect.

The reason why the McGurk effect varies across languages is contested. Harrison (Harrison, 2016) used the McGurk task to assess audio-visual integration in Spanish learners of English and native controls using the /s/-/ʃ/ distinction which does not exist in Spanish. Non-native English speakers showed the McGurk effect when misperceptions were plausible. However, in implausible conditions, such as audio /g/ paired with video /b/, resulting in the sound /bg/, non-native participants did not show the McGurk effect. The McGurk effect was weaker for native English controls compared to Spanish learners of English. The weaker McGurk for native English speaker was not expected, although they showed an increase in reaction times for implausible items. Harrison suggested that non-native speakers have to be taught visual information to process non-native speech but the /s/-/ʃ/ distinction is too well developed in native speakers and does not gain benefit from visual information.

Some researchers argue that cross-linguistic differences in the McGurk effect reflect confounding variables such as limited sets of minimal pair stimuli needed to elicit a reliable McGurk effect. For example, Hazan and colleagues (2010) manipulated language-specific factors such as visually different contrasts in native and non-native languages and non-native and phonemic inventories in relation to visual and auditory contrasts as well as visual attributes of speech relative to auditory attributes of languages. They also manipulated speaker and listener-related variables: visual salience, visual bias of a given perceiver, and clarity of a voice. They used five speakers of two languages (Australian English and Mandarin Chinese) as well as numerous iterations of congruent and incongruent AV, AO, and VO stimuli. They also tested the effect of auditory (pink noise) and visual (blurring) degradation on performance. There were three groups of participants: Australian, British English, and Mandarin Chinese participants. The first two groups were included to gauge the
effect of accent on perception of a native language. The results showed no differences between Australian and British participants. English-speakers utilised visual attributes of non-native speech more than Chinese participants in the AV condition but the most striking finding was that the effect of visual cues varied across participants, showing considerable inter-subject variability. The researchers concluded that utilising visual and auditory attributes of speech is a flexible process. They emphasized that random factors (L1 and L2 features and individual differences) moderate the perception of visual aspects of speech. Are these conclusions valid outside of the McGurk effect paradigm?

This question is relevant because the reliability of the McGurk effect has been criticised lately. For example, Basu-Mallick and colleagues (2015) report significant individual differences ranging from 0% to 100% of participants who do not show the effect. They conducted two experiments using a large sample of English-speaking participants. A sub-selection of the sample was re-tested following a 12-month period. Not only did the results showed that McGurk effects is perceived differently by different individuals, they also observed significant variability from (17% to 58%) of how frequent McGurk effect was perceived across 12 different McGurk stimuli. Moreover, they argued that the assumptions of normal distribution of responses are not valid as about 77% of participants deviated from the normal distribution and almost never or almost always perceived the McGurk illusion. In contrast, there was almost no intra-individual variability between the first and second (12-month period) testing. In the final experiment, they used 8 new McGurk stimuli in an open-choice vs forced-choice design. The results showed that the forced-choice design resulted in a greater frequency of McGurk effect (18% greater) but similar inter-individual variability. The researchers argued that the individual differences should therefore be taken into account when interpreting McGurk-based studies.

Recently, there have been more questions raised regarding the McGurk effect being used as a measure of audiovisual integration in speech perception. Van Engen and colleagues (2017) looked at how the individual McGurk effect is related to the ability to identify sentences in noise in audio-only and AV modes. The results failed to show a relationship between participants' McGurk effect size and their ability to extract visual cues when presented with sentences in noise. The researchers concluded that McGurk susceptibility may not be a valid measure of AV integration (Van Engen et al., 2017).

Furthermore, the McGurk effect does not provide an ecologically valid paradigm, as under most naturalistic conditions there is no conflict between auditory and visual input. Despite
these criticisms, the contribution of the McGurk effect cannot be underestimated, since it shows the contribution of visual speech under clear speech conditions. However, it is not obvious how much benefit AV speech provides compared to auditory-only, especially when processing non-native speech. The next section will focus on this question.

1.3 AV benefit effect

Early studies, such as that of Reisberg and colleagues (Reisberg, 1978), demonstrated that shadowing (repeating) of non-native passages was improved by the provision of visual cues. Catford and Pisoni’s (Catford & Pisoni, 1970) results showed that seeing a person articulating non-native speech facilitates accurate pronunciation when compared to hearing sounds without seeing the speaker’s face. More recent studies used a training paradigm to investigate AV benefit on L2 perception. The study by Hardison (Hardison, 2003b) investigated whether audio-visual training results in better learning of non-native speech sounds compared to auditory-only training. She hypothesised that visual cues would facilitate Japanese speakers’ discrimination of /r/ and /l/ to a different degree depending on their position and context. She was also interested in examining the degree to which audio-visual training enhances generalisation of performance to new stimuli and new speakers. She examined both perception and production performance in Japanese participants who were at an intermediate level of English proficiency. Half of them received AV, A-only, and Video-only training while the other half had A-only training. A control group received no training. The /r/ and /l/ sounds were presented in different positions (initial vs. final) and with various adjacent vowels (rounded vs. unrounded). The stimuli were recorded by three female and two male English native speakers. The training took place over 15 sessions, spread over 3 weeks. The results showed that both experimental groups improve in accuracy with training compared to the control group, not only in the perceptual domain but also in the production of difficult contrasts. Moreover, AV training enhanced perceptual accuracy more than A-only training. In line with her hypothesis, Hardison found that AV training facilitated perception and production of /r/ and /l/ contrasts when they were presented in the initial position with rounded vowels which is the most difficult for Japanese speakers. There were also significant differences in the participants’ performance across talkers during the training phase. Furthermore, post-training tests of generalisation showed that accuracy was influenced by talker familiarity in the Visual-only condition only.

Similar questions were investigated for the Korean learners of English (Hardison, 2003b). The same number of participants was divided into two groups: AV and A-only training and further
divided into multiple versus one-talker training. The results replicated the finding of a significant improvement in perception and production in both experimental groups for non-native speakers compared to controls, as well as better performance for AV training compared to A-only training. A significant effect of talker was also found with lower identification accuracy for the single-talker group when tested with unfamiliar talkers. Unlike the results with Japanese learners however, Korean participants performance was better when /r/ and /l/ were present in the initial rather than in the final position, as predicted by the features of Korean phonology. Korean speakers also experienced difficulties with the rounded vowel context similar to the Japanese participants. Hardison concluded that training in the AV modality is superior than training in the A-only modality, particularly in non-native positions, such as in the rounded vowel context and in the initial position for Japanese speakers and in the final position for Korean speakers. Hardison (2003) has also argued that training difficult contrasts such as /r/-/l/ generalises to early word identification in connected speech. This conclusion would be relevant to the L2 acquisition field, however, some studies caution against this conclusion.

Specifically, Hazan and colleagues (2005) recruited Japanese participants who had been learning English in Japan for 6 years and native English controls. Participants were presented with consonants /p/, /b/ and /v/ embedded in nonsense words to test consonant identification with little influence of lexical knowledge. They were then presented with pairs of real words with /b/ and /v/ in one pair, and /b/ and /p/ in the other. Stimuli were presented in AO, VO, and AV conditions. The results showed that Japanese learners of English showed much lower identification of English labial (/b/-/p/) and labiodental (/v/) consonants in the pre-test VO condition compared to native English speakers. Whilst Japanese participants showed the AV benefit effect for that labial/labiodental contrast, they failed to show the AV effect following the perceptual training of the /l/-/r/ contrast. Clearly, there is a need to replicate the AV benefit effect both across and within languages.

The issues was addressed with Spanish learners of English (Hazan et al., 2006). The focus was on the phonetic contrasts in English that are difficult for native Spanish speakers and found that visual information facilitates learning of non-native speech. Specifically, visual cues affected the assimilation of English voiced and voiceless fricatives to Spanish voiceless phonemes and English voiced and voiceless stops into Spanish voiceless counterparts. They reported that the utility of visual cues differs for native English speakers and native Spanish learners of English. AV presentation improved identification of English consonants for native speakers but only improved vowel-consonant (VC) identification for non-native speakers. Despite beneficial effects of AV modality in all groups, presentation of visual cues alone did
not improve performance. The researchers argued that Spanish participants ignored the visual cues that are associated with non-native contrasts that are phonemic in English but allophonic in Spanish because such contrasts are not native.

Others have argued that an AV benefit effect is evident for the contrasts that are assimilated auditorily in bilinguals. For instance, Burfin and colleagues (2014) presented French native speakers with the Spanish inter-dental fricative phoneme /θ/ that does not exist in French. When presented auditorily, the French participants assimilated the phoneme to the closest native category /ʃ/. However, when the same phoneme was presented audio-visually, they correctly identified up to 80-90% accurately. In a follow up study, a phoneme identification experiment was used, with a Bengali dental-retroflex contrast that is absent in the native language of participants. An ABX design was utilised (A token, B token, X-stimulus) and presented their stimuli in audio-visual and auditory-only mode. Participants were requested to press one key if the A token was the same as X or another key if B is the same as X. The results showed that in the auditory-only condition, the listeners assimilated the Bengali dental contrast to their native language’s dental contrast. In contrast, in the AV condition, they were able to overcome phonological deafness to identify Bengali dental contrasts. The audio-visual benefit effect was greater for native compared to non-native phonemes. However, they used bilinguals from 9 different languages, which may have masked the inter-linguistic differences.

One question remains whether similar conclusions follow from language pairs that are rarely encountered or less familiar to participants. Moreover, there is a need to demonstrate an AV benefit effect with naïve participants rather than bilinguals. Davis and Kim (Davis & Kim, 2001) investigated how naïve participants (English speakers) with no knowledge of Korean process non-native AV presented Korean speech. In the first phase participants saw the upper or lower half of a face pronouncing a phrase in Korean. Each phrase was repeated three times before participants were requested to repeat the phrase aloud. In the second phase, participants were requested to make a forced-choice, indicating whether they recognised the phrase. Participants repeated and recognised a phrase in Korean more accurately after seeing the lower half of a face. In a subsequent study (Davis & Kim, 2004), the effects of viewing the speaker face were investigated further. English speaking participants were asked to monitor target Korean syllables in phrases. They found an AV benefit effect for non-native speakers compared to AO face presentation without motion.

Gullberg and colleagues (Gullberg, Roberts, & Dimroth, 2012) asked if naïve listeners are able to access lexical knowledge after a minimal exposure to a new language. Naïve Dutch participants were exposed to AV input in Mandarin manipulating word frequency and gesture frequency. Results showed that adult listeners can learn complex AV input in an unknown language by using gestural cues, word frequency, and syllable structure.
In sum, AV presentation gives a benefit to non-native speech perception and production when compared to AO and VO speech conditions but this benefit is not clearly demonstrated with naïve participants outside of a training paradigm. Furthermore, several studies failed to show an AV benefit or report that the effect varies across languages. Those studies highlight the need for replication of the AV benefit effect in different languages and different experimental conditions.

1.3.1 Language-specific AV benefit

Taitelbaum-Sweed and Fostick (2016) investigated AV perception over the lifespan. They presented different age groups with meaningful and nonsense words pronounced by the same female Hebrew speaker in AO, VO, and AV modes. The results showed the AV benefit effect was smaller for Hebrew speaking children than English speaking children. They attributed the differences to fewer vowels and consonants in Hebrew compared to English and argued that mode of presentation interacts with chronological age as well as language features. In the VO condition, older children were more accurate than younger children whereas a beneficial effect was seen in the AV condition by the age of 8 and then decreased with age. They urged future researchers to investigate the interactions between AV benefits and language features.

Some studies find no AV benefit at all. For example, Hazan and colleagues (2005) failed to replicate Hardison’s (Hardison, 2003a) results. In that study, perception of difficult contrast /l/-/t/ for Japanese learners of English was investigated using audio, visual, and audio-visual conditions. Proficiency in English ranged from low to lower-intermediate. Results revealed that identification of the /l/-/t/ contrast was no better with AV mode of presentation than AO and was significantly worse in the VO condition compared to the AV and AO conditions. A training phase was also carried out with a subtest of 41 participants. Half of the participants had audio-visual training while the other half had auditory only training. Results revealed that both groups showed a significant improvement with training, however, audio-visual training resulted in greater improvement in a smaller proportion of participants (18.5%) compared to the auditory only training. While there was no evidence of audio-visual benefit, the audio-visual training group improved in the VO condition between pre- and post-test. The researchers concluded that training enhanced sensitivity to visual cues. Moreover, they argued that participants with better scores in the VO condition at pre-test are more likely to have ‘visual awareness’. Contrary to prediction, the analysis found no differences between groups at post-test. Unfortunately, the visual-only condition was utilised for the pre-test only.
To address this issue further, Hazan and colleagues (2006) replicated their 2005 findings with the addition of a VO condition by focusing on a single contrast of /p/-/b/-/v/ in Spanish learners of English. They were interested in whether visual cues provide information about place and manner of articulation, and they predicted that the visual distinction between /b/ and /v/ would facilitate performance while limited information would be provided by visual cues on voicing distinction in /b/-/p/ contrasts. Participants with low intermediate level of English language proficiency participated in three conditions: AV, A-only, and V-only. The results showed that identification in the AV condition was no better than the AO condition. Moreover, contrary to prediction there was no difference in confusion of place and manner of articulation between the AV and AO conditions. Comparisons between the AO and VO conditions showed that a subgroup of participants acquired the ability to distinguish /b/-/v/ contrasts, regardless of whether AO or VO cues were utilised, in contrast to participants who were unable to acquire this ability when acoustic cues were present. The researchers suggested that sensitivity to non-native contrasts may not be present at early stages of second language acquisition for acoustic information and for visual cues too, as shown by lack of benefit when visual cues are present. This conclusion carries implication for L2 language acquisition teaching but further replication with other languages is necessary.

The aforementioned study highlighted the fact that the AV benefit depends on the degree of exposure to audio-visual attributes and proficiency level in L2. Werker and colleagues (1992) compared English and French-speaking Canadian participants on their perception of a voiced interdental consonant in the context of the vowel /a/, which is challenging for native French speakers. The researchers first obtained a baseline of auditory perception of difficult contrast by 10 English speakers and 10 French speakers with an intermediate level of English proficiency. They found that while both groups did not differ in perception of the phones that are used in both English and French, perception of the interdental stimulus in French speakers was significantly worse than in English speakers. In the second part of the study, Werker and colleagues were interested in tracing how level of proficiency in English affects utilisation of visual cues. Specifically, they predicted that the more proficient the participants are the more likely they are to use visual cues. In this part AO, VO, and AV conditions were included for each participant. Surprisingly, the researchers found no difference between participants with different levels of English proficiency in the A-only condition but found the participants who were less proficient in English were less influenced by visual information. This suggests that there may be a different trajectory for acquisition of the auditory and visual cues in a non-native speech.
Visual perceptual assimilation of non-native speech

The issue of language proficiency was investigated further in the study of Harrison (2016). She examined whether L2 learners of English notice visual details of L2, comparing the initial stage of language acquisition and intensive L2 training. She chose fricative English sounds "s" (as in “save”) and "sh" (as in “shave”) presented to English L2 learners and English control participants. There were 21 single syllable words in AV congruent, AV incongruent, AO, and VO conditions. Results showed that non-native and native speakers were near perfect in all conditions except the VO condition. She concluded that auditory input is primary and visual cues have little effect on discrimination of /s/ and /ʃ/. For non-native speakers, performance did not improve in AO and AV conditions after a period of intensive training in L2 but became faster in the VO condition. This finding needs to be replicated with other languages and contexts.

Language-specific AV benefit effects were also reported by Burnham and colleagues (2015). They compared three tonal languages (Thai, Cantonese and Mandarin) to a non-tonal language (English) and a pitch-accented language (Swedish). In the first experiment they presented participants with 6 Thai syllables each with one of 5 Thai tones pronounced by a single female Thai speaker. Stimuli were presented in AO, VO, and AV conditions. The AO condition had two manipulations -noisy (Thai speech bubble) and clear. Participants viewed two videos of a speaker and were requested to indicate whether two tones were the same. The results showed an interaction between benefit and language, with native Thai speakers performing better than all groups in AO and AV conditions. Mandarin speakers were better than Cantonese, who were in turn better than Swedish (pitch-accent language), who were better than English speakers (non-tonal and non-pitch accent). Results were interpreted as evidence of language-specific benefits in speech perception e.g. in the VO condition, English speakers performed better than all other groups reversing other effects. The researchers argued that visual attributes of speech are used for processing tones regardless of language-specific experience. In contrast, language experience has an effect on the VO condition in clear contexts. Thai, Cantonese, and Swedish listeners all perceived VO tones better than chance, whereas Mandarin listeners did not. English listeners perceived a VO tone better than chance in clear and noisy conditions and outperformed all other groups. The researchers concluded that the AV advantage in speech perception is constrained by language specific features. Thus, English speakers were not able to integrate auditory and visual tone information as effectively as other language groups.

Burnham and colleagues (2015) also tested tone discrimination in different auditory contexts: with pure speech, filtered speech and violin sounds. They added low-pass filtered speech or
violin sounds to Thai tones and varied inter-stimulus intervals under the presumption that longer (1500ms) presentation involves deeper processing. They used a same-different paradigm and 4 groups: Thai, Cantonese, Swedish and English speakers. They predicted that English listeners would discriminate F0 patterns better if they were presented in a non-speech context than a speech context. Results showed that the performance of English speakers improved as the tone stimuli became less speech-like. No such pattern was observed for other language groups. Burnham and colleagues argued for a language specific effect originating in infancy via perceptual narrowing for lexical tones.

Debate continues about cross-linguistic AV effects. Studies show that the visual attributes of speech do not benefit and may hinder speech perception under certain conditions. Kawase and colleagues (2014) presented English syllables produced by Japanese learners of English versus native English speakers of East-Asian descent. They presented stimuli in AO, VO and AV modes. The consonants included speaker contrasts that are difficult for Japanese participants (e.g., /v/ and /θ/, and /l-ɹ/ contrasts). The researchers found that native English speakers perceived production of Japanese-like phonemes (/b, s, l/) better than the phonemes that do not exist in Japanese (/v, θ, ɹ/). This was consistent across all modes of presentation. Although visually distinguishable non-Japanese phonemes /v, θ/ were perceived significantly better when presented in the AV compared to the AO condition, the effect was reversed when perceiving the non-Japanese phoneme /ɹ/. The researchers concluded that if Japanese speakers pronounce English contrasts incorrectly, the visual information from that production hinders AV perception for native English speakers. Thus, visual cues from the productions of non-native speakers can generate an AV detriment effect.

The issue of an AV detrimental effect for the perception of accented speech was investigated by Banks and colleagues (2015). This ability to ‘tune in’ to accented speech is a part of many people’s every day speech perception and is a common perceptual adaptation which is robust, present throughout life, and occurs with unfamiliar accents. In the first experiment, participants had training with a novel accent in AV and AO conditions (with or without background noise) and in a VO condition. In the second experiment the participants listened to a non-native accent in AV or AO mode to see whether a greater amount of continuous exposure to AV stimuli without training produces a larger AV benefit effect. Results showed that although there was an AV benefit for the recognition of accented speech in noise, neither experiment gave an AV benefit for perceptual adaptation of accented speech when compared to AO stimuli. These results were similar to a study by Janse and Adank (2012, cited in Banks et al 2015), who report no AV benefit for perceptual adaptation to accented sentences in older individuals.
A speaker’s ethnic background can also produce a bias in the perceiver. Yi and colleagues (2013) were interested in the question of how speakers perceive their native-language accent and whether AV presentation helps perception. They presented native English speakers with AV stimuli of native English (Caucasian-looking) and native Korean (Asian-looking) speakers producing sentences in English. The results showed an AV benefit effect but the effect was smaller for the English sentences produced by a Korean speaker than an English speaker. The results suggest that participants exhibit visual bias that is based on ethnicity. However, an alternative hypothesis is that the bias results from a lack of experience with non-native speaker as an interlocutor. When native English speakers saw Korean faces pronouncing English sentences, they rated them as more accented suggesting lack of familiarity. The match between the ethnicity of the speaker and participants highlights a potential effect of context on the size of the AV benefit effect.

The current study will avoid this confounding variable because of the Caucasian ethnicity of the Russian language speaker. It also aims at contributing to the debate of whether the degree of AV benefit is language-specific and the potential reason for these differences. One factor that has been proposed to account for cross-linguistic differences in the AV benefit effect is speech synchrony. Therefore, we will now turn to an overview of this area and its relevance to the study in question.

1.3.2 Speech synchrony

The issue of how synchrony between audio and visual input streams impacts on AV speech perception dates back to an early study by Koenig (1965). In that study the participants were tested with low-pass-filtered speech consisting of single words and sentences. Speech comprehension was not disrupted until the sound delay exceeded 240ms. In another experiment, participants were presented with a sound track in which some words were replaced by rectangular pulses corresponding to the closing of the talker’s vocal folds (MacLeod & Summerfield, 1987). This manipulation improved identification of words relative to the silent speech-reading condition, but a delay of the sound track by 80 ms disrupted performance and a delay of 160ms resulted in a chance performance. When sentences were presented audio-visualy using audio delays of up to 300ms, delays greater than 120 ms interfered with performance. In another study, shadowing of a difficult passage was more accurate when the face of the speaker was visible but this effect disappeared when asynchrony of the audible and visible speech was half a second (Arnold & Hill, 2001).
Smeele and colleagues (1998) used seven asynchronous conditions. They found that performance was most accurate when two sources of information were consistent. The advantage of the two sources of information over just one was robust at asynchronies up to 250 ms. Consistent information did not help at the long lag of 533 ms. The researchers concluded that long delays disrupted integration of visual and auditory inputs.

Kim and Davis (2004) also investigated the impact of temporal asynchrony on the AV benefit effect. They showed that auditory speech detection in noise was facilitated by seeing the lower half of the face but that this facilitation effect was not present when there was sufficient asynchrony between AV signals. In a recent study, Kim and Davis (2014) examined the effect of visual speech’s timing on auditory speech by pairing speech/non-speech with a talker’s static/moving face. With the moving face, there were two conditions: full face version and modified face version with only peri-oral motion available. Whilst the visual cue of peri-oral motion was only matched with speech stimuli, its provision resulted in better discrimination times for both speech and non-speech compared to the static face. Further, a facilitation effect on discrimination was found for the full mouth region version compared to the peri-oral condition but this effect only occurred for the speech stimuli. Kim and Davis argued that by providing a visual cue the auditory processing becomes ready or re-organised (re-set) for processing of the auditory stimuli. This re-setting results in increased sensitivity to the auditory input. This idea is similar to that proposed in Thorne and colleagues (2011, cited in Kim and Davis 2014) who demonstrated that pure tone frequency was discriminated better when a visual prime (white rectangle) was presented. They demonstrated that the visual primes changed the oscillatory activity of alpha and theta frequencies of the auditory cortex.

Several studies investigate the underlying mechanism of the temporal relationship between auditory and visual stimuli. For example, Lewkowicz (1996) identified an “inter-sensory synchrony window” during which it is possible to integrate de-synchronised auditory and visual stimuli. Several studies investigated this effect (for a review see Conrey and Pisoni 2006) and can be summarised by three main findings: first, the window lasts several hundred milliseconds; second, the time window is not symmetrical and visual-leading asynchronies are more challenging. Third, there is considerable inter-individual variability in detection of the inter-sensory window.

The finding of individual variability in detection of the AV window has received considerable attention. One reason is that better detection of AV synchrony may be associated with better AV speech perception. For example, Conrey and Pisoni (2006) investigated the detection of AV synchrony in speech and non-speech perception by measuring sentence intelligibility of
the same participants in three modes: AO, VO, and AV. Auditory signals were degraded by time reversed division of speech-signals into 80-ms long segments and then reassembled. This was done to avoid ceiling effects in AO and AV speech. Results showed significant within-participant correlations between AV synchrony detection for speech and non-speech signals. Furthermore, people who were better at identifying AV sentences had a smaller AV asynchrony window - that is they were better at detecting timing differences between auditory and visual signals. However, there was no relationship between VO speech perception scores and AV asynchrony detection. Similar results are reported in Grant and Seitz (1998, cited in Conrey and Pisoni, 2006).

Further understanding of the temporal relationship between auditory and visual speech comes from research by Chandrasekaran and colleagues (2009). They identified some natural statistics of audiovisual speech by describing 4 major characteristics. First, there is a robust correlation between motor mouth opening gestures and the acoustic envelope. Second, correlation is maximal in two regions: one around 300–800 Hz (first vocal tract resonance (formant) F1 and the other around 3000 Hz (F2 and F3). Third, temporal co-modulations of the mouth and acoustic stimuli appear in the 2–7 Hz frequency range that correspond to syllabic rhythm. Fourth, the timing of visual articulation relative to the onset of the auditory stimulus is consistently between 100 and 300 ms in advance of speech.

Such findings led to research devoted to the question of the mechanisms of auditory speech lagging the visual signals. Schwartz and Savariaux (2014) argued that this lag of an auditory signal behind the visual signal is only specific under certain experimental conditions - at the start of speech utterances or what they call preparatory sequences of speech. They defined them as the gestures that are visible but produce no sound. In contrast, they made a distinction of chained sequences of speech, or so called co-modulatory gestures of speech. They defined them as the visual gestures which produce both visual and auditory signals more or less in synchrony and as argued by Schwartz and Savariaux (2014) these types of co-modulatory gestures are far more common in speech. In their experiment they presented audio-visual data on plosive-vowel syllables (pa, ta, ka, ba, da, ga, ma, na), produced either in isolation or in sequence. Their results showed that when these syllables were produced on their own, preparatory gestures provided audiovisual asynchronies similar to those observed in previous studies, such as that of Chandrasekaran et al. (2009). In contrast, the chained sequences that provided co-modulatory gestures, results in precise audiovisual synchrony. Schwartz and Savariaux argued that the natural coordination between visual articulation and auditory sound produce both lead and lag of the visual input. They further argued that small audio leads may be produced as a result of the absence of visibility for certain auditory stimuli whilst large
video leads may be produced as a result of labial anticipatory co-articulation. Finally, they stressed that visual input may still improve predictions, even if auditory and visual input are not asynchronous.

How is the finding of a temporal relationship between auditory and visual speech relevant to L2 audio-visual speech processing? Navarra and colleagues (2010) compared linguistic processing of a familiar language in contrast to an unfamiliar one. They asked native speakers of English and Spanish to make judgments of visual and auditory speech sentence synchrony. They reasoned that if language familiarity is as important as visual saliency, there should be an increase in the typical visual lead when presented with an unfamiliar language. Indeed, the results showed that visual information has to be presented earlier in time for an unfamiliar language compared to the native language for successful AV integration. They predicted that if a visual signal precedes an auditory signal, the resulting audio-visual signal is perceived as more synchronous in the native language than in an unfamiliar language. In contrast, the “visual anticipatory effect” showed no difference in bilingual Spanish-English speakers. The researchers proposed that attention in bilinguals may be differentially distributed across auditory and visual streams when people process L1 versus L2. Specifically, they argued that as the availability of attention resources diminishes, so does the influence of visual information on speech perception. Thus, they reasoned that attention is distributed differentially in native and non-native languages for bilingual speakers and this in turn impacts on visual and auditory integration. Attentional resources would potentially point to a top-down process as opposed to a bottom-up process when using specific visual gestures. However, the researchers pointed out that both processes could be recruited.

In sum, input synchrony is important in AV speech perception and the ‘visual anticipatory effect’ is influenced by whether stimuli are presented in a native or a non-native language. This issue is informative to the question of how visual aspects of speech impact on L2 speech processing. Furthermore, another question is whether a specific prediction could be made about audio-visual perception of L2 when the second language is characterised by a particular pattern of inherent asynchrony. For instance, one study showed that English has greater asynchrony between visual and auditory speech units compared to Russian (Karpov, Krnoul, Zelezny, & Ronzhin, 2013). It is therefore expected that native English speakers will be sensitive to such information during the AV processing of Russian natural stimuli.

This question will be partially addressed in the current study by using Russian language, not previously explored despite its inherent patterns of asynchrony. Given the evidence reviewed, it is anticipated that asynchronies might contribute to visual and AV speech perception by
native English speakers who are not familiar with Russian. Whilst the current study will not manipulate synchrony between different input streams and instead relies on the natural inherent differences between Russian and English language, another type of manipulation will be used. Specifically, the current study will use different rates of speech presentation. The reason for this will be justified following a review of the relevant research.

1.3.3 Rate-dependent speech perception

Listeners adapt successfully in real life to variations in speech rate. Perceptual adaptation in speech refers to the ability to adjust to a wide variability of incoming signals, such as different talkers, volume, rate, accent, background noise, and so on. There are methodological issues in relation to experimental manipulations of time-modified speech. Two rate manipulations can be used: in the first, a speaker is recorded at a natural rate and recordings are manipulated by linear compression or expansion using a digital processing technique (Dilley & Pitt, 2010; Reinisch & Sjerps, 2013). In the second method, a speaker is asked to produce the experimental stimuli at a certain rate, e.g. fast, slow, or normal (Newman & Sawusch, 2009). There is an issue with the second method, especially if speakers are requested to produce naturally fast speech (even moderately) as this produces segmental reductions and deletions and creates articulatory undershoot (Ernestus, 2014). Segmental deletion refers to a complete absence of a segment whilst segmental reduction refers to an incompletely realised segment. Vowels are more susceptible to shortening than consonants in perception and unstressed syllables are shortened and deleted more than stressed syllables. Importantly, speech with more instances of reductions also results in more deletions (Ernestus, 2014).

Normalization is a process whereby listeners compensate for segmental reduction/shortening and deletion. A number of studies show that an ambiguous sound is interpreted as perceived for relatively longer following a fast context sentence than following a slow context sentence (Reinisch & Sjerps, 2013). Recently, Reinisch (2016) used German words bannen and bahnen which differ minimally in the /a/-/a:/ vowel duration. They asked listeners to categorise them and found that they performed the task depending on the overall sentence duration and speech speed. Participants reported hearing more long /a:/ responses in a fast speech context, which resulted in more segmental reductions and deletions compared to a normal rate which was linearly compressed to the same duration as a naturally spoken fast sentence. They advocated using linearly compressed speech, rather than naturally produced speech with modified rate in order to control experimental manipulations.
There is a rapid adaptation to compressed speech over the course of exposure, resulting in improved performance until reaching a plateau (Mehler et al., 1993a). For instance, in the study of Golomb and colleagues (Golomb, Peelle, & Wingfield, 2007) participants show quick adaptation to learning 20 time-compressed sentences over four sessions occurring once a week, showing stable improvement each time. In the study of Banai and Lavner (Banai & Lavner, 2012) participants were trained to identify time-compressed speech in a second language, and pre-to post-test results were contrasted between participants and untrained controls. Results showed that training resulted in better perception compared to untrained controls but there was no generalisation to novel stimuli. They suggested that perceptual adaptation of time-compressed speech is less effective in a non-native language. Indeed, studies report highly experienced speakers of a second-language perceive speech in a non-native language as too fast (Zhao, 1997). A compromised ability to perceive sentences that are presented rapidly in noise is also reported for bilinguals (Bradlow & Bent, 2002; Rosenhouse, Haik, & Kishon-Rabin, 2006). Some studies report that increased speech rate hinders perception of non-native speech compared to native speech in bilinguals (Rosenhouse, Haik, & Kishon-Rabin, 2006). Native speakers of Arabic who spoke Hebrew as a second language for at least ten years, were presented with sentences at a slow and fast rate either in clear or babble noise. Target recognition in each sentence was measured. Results showed that Arabic-dominant bilinguals showed reduction of 15% in their recognition of Arabic words when those were presented in fast speech. In contrast, the same native speakers of Arabic showed reduction of 30% in their recognition of Hebrew words presented in a fast rate.

What about slow speech? Several studies show that slow speech has an increased number and duration of pauses (Bradlow & Bent, 2002; Picheny, Durlach, & Braida, 1986; Smiljanic & Bradlow, 2008). Categorical perception of phonemes is improved by slow speech because of increased syllable duration and accompanying changes in segmental cues (J. Miller & Volaitis, 1989; Joanne L. Miller & Dexter, 1988; J. L. Miller, Green, & Reeves, 1986). Studies of cognitive processes show that fast speech demands processing resources (Wingfield, 1996). Others hypothesised that matching incoming stimuli to prototypes in long-term memory takes time and is more challenging when incoming stimulus is fast as compared to slow (Goldinger, Pisoni, & Logan, 1991; Nygaard et al., 1994; Wingfield, 1996).

Despite putative benefits of slow speech, it is debatable whether processing non-native speech at a slow rate is more effective. Some studies suggested that a slower rate facilitates non-native listening comprehension (Chaudron, 1982; Derwing, 1990). In contrast, in the study of Derwing (1990) native participants viewed a short film and then asked to describe the film to
a listener who was either a native or a non-native speaker. Comprehension scores showed that faster speech rates reduced comprehension. However, the results did not support the expectation that slower presentation necessarily results in better comprehension. Another study demonstrated that comprehension was better with a slower speed but only by non-native listeners with low proficiency (Blau, 1991). Furthermore, for less proficient bilinguals, slow and clear speech is less intelligible than for monolinguals (Cutler, Weber, Smits, & Cooper, 2004; Flege, Bohn, & Jang, 1997; Jia, Strange, Wu, Collado, & Guan, 2006; MacKain, Best, & Strange, 1981). Processing of slow speech approaches levels of monolinguals when bilinguals became proficient (Schmidt & Flege, 1995).

Several cross-linguistic studies used a variable rate of stimulus presentation. One such study by Shia and Farooq (2011) investigated how temporal manipulation affects monolingual and bilingual listeners’ recognition of non-native English passages. It was hypothesized that the effect of temporal manipulation would be greater for non-native than native listeners, especially in the presence of noise. Participants were presented with English passages at five different rates. For the two faster conditions, speech was time compressed by 15% and 30%, respectively. For the two slow conditions, the duration of each sentence was lengthened by 1.15 and 1.3 times, equivalent to 87% and 77% of the original speech rate. Sentences were presented both in quiet and in noise. Results showed that bilinguals, including English-dominant bilinguals, recognised fewer targets compared to English monolinguals. However, the effects of speech rate on target word recognition were not significantly different between groups, with all groups showing that speech rate together with competing noise negatively affect performance. The benefit of slow speech rate for bilingual individuals was only evident when speech expansion was beyond 30% of the original rate.

In sum, there is still debate in the literature regarding perceptual speech rate adaptation for non-native speakers. More specifically, it is unclear whether slow rate of speech presentation is useful for non-native speech perception, including native participants. The latter question has clear implication for the 2nd language acquisition teaching. An additional question is whether audio-visual speech benefits or hinders processing of rate-modified speech? This is one of the theoretical questions that motivates the present thesis.

Some studies show that audio-visual speech enhanced perception of speeded-up speech. In the study of Massaro and colleagues (1998) one-syllable words were presented either at a normal rate or three times the normal rate in auditory, visual, and bimodal conditions. At the normal rate of presentation, auditory identification was high with no facilitatory effect of bimodal perception. When presentation was increased unimodally, accuracy was 55% in auditory and
Visual perceptual assimilation of non-native speech

45% for visual (lip-reading). In the bimodal condition, accuracy improved to 73%. Whilst the study of Massaro and colleagues (1998) suggests an AV benefit in perception of speeded-up speech, it is unclear if there is an AV benefit for other rate-modified speech, such as slow presentation and whether the AV benefit for time-modified speech is language-specific.

Why is the question of language-dependence of rate-modified speech important? Besides its implication to the field of 2nd language acquisition teaching, there is the implication for theoretical accounts of speech perception and specifically the issue of whether perceptual adaptation to temporally modified speech occurs at a high-level perceptual stage (post-lexical or semantic) or not. If the latter hypothesis were correct, then adaptation to an unknown language would be difficult, if not impossible. Indeed, several studies support the hypothesis. For instance, Dupoux and Green (Dupoux & Green, 1997) presented participants with 20 sentences compressed to 38% of their original duration. The results showed that identification improved from an initial 30% to 39% for final sentences. Similarly, when sentences were compressed to 45% of the original duration, 78% of sentences could be identified. Importantly, participants adapted to the compressed speech of a male and performance generalised to a female speaker. They interpreted their results as evidence that adaptation occurs not at the perceptual-level but at the abstract level, implying that talker-specific acoustic characteristics do not play a major role. However, Altmann and Young (Altmann & Young, 1993) reported that processing of compressed sentences could be improved after listening to sentences composed of non-words. Such results support the hypothesis that adaptation occurs at a late perceptual stage. If this were correct, we would not expect adaptation to occur to the same extent for an unknown language.

However, several studies show that adaptation to compressed speech can occur to an unknown language. Mehler and colleagues (Mehler et al., 1993b) argued that languages could be classified according to their rhythmic qualities. Romance languages such as French, Spanish, Catalan, Italian, and Portuguese belong to the syllable timed class; Germanic languages such as English and Dutch belong to the stress-timed class; and Sino-Tibetan languages such as Japanese are classified as the mora-timed class. Stemming from this classification, they proposed that processing of L2 is easier if it belongs to the same rhythmic group as L1. Mehler and colleagues (1993) reported data supporting this proposal. They had participants listen to time-compressed sentences at different rates (uncompressed, 50%, and 40% of the original length). They were then tested with the compressed sentences. Spanish-Catalan and other bilinguals showed perceptual adaptation to time-compressed speech in their non-native language, but monolingual Spanish participants also showed an adaptation effect.
Visual perceptual assimilation of non-native speech

when being exposed to compressed Catalan, a language they did not understand. In contrast, no such effect was observed with French and English participants.

Sebastian-Galles and colleagues (Sebastián-Gallés et al., 2000) further investigated whether processing of time-compressed speech is language-dependent. They recruited Catalan-Spanish and French-English speakers as pairs of languages for comparison because Catalan-Spanish have similar phonological properties and French-English have different phonological properties. They reasoned that if higher-level semantic processing has a role in the adaptation to compressed speech, then bilinguals (both Catalan-Spanish and French-English) would be expected to show generalisation from one language to another. In contrast, if low-level phonological perceptual factors are responsible for adaptation to compressed speech, than the phonological proximity of the languages will be expected to cause generalisation of adaptation in monolingual Spanish and Catalan speakers but not in monolingual or bilingual French-English subjects. Sebastian-Galles and colleagues gave participants a set of ten habituation/prime sentences that were time compressed (38-50%) in one language, followed by a set of time compressed (40% or normal rate) test sentences in the same language or in another language. Results showed that comprehension of compressed speech could be enhanced by exposure to another language, but that this effect is language-dependent. Specifically, Catalan/Spanish speakers showed improved comprehension of compressed speech in both languages following exposure to the other language. Moreover, the authors found generalisation from Catalan to Spanish for monolingual Spanish speakers. There was no generalisation between English and French for bilingual speakers or from French to English for monolingual English speakers but remarkably adaptation from Dutch to English. Contrary to the prediction of more adaptation to compressed speech between Spanish and Catalan than between English-French or English-Dutch, the results showed adaptation to compressed speech does not require lexical overlap since monolingual Spanish speakers who do not understand Catalan improved as much as Spanish/Catalan bilinguals. Furthermore, the phonological overlap between Dutch and English was sufficient to produce the same effects. The researchers suggested that although comprehension (lexical overlap) might be sufficient to enhance adaptation, it is not essential. At the same time the capture of low-level acoustic properties of fast speech does not seem to be necessary as evident by the adaption effect even when processing a foreign language.

In contrast, Sebastian-Galles and colleagues (2000) suggested that the necessary level of representation is pre-lexical, which is intermediate between the lexical level and the acoustic/phonetic level. They cited Cutler et al. (Cutler et al., 2004) who suggested that French and English have contrasting pre-lexical representations whereas Spanish and Catalan
Visual perceptual assimilation of non-native speech

share pre-lexical representations. Therefore, the finding that adaptation to compressed speech is dependent on language similarity is compatible with an influence from pre-lexical properties on non-native speech perception. Moreover, Sebastian-Galles and colleagues argued that due to the predictability of syllable stress in Catalan and Spanish (penultimate syllable), participants were able to rely on the regularities in speech without understanding sentences. Some French syllables have predictable stress patterns but English does not have a predictable stress pattern. Therefore, French (non-native) speech will not transfer to English native speakers in the same way as Catalan does for Spanish (and vice versa).

In a follow-up study, Sebastian-Galles and colleagues (Sebastián-Gallés et al., 2000) asked if adaptation to compressed speech can be linked to lexical rather than phonological/rhythmic factors. They argued that Catalan-Spanish and English-Dutch word pairs are not only similar in phonologies and rhythm but have overlapping lexical representations (cognates), morphology, and syntax. In contrast French and English languages do not share such features. In their study, Spanish listeners were adapted to Spanish, Italian, French, English, and Japanese. The same listeners were tested with Greek sentences compressed at different speeds. The results demonstrated adaptation between Southern European languages, namely Spanish, Italian, and Greek but no adaptation between English and Japanese, with French resulting in intermediate adaptation. The researchers concluded that perceptual adaptation to compressed speech occurs mostly at the phonological rather than the lexical level. Moreover, not only the rhythmic properties of the language but also the vowel and/or the lexical stress pattern contribute to these effects. The researchers rejected the alternative hypothesis that adaptation could be explained in terms of rhythm and argued that it must include phonological features, such as vowel properties and stress. They also rejected the hypothesis that lexical information that promotes adaptation to time-compressed speech, postulating instead that this occurs primarily at a phonological level.

An alternative view is that speaking rate influences processing at different levels. Bosker and Reinisch (Bosker & Reinisch, 2017) tested the hypothesis that unfamiliar languages sound faster than native one, using implicit rate perception. They used the fact that there is normalization of speech rate with ambiguous vowels, such as short /a/ and long /a:/ is interpreted as /a:/ following a fast but as /a/ following a slow sentence. They tested whether a tendency to perceive a long /a:/ is more common with a second language. They presented Dutch and German listeners with fast and slow sentences in both languages spoken by the same bilingual speaker. In the follow up, they listened to non-words that contained vowels from an /a-a:/ continuum. Their results demonstrated that there was a consistent effect of rate normalization for both listener groups. The effect was not due to long-term spectral
Visual perceptual assimilation of non-native speech

properties, duration, and number of realised segments. The results were interpreted as evidence that listeners use pre-existing knowledge. Bosker and Reinisch (2017) argued that accented non-native sentences are harder to process than native sentences and this increase in cognitive load may speed up time perception.

In sum, normalisation of speaking rate occurs at multiple levels of processing and perception based on studies of time-modified non-native speech. This contributes to our understanding of speech perceptual adaptation both in native and non-native speech. More studies exploring the issue within the audio-visual domain and putative AV benefits are necessary, as the majority studies are restricted to auditory speech perception. One aim of the current study is to explore the effect of time-modified speech on AV perception of native and non-native speech using Russian, to extend the body of knowledge by pairing time-modified stimuli with the paradigm of cross-modal perceptual matching. Time-modified speech paradigms allow inferences about level of speech processing. However, the cross-modal perceptual matching paradigm tests whether native and non-native speech representations are amodal or modality specific. We describe this paradigm in greater detail below.

1.3.4 Cross-modal repetition priming

In the cross-modal repetition priming paradigm, a stimulus in one modality (prime) is presented followed by the corresponding/different stimulus in the other modality (target) and participants are asked to indicate whether they match or not. It is commonly inferred that changes in response time or accuracy reflect the cognitive processing required for a given task (Bowers & Marsolek, 2003). This paradigm allows examination of whether speech representations of L1 and L2 are amodal or not.

In the general area of cross-modal matching, several experiments conducted by Lachs and colleagues (2004) showed that participants were able to match a video of a speaker's face to the corresponding voice when visual and auditory stimuli were presented separately in time. However, participants were unable to match the vocal source with a static face or when auditory stimuli were temporally reversed. Similar results were obtained by Kamachi and colleagues (2003), who found that a cross-modality sentence matching task can be accomplished despite differences in the linguistic content of two signals. They used a delayed matching to sample task, XAB to demonstrate that participants were able to match the video of an unfamiliar face, X, to an unfamiliar voice, A or B, and vice versa, but only when stimuli were dynamic and not temporally reversed. They argued that time-varying information plays
a critical role in the capacity to perform a cross-modality matching task whilst differences in linguistic content of sentences across modalities did not produce a significant effect. They concluded that identity-specific information is not as closely related to particular utterances as time specific dynamic information.

The cross-modal matching paradigm allows examination of the possibility that hearing a word initiates mental visualisation of articulatory organs, followed by comparing them to the seen articulation. Kaganovich and colleagues (Kaganovich, Schumaker, & Rowland, 2016) argued that visual articulation unfolds over time and it is possible to detect incongruence between the auditory target and silent articulation at a sub-lexical level (i.e., syllabic and/or phonological) prior to the completion of the entire word’s articulation. However, final segments may be crucial (e.g., beam vs. beet as cited in Kaganovitch et al, 2016), necessitating processing of the entire articulatory input at the lexical level. In order to tease out the contribution at different levels of processing, the researchers combined behavioural measures and event-related potentials recordings (ERPs) with a cross-modal repetition priming task. They presented an auditory prime (word), first asking participants whether the target (silent visual articulation) matched the auditory prime. Half of the trials matched (congruent) whilst the other half did not (incongruent). Kaganovitch et al. exploited the fact that during incongruent trials the N400 and the late positive complex (LPC) component of ERP can be detected. The N400 amplitude is shown to be sensitive to incongruence of prime and targets and it precedes the behavioural response when the word is recognized. Whilst the N400 is considered to be a measure of sub-lexical processing, the LPC ERP component belongs to a family of relatively late positive deflections in the ERP waveform. They conducted two experiments: the first included matching a visual target to an auditory prime combined with EEG recording. Within this first experiment, there was a separate lip-reading task for each participant. In the second experiment, participants were asked to listen to words in noise (embedded in two-talker babble masker). In the second experiment, there were two conditions: Audio-only and Audio-Visual. The results showed that incongruent trials elicited early N400s and large LPCs but only the N400 was significantly correlated with improvement in detection of words in noise in the AV condition. That is, individuals with larger N400 differences between incongruent and congruent trials showed greater AV benefit (better detection of words in noise in the AV condition compared to the AO condition). The researchers also found that individuals with smaller LPC to congruent articulations were more accurate on the lip-reading component of the matching task. However, there was no correlation between the behavioural measures of lip-reading accuracy from Experiment 1 and the AV benefit effect in detecting words in noise in Experiment 2. Kaganovich et al interpreted their results as supporting the possibility that participants hearing a word initiated a mental visualisation of articulatory organs that was
then mapped to seen articulation. Thus, individual differences in the N400 were the result of individual differences in the strength of memory traces for correlation between sounds and speech articulatory gestures. The strength of such memory traces would affect how well people can identify words embedded in noise in AV mode.

Kaganovitch et al. (2016) also proposed that auditory working memory is actively recruited in order to match the auditory input with visual articulation and that individual differences in the N400 amplitude results from how well these auditory or sub-vocal processes are maintained during the matching task. This possibility was partially addressed in a study by Van De Zande and colleagues (2014) with long-term cross-modal repetition priming. They demonstrated that even in long-term priming, exposure to auditory-only words as primes facilitated subsequent VO recognition of the same target words. Facilitation was found even if different speakers produce prime and target words. The researchers concluded that lexical phonological representations are amodal i.e. shared for auditory and visual speech.

The results of Van De Zande and colleagues (2014) strike accord with findings from Buchwald et al (2009) by showing that VO exposure primes auditory word recognition. In that study the researchers used visual speech lexical primes and auditory lexical targets. They found that participants identified words presented auditorily in noise better when they were preceded by a video clip of the same lexical prime. This effect was also maintained across different speakers. The researchers therefore concluded that neither visual nor auditory speech perception is modality-specific or instance-specific. These results were similar to those of Dodd and colleagues (1988) who used a semantic categorisation task with VO prime and auditory targets. They also argued that visual prime and auditory targets activate the same amodal representation. One question that follows about amodal speech representation is whether the same is true of non-native speech?

In the studies by Sanchez-García and colleagues (Sánchez-García, Enns, & Soto-Faraco, 2013) an AV matching task was used to test Spanish and English speakers in their native language. In the follow-up study, Sanchez-García et al (2013) also included a group of Spanish, English and German native speakers tested in a non-native language. In both studies, the matching task, a VO condition (sentences recorded by a Spanish-English bilingual male and played silently) presented, followed by an AV fragment. Participants were asked to make a match/mismatch judgment. There were several conditions: visual context could be continuous with visual modality in the AV target fragment (i.e. intra-modally continuous), continuous with the auditory modality in the AV target fragment (i.e. cross-modally continuous), or with neither of the two modalities in the target fragment (i.e. discontinuous).
For example, during intra-modal continuous trials, the visual context continued into the visual modality of the target but not the auditory modality. The results showed that if participants were given a visual-only context followed by a congruent native AV stimulus, responses were faster i.e. they showed an AV benefit effect. However, this did not sustain for non-native stimuli. The researchers therefore argued that the visual attributes of speech are redundant for anticipatory planning when the visual attributes do not exist in the listener’s native language. These results are yet to be replicated with stimuli at the syllable and word level, as well with other languages.

The current study addresses shortcomings from the previous studies of Sanchez-García and colleagues (2013), but also combines cross-modal repetition priming design with the use of a printed prime. The reason for integration of a printed prime becomes apparent when reviewing the studies which demonstrated the usefulness of this paradigm. Therefore, the experiments that have used a printed prime in their cross-modal matching paradigm will be reviewed below.

1.3.5 Orthographic primes

There is a vast body of data showing that reading impacts on spoken word recognition (Okano, Grainger, & Holcomb, 2016). Studies show that participants take longer to recognise spoken words if stimuli do not map onto orthography (Ziegler & Ferrand, 1998). Many experiments also show that spoken word lexical decision is affected by orthographic features, e.g., proximity (Grainger, Muneaux, Farioli, & Ziegler, 2005; Ziegler, Muneaux, & Grainger, 2003). The study by Lansing (1995) used a prime (a printed word which was semantically related or unrelated) followed by a target (word presented in VO mode). Their results confirmed that semantically related primes facilitated speech reading of targets. However, for unrelated prime-target pairings, sensory factors such as word visibility (number of visually distinct phonemes in a word) were more important. The researchers concluded that presenting a written prime provides an abstract model of a word and facilitates lip-reading by matching the observed articulation with the abstract model. They argued that when prime and target do not match, then lip reading is more difficult and thus hindered. It is not yet known if these mechanisms emerge with non-native speech and this issue will be addressed here. One relevant question here is what type of printed prime to use when assessing non-native speech processing in the early stages of foreign language contact? Cognates are possible candidates.
1.3.6 Cognate effects

Cognates (if they exist in a given pair of languages) are often the first words to be acquired in the early stages of second language acquisition (Desmet & Duyck, 2007). Cognates can be defined as words that share form and meaning across languages. For example, the English word cigarette and the Russian word “sigareta” (White, 2015). Cognates are more typical within a family of languages e.g. Spanish–Italian or Dutch–English, including identical written words that are pronounced differently in the two languages, e.g. English–French. Quasi-cognates can be phonologically and semantically similar yet written differently. Cognates can also be found in languages from different families that use different scripts, for example English–Hebrew (Gollan, Forster, & Frost, 1997) and English–Korean (Kim & Davis, 2003). They can be present with different alphabetical scripts, e.g. Russian and English although some Cyrillic and Roman letters look similar because of shared origins in the Greek alphabet (Elgort, 2013). Cognates have an effect on word recognition (Lemhofer & Dijkstra, 2004) and translations, as demonstrated by Boada et al (2013). Cognates have an effect on priming (Lalor & Kirsner, 2001) using Latin and Cyrillic alphabets (Sherkina-Lieber, 2004). Written cognates can have both facilitatory and inhibitory effects on word recognition, depending on the composition of stimuli. Furthermore, false cognates (also called “false friends”) are words that share their form across languages but have different semantic meaning. These false friends also show a similar pattern of facilitation or inhibition depending on the stimulus list composition. For instance, when a lexical decision task was given to Dutch-English bilinguals with a list of cognates, false friends, and control words, their recognition time showed a facilitation effect for cognates. However, when cognates were excluded from the list and Dutch words were added, there were strong inhibition effects for false friends (van Hell & Dijkstra, 2002). Similar inhibition results for false cognates were found by Van Heuven and colleagues (van Heuven & Dijkstra, 2010) in a functional magnetic resonance imaging (fMRI) study.

A Russian-English cross linguistic study of cognates was conducted by Temnikova and Nagel (2015). Russian-speaking participants who had been studying English for a mean duration of 11 years, were presented with Russian primes (related or unrelated) and English targets (cognates versus non-cognates). Using a lexical decision task, participants had to judge if a letter string was an English word or not. They found no facilitation from written cognates. On the contrary, they found that cognates significantly slowed identification times. This effect was independent of the Russian speakers’ English language proficiency. They advocated further studies into the inhibitory cognate effect.
Midgley, Holcomb and Grainger (2011) argued that cognate effects may not be observable in participants with lower levels of proficiency using behavioural studies and they utilised EEG measures instead. They assumed that any cognate advantage observed when processing written words is due to activation of pre-existing form-meaning associations in the native language (L1). They presented native English speakers who were at various stages of French L2 acquisition with words (cognates and non-cognates) in English and French. At the same time, they recorded event-related potentials (ERP) such as N400 that reflects lexical and semantic processing. Increased amplitude in the N400 component is associated with increased processing difficulty (Lotze, Tune, Schlesewsky, & Bornkessel-Schlesewsky, 2011). Cognates produced different patterns to non-cognates in N400 suggesting they were more difficult to process. Importantly, this effect was present in native language processing. The results also showed that the English cognates showed early differences in N400 when compared to non-cognates. In contrast, non-English items showed later differences between cognates and non-cognates. The researchers concluded that different timing of N400 components reflects a different mechanism compared to the L2 cognate effect. Furthermore, they argued that cognates have a special status for L2 language learners, especially during early L2 acquisition. The authors also reported a reversed cognate effect in posterior sites at around 300 msec post-stimulus onset during L2 processing. They argued one reason for a reversed cognate effect is conflict in mapping from orthography to phonology because the same orthographic pattern maps onto two distinct pronunciations. Furthermore, they hypothesised that L1 pronunciation should be more dominant compared to L2. A different explanation of the reversed cognate effect is possible code-switching from L2 to L1 generating a switch cost.

Similar conclusions were derived from a study by Bice and Kroll (2015). They recorded ERP whilst native English speakers who were in the early stages of learning Spanish performed a lexical decision task in English that included cognates. The results showed that even beginning learners demonstrate a reduced N400 for cognates compared to non-cognates and this effect is stronger in intermediate learners.

Kouačević (2012) presented Croatian-English speakers with cognate and non-cognate words in a lexical decision task. Their results showed that reaction times to cognates were slower than non-cognates. There was no cognate effect on response latency to English words. They interpreted the results in terms of inhibition of the cognate effect as a result of the participants performing a language decision task prior to performing the lexical decision task. They also argued that the less proficient language exerts an inhibitory effect on the native language, leading to inhibition.
In summary, written cognates influence lexical access in the early stages of non-native speech acquisition. This permits examination of speech processing models and assumed connections with the semantic level of speech processing. To the best of our knowledge no study has yet examined VO and AV processing of cognate words in the early stages of L2 exposure. Thus, the current study uses English-Russian cognates in the experimental design. As cognates are presumed to access the semantic level of word processing, this manipulation will also allow an examination of theoretical models of speech perception. These theories will be reviewed below, with specific emphasis on whether models incorporate VO and AV aspects of native and non-native speech processing.

1.4 Theories of Speech Perception

Psychoacoustic theories of speech perception share the idea that speech is best explained by the psychophysics of sounds, rather than language-specific processes. For instance, the theory of acoustic invariance assumes that all phonetic features contain an invariant acoustic pattern. This pattern constrains speech perception (Blumstein & Stevens, 1979). However, such theories cannot account for the influence of higher-order linguistic context and visual cues on speech processing. It is also unlikely that acoustic properties remain invariant across different speakers, phonetic contexts, languages, or even an individual lifespan (Stathopoulos, Huber, & Sussman, 2011).

Pattern Recognition Theories do not assume the integration of visual and auditory information either. For example, the Auditory Dominance Model (De Gelder & Vroomen, 1992; Sekiyama & Tohkura, 1991) assumes that visual speech is a secondary constraint, and only occurs when auditory speech is not intelligible. Intelligibility is not clearly defined in this model, though. Moreover, all-or-none auditory identification contrasts with the idea of intelligibility being a continuous measure.

Additive Speech Perception models (Massaro, 1987; Massaro et al., 1993) do assume that the sources of information are added together at the integration stage. Pre-labeling theory (Braida, 1991) holds that information from multiple sources is integrated prior to labelling. According to this model, a stimulus in a given modality is located in a multidimensional space, with a response center corresponding to a prototype. In bimodal presentation, the multidimensional space is the combination of space presentations. A response depends on the response center (prototype) closest to the location of the stimulus.
All the aforementioned models treat two sources of audio-visual information as independent of one another at the initial evaluation stage. One exception is the TRACE model of speech perception (McClelland & Elman, 1986). The TRACE model is based on the principles of interactive activation models and assumes that speech processing occurs via excitatory and inhibitory interactions among a set of units. These units are divided according to three levels: feature, phoneme, and word. Together with bottom-up activation from features to phonemes, and from phonemes to words, there is top-down activation. In addition, activation of some units at one particular level can inhibit other units at the same level. De Gelder and Vroomen (De Gelder & Vroomen, 1992) added visual input features to auditory input features to account for audio-visual speech perception. This model proposes feed-forward but also a feedback processing. It also assumes that features in one modality can activate features in the other modality. None of the models make specific predictions about processing of visual aspects of non-native speech or AV speech perception in L2. The Native Language Magnet (NLM) model of Kuhl and colleagues (Grieser & Kuhl, 1989; P. K. Kuhl, 1991) by contrast does addresses some of these issues, albeit not directly in the audio-visual domain.

1.4.1 Native Language Magnet (NLM) Model

Kuhl and colleagues (Grieser & Kuhl, 1989; P. K. Kuhl, 1991) define phonetic categories of a native language as prototypes that function as perceptual magnets for sounds, assimilating new sounds according to a prototype match. This results in a warping of the psychoacoustic space around prototypes so that the acoustic space accommodates and strengthens prototypes and accommodates poor examples of a phonetic category.

NLM theory assumes that infants have an innate ability to perceive differences between the sounds of phonetic categories and to discriminate stimuli according to these categories. These processes reflect auditory processing mechanisms that are based on natural psychophysical boundaries. However, by 6 months of age magnet effects begin to emerge as a consequence of exposure to a specific linguistic environment. Language-specific experience allows infants to derive language specific category representations (magnets) guided by the limits of a unique boundary. According to NLM theory, categories attract sounds that are distinguished by a single magnet and magnet effects precede and underlie changes in perception of non-native language features. NLM theory accounts for category effects observed in L2 adult learners because it assumes that native-language magnets distort the perceptual space, resulting in
attraction towards L1 “similar” sounds. NLM theory therefore makes assumptions about how speech in a foreign-language is identified according to acoustic proximity to a native magnet.

Evidence for NLM theory is mixed. In their original study, Grieser and Kuhl (1989) examined development of native phonetic prototypes in 6 month-old infants. The researchers focused on the perception of the synthetic English vowel /i/. Drawing from average production results obtained by Peterson and Barney (Peterson & Barney, 1951), they derived the best and worse prototypes of the vowel /i/ and then designed a synthetic stimuli set for each one which deviated from the reference points in equal steps (e.g., 30, 60, 90, and 120 mels). Their design included a discrimination paradigm consisting of a stimulus presented either with itself or with a different stimulus. In order to adapt this design to infants, the researchers trained infants to turn their head when two stimuli were different. The results showed that infants had more difficulty discriminating vowels when they were presented with a prototype stimuli set compared to a non-prototype set. Moreover, discrimination was worse when stimuli were close to a reference point, such as when deviating by 30 and 60-mels, and this effect was more pronounced for prototype stimuli than for non-prototype stimuli. The researchers also correlated performance with a goodness of stimuli rating derived from adults, and showed that infants perform worse with exemplars that were rated as good prototypes of the vowel /i/ compared to those that were rated as poor.

Kuhl (1991) extended her work by comparing performance of three groups: adults, 6-months babies, and Rhesus monkeys. The results replicated the findings of Grieser and Kuhl (1989) with both adults and 6-months babies with no effect for monkeys. The researchers concluded that the observed effect in humans is due to their experience with language rather than being just an auditory phenomenon.

To investigate the impact of language exposure on native speech development, Kuhl and colleagues (P. K. Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992) conducted a cross-cultural study with American and Swedish 6-month old infants using a head turning paradigm. In addition to English prototype and non-prototype stimuli sets, the corresponding Swedish stimuli sets were incorporated. The results replicated previous findings, showing poor discrimination around referents and worse performance for prototype than non-prototype stimuli, but the effect was maintained for the native language only and not for non-native language stimuli. Kuhl et al. (1992) concluded that by 6 months of age infants have developed prototypes for native speech, despite not fully developing other aspects of speech perception by that age.
The results of the above studies lead to the theoretical argument that phonetic prototypes exert an effect on surrounding stimuli and serve as a perceptual magnet. The magnet is assumed to warp psychoacoustic space around a prototype so that the space is "shrunk" around prototypes and "stretched" around non-prototypes or poor exemplars of a phonetic category. However, this argument was undermined by the criticism that no identification component was included in the experimental design of the aforementioned studies. This meant that although stimuli were rated for their “goodness” by adults, raters may not have perceived them as belonging to the same phonetic category of /i/. To address this criticism, subsequent experiments used a stimuli identification task in addition to a rating of goodness task. Iverson and Kuhl (Iverson & Kuhl, 1995) found that only 55% of raters identified non-prototypes as /i/ and chose a stimulus 75 mels to the left of the original prototype as the best exemplar of the vowel /i/. Even with a correction, the study replicated previous findings by showing that an unbiased measure of sensitivity (d’) was worse for the prototype (the one selected by adults) than for the non-prototype in infants.

Sussman and Lauckner-Morano (Sussman & Lauckner-Morano, 1995) also included an identification task in addition to the discrimination task. Their participants labeled non-prototype of /i/ as belonging to this vowel category only 20% of the time. As a result, the researchers used non-prototypes that were used as poor exemplars of the vowel /i/ category at least 50% of the time. They also altered the paradigm whereby the stimuli varied from reference points in 15 mels instead of 30 mels and used five reference points compared to two (prototype and non-prototype), with the three addition stimuli being a 75-up mel non-prototype with increasing F1 and decreasing F2 values, a 75-down mel non-prototype with decreasing F1 and increasing F1, and 67-mel non-prototype with decreasing F1 and increasing F2 values. The results showed that the prototype group had worse performance, especially for the 30 and 45-mel stimuli compared to 75-down group had the best performance.

Lively and Pisoni (Lively & Pisoni, 1997) reported that the perceptual magnet effect is weaker in adults than infants, especially for vowels that are more distal to referents. In three experiments using the original Kuhl’s (1991) synthetic tokens of /i/ they divided raters into two groups whereby one group rated prototype-centered set of tokens for goodness while the other group rated the non-prototype set. Moreover, half of each group were presented with the written word "peep" so that they could compare a token to a good example of /i/ in the word, while the other half of raters were asked to imagine the word "peep" without seeing it in print. The results replicated the previous studies of phonetic prototypes by showing that some tokens were judged better than others. However, although the study used exactly the same
stimuli and procedure as Kuhl (1991), they did not replicate the magnet effect. Instead, Lively and Pisoni (1997) focused on individual differences among ratings and found that ratings of prototype and non-prototype sets were not consistent across participants. To investigate whether goodness rating change across context, Lively and Pisoni derived individualised prototype and non-prototypes based on a participant's own goodness rating and then synthesised a set of tokens centered around them. Each participant was then presented with the same-different discrimination task used in Kuhl (1991) and their own individualised stimuli set. The results failed to detect any magnet effects with either stimuli set. In addition, the results showed that goodness rating varies with acoustic context. The researchers compared their findings with Kuhl's (1991) and argued that some members of the /i/ category were reject by their raters. They confirmed their hypothesis by showing that non-prototype tokens were categorised as /i/ in only 54% of the trials. They criticised the validity of the magnet effect in the same-different discrimination paradigm from Kuhl’s original study (Kuhl, 1991) as raters may have made comparisons between categories rather than comparing tokens within the single /i/ category.

Frieda and colleagues (Frieda, Walley, Flege, & Sloane, 1999) examined individual differences in selecting a prototype, and the impact of these individual differences on the magnet effect. Participants were selected if they spoke different dialects of American English. In the first experiment, the /i/ category was identified for each participant individually by asking them to identify a stimulus as either /i/ or not, obtaining an outer limit of the category. Each participant was then presented with a grid on a computer screen, with each box representing a different sound. The participant then moved around the grid using the mouse to locate the best prototype of /i/. Each participant had multiple trials and only participants who consistently chose the same prototype were selected for the second experiment. Consistent with the variability assumption, there were individual differences in perception of /i/ category limits. Moreover, 16 out of 24 participants chose different prototypes, highlighting differences in the localisation of a prototype. In the second experiment, researchers tested the perceptual magnet effect directly. Importantly, they included a foreign vowel (Swedish vowel /y/) condition and hypothesised that it would function as a non-prototype. The results showed that at least half of the participants preferred the same stimulus as that chosen early as a prototype. Moreover, discrimination was worse in the prototype condition compared to the non-prototype condition, consistent with the previous findings. However, they failed to find any evidence of warping of perceptual space around prototypes and non-prototypes respectively.

Other studies also failed to show warping of perceptual space around prototypes. Thyer and Hickson (Thyer, Hickson, & Dodd, 2000) presented native speaking Australian English
participants with five synthetic Australian vowel categories and synthesised a prototype and a non-prototype stimuli set for each vowel. In the first experiment, participants rated each token for goodness. The results were consistent with previous studies showing some internal structure of a phonetic category, with the prototypes and their closest variants receiving the highest rating. However, there was no evidence that ratings declined with distance from the prototype in perceptual space. There were also some inconsistencies in rating responses. In the second experiment the participants identified each token as belonging to one phonetic category. The results showed that for three out of five categories, identification accuracy of prototypes was good (97%) and decreased around the category boundary as expected. There was also a significant correlation between the participants' rating and their identification accuracy. In the last experiment, a small subset of prototype and non-prototype stimuli was presented for identification. There were also two groups of participants: a naïve group and a phonetically trained group. The researchers found no difference in identification of vowel categories between trained and untrained participants. However, participants identified a smaller set of stimuli differently to the full set by accepting poorer quality tokens as category members. When the participants performed same/different judgments comparing each of five vowel stimuli to the prototype and to the non-prototype, there was no evidence of warping of perceptual space in any of the five vowel categories.

Further questions about the theory were raised in the study of Iverson and Kuhl (Iverson & Kuhl, 2000) investigated whether phoneme boundary and perceptual magnet effects are identical. In the first experiment, the participants identified stimuli as /i/ or /e/ and rated goodness. The results replicated Iverson and Kuhl’s (Iverson & Kuhl, 1995) results, showing that participants were consistent in their categorisation of the /i/ category but their consistency for the /e/ category was reduced. However, there was a relationship between location of the best /i/ exemplar and the highest goodness rating for each participant. In the second experiment, the researchers divided their participants into groups and varied context and stimulus interval: 30-mel roving, 30-mel fixed, 60-mel roving, and 60-mel fixed. The participants performed same/different task. Results showed an inverse relationship between participants' sensitivity and their goodness rating, suggesting that the perceptual magnet effect was present regardless of the context and stimulus interval. In order to investigate any phoneme boundary effect, the researchers examined sensitivity peaks near the phoneme boundaries and correlation between sensitivity and the degree to which the stimulus within each pair is discriminated differently.

In sum, results with native speech processing suggest there is possible application of NLM theory to non-native speech processing. However, the theory has been criticized for a focus
around a single L1 category (Reid et al., 2015). NLM theory makes few specific predictions in relation to how visual attributes of speech are processed in non-native speech and is therefore limited in its utility at least prima facie. We will therefore turn to a model which is capable of making such specific predictions.

1.4.2 The Perceptual Assimilation Model (PAM)

The Perceptual Assimilation Model (PAM) (Best, 1995) is based on a realist approach to perception. The model assumes that articulatory gestures are perceptual primitives for perceiving native and non-native speech input. Articulatory gestures can be perceived without accessing knowledge of the vocal tract, acoustic features or mental representations including prototypes. Critically, PAM assumes that the patterns of perceptual assimilation for native and non-native speech segments will be shaped by the experience and familiarity with speech input. The model makes specific predictions about the possible patterns of assimilation as follow:

1. Two-category (TC), whereby non-native segments are assimilated to a different native category, leading to good discrimination.
2. Category-Goodness (CG), whereby non-native sound contrasts are assimilated to a native category but differ in degree of difference from a form. Discrimination depends on category goodness differences to this form.
3. Single-Category (SC), whereby non-native sounds are assimilated to the same native category and equally differ from the native form. The pattern of discrimination is likely to be poor but could be just above chance level.
4. Both uncategorisable (UU Type), whereby both non-native sounds are outside of native category but still within the phonetic space, and thus may be discriminated as poor or fair, depending on their location in the phonological space.
5. Uncategorised versus Categorised (UC Type) whereby one non-native sound is assimilated into a native category whilst the other falls in a phonetic space outside, thus resulting in very good discrimination.
6. Nonassimilable (NA Type) whereby both non-native categories fall outside phonetic space and are heard as non-speech, resulting in acceptable discrimination.

Hazan and colleagues (Hazan et al., 2006) tested PAM by assessing whether visual information facilitates auditory training of non-native speech. They focused on English phonetic contrasts that are difficult for Spanish speakers and then identified patterns of
assimilation, specifically assimilation of English voiced and voiceless fricatives to Spanish voiceless phoneme (/z/ → /s/, /d/ → /t/), and assimilating both English voiced and voiceless stops into Spanish voiceless counterparts. They asked whether visual cues affect perception of assimilation patterns and use of visual information differs between native speakers and Spanish learners of English. A total of 36 participants were included in the study, with 28 of the sample spending less than 2 months in an English speaking country. In addition, 12 native speakers of English also took part in the experiment. The participants were requested to identify syllables presented in mild background noise under AO and AV conditions. The results showed that the AO condition improved consonant identification in native English speakers and Spanish learners of English, and both groups obtained most improvement in the vowel-consonant (VC) context as opposed to VCV, whereby the VC context was the most difficult to identify in AO conditions. Despite enhanced effects of AV presentation in both groups, Spanish learners of English continued to show the same patterns of assimilation in the AV condition as in the AO condition and visual cues did not improve performance. The researchers argued that Spanish participants ignored visual cues associated with non-native contrasts that are phonemic in English but allophonic in Spanish because they are not represented in the native language. Since the study’s focus was on the very specific contrasts, it is hard to make generalisation to other contrasts and languages.

PAM predicts that unfamiliar category sounds are more likely acquired as a new category by non-native speakers (Best & Tyler, 2007). Faris, Best and Tyler (2016) refined the concept of unfamiliar category (UU) to 3 types: focalized (similar to a single L1 category), clustered (similar to more than one L1 category), and dispersed (not similar to any L1 categories). Faris et al. presented Egyptian Arabic speakers residing in Egypt with spoken Australian English vowels and asked them to rate them. They specifically chose listeners with minimal exposure to English. Participants were presented with a grid containing CVC or CV keywords in Arabic, representing Arabic vowels and diphthongs transcribed as vowel-glide sequence and asked to match an unfamiliar vowel with a similar vowel from the grid. They found only 3 out of 19 Australian English vowels were consistently assimilated to Egyptian Arabic categories. The researchers argued that non-native assimilations reflect the detection of phonetic information in non-native phones that are phonologically meaningful in the native language consistent with the assumptions of PAM. Faris and colleagues did not investigate audio-visual speech, though.

In contrast, Reid et al. (2015) applied a cross-language category assimilation paradigm to the AV domain by asking participants to classify instances of L2 into L1 phonemic categories (specifically lexical tones) and rating them for goodness of fit. Thai, Cantonese, and Mandarin
speakers (all tonal languages) were asked to map Thai tones onto their own native lexical tone categories, and English speakers (non-tonal language) were asked to do the same, but with either native intonation categories (prosodic e.g. statement or question) or using symbols as a response. Stimuli were presented in AO, AV and VO conditions using a discrimination task. In the category assimilation task participants heard the same native Thai female pronouncing a syllable with one of five Thai tones and were instructed to match the character/intonation type/symbol. The results showed that, for all groups, VO presentation resulted in poorer performance and there was no AV benefit. Reid at al. (2015) speculated that visual correlates of consonants and vowels may be stronger than the visual correlates for tones and suggested the PAM model should be modified to include non-segmental and non-auditory features.

Other studies tested PAM with AV speech perception, such as perception of English fricatives (Wang, Behne, & Jiang, 2009). They contrasted the perception of fricatives with a visually distinct place of articulation (labiodental, interdental alveolar) in Mandarin Chinese and Korean learners of English compared with native English speakers. The groups were selected because interdental fricatives are absent and prevocalic voiceless labiodentals and alveolars are present in Mandarin whereas neither labiodental nor interdental fricatives are present but voiceless alveolars are absent in Korean. Guided by the assumptions of PAM, the researchers expected Korean participants to perform better than Mandarin speaking participants because the front visual attributes are more crowded for Mandarin in the perceptual space around interdental fricatives. Results showed that for the labiodentals (which do not exist in Korean), Korean participants made more errors in the visual domain compared to Mandarin or English controls but fewer errors when stimuli were presented in AO and AV conditions. For the difficult interdentals, which do not occur in either Mandarin or Korean, non-native groups performed worse than English controls but showed an AV benefit effect. Importantly, the study showed differences between the non-native groups, with the Mandarin group showing worse performance when identifying interdentals in the auditory and AV conditions compared to Koreans. Furthermore, Mandarin speakers showed greater AV-fusion with the incongruent AV items compared with their Korean counterparts. Wang and colleagues concluded that whilst non-native listeners do utilize visual attributes of non-native speech, they differ from native speakers in their weighing of visual and auditory attributes, depending on whether they exist in their native language and concluded that the acquisition of non-native visual attributes parallels that of auditory non-native speech perception insomuch as new non-native categories need to be established by learning associations between non-native attributes and the corresponding non-native sounds. One weakness of the study was that the experimental groups differed in their length of stay in an English speaking country and their level of proficiency was not controlled. Clearly, PAM can account for differential effects of language
on the AV benefit effect in non-native speech processing and therefore it is able to generate a set of hypotheses prima facie that motivates the work to be reported in this thesis.

1.4.3 The Speech Learning Model

The Speech Learning Model (SLM) (Flege et al., 1997) was developed to explain why non-native listeners have difficulty pronouncing non-native sounds. Accents become stronger with age of second language acquisition. Flege hypothesised that non-native production errors are caused partially by perceptual errors. SLM assumes that processes used to learn the sounds of a native language, including category formation, are applied naturally to non-native listening. Language-specific speech sounds can be isolated in representations called phonetic categories. Phonetic categories established in childhood for native speech sounds evolve to reflect the properties of all phonemes learned and speakers of two or more languages maintain contrasts between native and non-native categories, which exist in a common phonological space that reflects exposure to any language learned.

A number of hypotheses can be derived from SLM. For example, sounds in native and non-native languages are related perceptually at a position-sensitive allophonic level, rather than at a more abstract phonemic level. Studies showing that certain allophones are better produced by non-native speakers than other allophones (Aoyama et al., 2004; Sheldon & Strange, 1982) support this hypothesis. A new phonetic category can be established for a non-native sound that differs phonetically from the closest native or non-native sound, if a non-native language speaker discerns phonetic differences between native and non-native speech.

Studies showing that non-native listeners perceive phonetic differences between languages support this prediction. The greater the perceived phonetic dissimilarity between a non-native sound and the closest native sound, the more likely phonetic differences between sounds are identified. Studies of native Japanese speakers of the Japanese /r/ compared to English /r/ and /l/ support this hypothesis (Sekiyama, 1997). Given that identification of phonetic differences between native and non-native sounds decreases as age of learning (AoL) increases, Flege (1987) has reported that bilingual speakers produce stops with similar VOT values in the native and non-native languages. The model assumes that native and non-native categories exist in a common phonological space and dispersed to maintain contrast.

Other evidence for the model comes from cross-linguistic studies of consonant perception e.g. in Cantonese (Chan, 2012) and Malay (Pilus, 2002). However, a few studies do not support
the model. For instance, in one such a study, the American English speakers had to learn German vowels. The results were not compatible with the Flege’ model because participants discriminated some instances of the same contrast between German vowels more easily that others (Kingston, 2003). Importantly, the model makes no prediction about non-native AV speech perception and is more concerned with the process of non-native speech acquisition in the auditory only domain. Therefore, we will now turn to the Motor Theory which can account for the audio-visual speech better than the SLM model.

1.4.4 Motor Theory

Motor speech theory assumes that speech production and perception depends on a specialized module located in the motor areas of the brain. In the revised form (Liberman & Mattingly, 1989) and in contrast to PAM, the theory assumes that listeners perceive intended gestures rather than the movement of articulators. Intended gestures are called perceptual primitives. The theory also assumes that speech perception necessarily involves the speech motor system - the so called motor theory of speech perception.

Empirical evidence for Motor Theory stems from a number of studies (Galantucci, Fowler, & Turvey, 2006). For instance, in the original study of Liberman and colleagues (Franklin S. Cooper, Alvin M. Liberman, & John M. Borst, 1951) results showed that two-formant synthetic syllables /di/ /du/ that are acoustically dissimilar due to co-articulation of the vowel are still perceived as the same /d/ consonant. Furthermore, different gestures resulted in the same acoustic structure due to co-articulation when voiceless stop consonants are cued by stop bursts. Liberman and colleagues concluded that perception therefore tracks articulation and that listeners parse acoustic signals along gestural lines, specifically the F0 contour caused by the voiceless consonant results in perception of voicelessness whilst the contour caused by vowel height contributes to perception of vowel height. Furthermore, motor theory accounts for McGurk type effects that are observed when the haptic feel of consonant gestures is substituted for facial movement (Fowler & Dekle, 1991).

Support for the Motor Theory comes from brain imaging studies using PET, fMRI and TMS (transcranial magnetic stimulation). PET studies show that when participants whispered nonsense syllables without hearing them (masked by white noise), the auditory cortex was still activated (Paus, Perry, Zatorre, Worsley, & Evans, 1996). Calvert and colleagues (Gemma A. Calvert, 1997) also found that the same auditory cortical areas are activated when silent speech or speech-like movements are viewed as those implicated in normal speech.
Transcranial magnetic stimulation of the motor cortex reveals that listening to lingual consonants is associated with enhanced muscle activity in the tongue (Fadiga, Roy, Fazio, & Craighero, 2009). Visual aspects of speech e.g. positioning of the jaw, shaping of the lips, motion of the cheeks are distributed over a broad area of the face, which is in turn “over” represented in the primary motor and sensory cortices (Vatikiotis-Bateson, Munhall, Kasahara, Garcia, & Yehia, 1996; Vatikiotis-Bateson & Yehia, 1996; Yehia, Rubin, & Vatikiotis-Bateson, 1997, 1998). Facial movements that result from the motion of the vocal tract are successfully extracted for speech perception, and 91% of the total variance observed in facial motion can be determined from vocal tract motion (Yehia et al., 1998). These studies support the idea that perception tracks articulation, as suggested by the Motor Theory.

However, the Motor Theory is not uncontested. Massaro and Light (Massaro & Light, 2004) investigated whether viewing articulatory organs facilitates the identification and production of difficult non-native phonetic contrasts. They developed Baldi, a computer-animated talking head and two training conditions: with the view of articulatory organs and without them. In the condition with the view of articulatory organs, four different head position were presented in a fixed order: a view from the back inwards, a side view of mouth only (static or dynamic), a transparent side view of whole face, and a transparent front view. In the condition without the view of articulatory organs, Baldi’s head was presented as non-transparent with the same sequence of viewing conditions. Each of 11 Japanese speakers completed intensive 3-week training with phonemes /l/ and /r/ under two conditions. Results revealed that although both viewing conditions produced some improvement in the perception and production of difficult English contrasts, there was no significant benefit for the condition that showed internal articulatory organs.

Criticism of Motor Theory persists. One limitation is that it cannot account for situations when multiple gestures correspond to a single sound (Schwartz, Basirat, Ménard, & Sato, 2012). Again, Motor Theory suffers from an inability to explain the influence of high-order linguistic context on speech perception. In addition, it fails to account for perception of speech in mutism, e.g. (Fourcin, 1968) and chimpanzees which understand speech but cannot produce it (Lyn et al.). A central assumption of Motor Theory is that the motor system is recruited during the perception of speech. However, critics argue that speech perception occurs independently of motor actions. Such criticisms preceded the development of the Fuzzy Logical Model of Perception (Massaro et al., 1998) which accounts for AV speech integration in non-native speech better than Motor Theory.
1.4.5 The Fuzzy Logical Model of Perception (FLMP)

The Fuzzy Logical Model of Perception (Massaro et al., 1998) assumes that prototypes and their associated features are the basis for speech perception. Information is evaluated about the degree to which the feature in the speech signal matches the prototype and the parameters of ideal values are fluid and represented as fuzzy values. The FLMP model assumes three stages of sensory processing: feature evaluation, feature integration and pattern classification and each stage is processed in relation to the degree of match between input and an ideal prototype. Fuzzy-truth values index the degree of match and provide a representation of matching between inputs. The model makes the critical prediction that speech will not be integrated with visible input if concurrent sensory features are contradictory. According to the FLMP, bimodal speech perception may be determined by visual and auditory sources of input and the contribution of one input modality will decrease if redundant information is afforded by another modality. The model predicts that audible speech cannot be integrated with visible speech if features are represented in different terms. Fuzzy-truth values can be used as a representation of the degree of a match. The model makes the additional prediction that informative features have a greater impact, and consequently, one feature may have a greater effect when a second feature is ambiguous. The model assumes that auditory and visual prototypes are independent, and the value of one source cannot change the value of the other source. For the purposes of the present study, FLMP is not specific to AO speech as it assumes that perceptual units from any modality can be summarised as ideal prototypes.

Support for the model comes from a variety of sources. Masaro and colleagues (Massaro et al., 1993) compared native English speaking to Japanese, Spanish, and Dutch bilinguals with variable English proficiency. They expected to see cross-linguistic differences in syllable recognition because of differences in phonotactics (possible sequencing of sounds) and phoneme inventories. For example, although all four languages have /b/ and /d/ segments, Japanese does not have the phoneme /th/, or /v/, American Spanish does not have /v/, and Dutch does not have /th/. Whereas English speakers frequently perceive /va/ and /tha/ when auditory and visible speech is varied along the /ba/ to /da/ continuum, other language speakers should perceive this continuum differently. Specifically, there should be differences in syllable recognition. English speakers frequently perceive /bda/ when visible /ba/ paired with an auditory /da/. This perception is supported by occurrence of /bd/ cluster at syllable, morpheme, and word boundaries in English. However, initial consonant clusters do not occur in the Kansai dialect of Japanese while consonant clusters are less likely to occur across word boundaries in Spanish. The cluster /bd/ does occur in Dutch but it is pronounced /pd/ or /pt/. Therefore, the researchers reasoned that bilinguals will be less likely to respond with /bda/
given a visible /ba/ and an auditory /da/. They also expected no differences between groups if participants are limited to only /ba/ and /da/ responses as all four languages contain them. The results showed no significant differences among groups, and replicated previous findings in terms of effects on both audible and visible speech. The same experiment was replicated with the full spectrum of stimuli and a free response task for three groups of participants (Smeele et al., 1998). The results showed that Japanese speakers tend to respond with /ba/ and /da/ but frequently gave /wa/ and /za/ as their response. The Dutch participants mostly responded with /ba/, /da/, /va/, /za/, /va/, /vha/, and /ma/. Although different responses were given, bilinguals in all groups were influenced by visible speech, and the contribution of visual aspects was larger when a source was ambiguous. Therefore, FLMP provided a reasonable description of the data.

FLMP can also account for cross-linguistic differences in segmental processing at a sentence level (Cutler, Weber, & Otake, 2006). FLMP integration of audible and visible speech has been demonstrated for all visibly distinctive consonants in English (Massaro et al., 1993) and vowels (Smeele et al., 1998). Massaro and colleagues (Massaro et al., 1998) reported differences in the AV processing of English consonants and vowels i.e. auditory information is less available for consonants, especially stop consonants than for vowels. Consonants and vowels also differ in visible aspects, with vowels having slower articulartory gestures and less specific articulatory position than stop consonants. The results suggest that visible speech has a larger impact on the perception of consonants than on vowels.

Massaro (Massaro, 1987) demonstrated that AV speech perception is not predicted by one source of information (auditory or visual). Participants were unable to filter out one modality of input and selectively focus attention on the other. Instead, response time and accuracy of speech identification reflected integration of visual and auditory input. Furthermore, participants have difficulty ignoring one source of information in AV speech. Massaro proposed that the integration of auditory and visual information in visible speech is involuntarily and universal across languages. The latter prediction was tested by Chen and Massaro (2004) who compared the perception of syllables by Mandarin speaking and English speaking participants using synthetic and auditory continua from /ba/ to /da/. They found that syllable identification was dependent on both visual and auditory input and proposed that the FLMP model could account for the results.

However, some evidence suggests that AV speech integration is not automatic and can be attenuated. For example, the magnitude of the McGurk effect is reduced if task demands increase in other perceptual channels (A. Alsius, Navarra, Campbell, & Soto-Faraco, 2005;
Tiippana, Andersen, & Sams, 2004). Notably, FLMP is not often tested with natural and ecologically valid L2 speech, as opposed to synthetic stimuli. Schwab and colleagues (Schwab, Nusbaum, & Pisoni, 1985) highlighted that synthetic speech is processed differently to natural speech. They showed that several sentences were required before listeners adapted to synthetically generated speech. In response to these issues a different theory called the Analysis-by-Synthesis model was developed.

1.4.6 Analysis-by-synthesis model.

The analysis-by-synthesis model was proposed by Halle and Stevens (Halle & Stevens, 1962). Based on this model, the salience of visual input as well as redundancy of visual and auditory inputs contributes to the final AV percept. The model, which has been applied to different perceptual domains, not just language, assumes that there are internal representations of the world that guide perception. Furthermore, the model assumes that incoming sensory stimuli are not just fed-forward but are constrained early by these internal representations. Thus, during initial sensory processing stages, errors between the sensory input and the internal predictions are recognised. In a number of combined behavioural and EEG-recording experiments, van Wassenhove and colleagues (van Wassenhove, Grant, & Poeppel, 2007) used congruent and McGurk incongruent stimuli. They asked participants to choose between [ka], [pa], and [ta] stimuli presented in auditory, visual, and AV conditions. They recorded two measures of auditory event-related potentials (ERPs)- N1 and P2. The results showed that visual speech differentially modulates early stages of auditory processing (50–100 ms). The results were interpreted as evidence of early integration. Processing time in auditory speech was impacted upon by how ambiguous visual speech was. For instance, /p/ is more visually salient than /k/, resulting in larger latency shifts in N1 and P2. They also showed that AV speech resulted in reduced auditory evoked potentials when compared to auditory speech alone. That is, they did not find supra-additive enhancement for AV speech. Instead they interpreted their results as evidence for supporting the notion of predictive coding in the context of the analysis-by-synthesis model. Building on the basis of this model, van Wassenhove and colleagues proposed that because visual speech input precedes auditory signals, it allows for on-line prediction or constraint of the incoming auditory input. Therefore, a given visual stimulus initiates a number of possible visemic representations and provides content against which the auditory stimulus is mapped. The strength of prediction derived from abstract visemic representations is dependent on the saliency of VO input. The researchers suggested that the analysis-by-synthesis component can form the basis for abstract representations of speech. This model includes the idea that distinctive features form the basis
of abstract representations. Van Wassenhove and colleagues acknowledged that attention may regulate how much weight is put on visually initiated predictors, possibly creating an attentional biasing effect and argued that temporal facilitation was evident regardless of the degree of visual saliency and that this implies that visual-based prediction is proposed to dominate the auditory input in the evaluation process. Furthermore, they proposed a deactivation mechanism in which stimulation of one modality inhibits the non-stimulated modality, thus providing a way to limit redundancy of cross-modal stimuli (Laurienti et al., 2003). Extending this idea, von Wassenhove and colleagues hypothesised that the internal predictions derived from visual stimuli narrow possibilities for information about place of articulation and this information is mapped into incoming visual stimuli. Therefore, deactivation of auditory cortices seen when visual input precedes auditory input, may reflect that the auditory cortical neuronal population had extracted information only in the exact and relevant frequency range. Finally, based on their results, the researchers proposed that there are two stages of AV speech integration: the first is the stage in which visual stimulus allows the prediction of the auditory input; and second stage is a perceptual unit stage, which is not dependent on specific features but rather is independent of modality.

1.4.7 Visual Place Articulatory Manner (VPAM) Theory

One of the most influential theories of AV speech perception was proposed by Summerfield (Summerfield, 1979) who suggested four alternative models. In the first two models, modality-specific and feature-based explanations are used for AV integration, while the latter two are gesture-based models. The first model assumes that AV integration occurs through the process of extracting discrete phonetic features, in which information about place of articulation is provided by the visual modality. At the same time, information about manner of articulation is provided by the auditory signal. This model is often referred to as Visual Place Articulatory Manner (VPAM), assumes that auditory and visual signals are translated into common symbolic codes prior to integration. The second model is that vectors describing the values of independent acoustical and optical parameters are integrated. The assumption is that exemplars of auditory and visual spectral information are compared and matched against incoming speech signals. The other models are gesture based models and related to the Motor Theory (Liberman & Mattingly, 1989). The third model is that cognitive representations of AV speech have a filter - the vocal tract. Hypothetical vocal tract configurations are compared to incoming stimuli. The fourth model is that the vocal tract provides articulatory dynamics obtained from auditory and visual channels combined together. Therefore, the underlying
assumption is cognitive processes of AV speech perception are independent of any particular sensory modality, or amodal, and occur in the early stages of speech perception.

Support for the gestural-based model of AV integration comes from a study of cross-modal effects on speech recognition. Rosenblum et al. (L. D. Rosenblum et al., 2007) conducted an experiment in which visual-only speech from a single talker was played for an hour, followed by an auditory-only condition. The results showed that participants were better at identifying speech in noise when it was the same talker as the one used in the visual-only condition. Rosenblum and colleagues hypothesised that the benefit in the AO condition was due to the participants being primed by the information extracted about the specific talker from the visual-only condition. Other experiments using cross-matching tasks by Lachs and Pisoni (L. Lachs & Pisoni, 2004) showed similar results.

Further possible support for the gesture-based model of AV speech integration comes from imaging studies. Specifically, functional magnetic resonance imaging (fMRI) studies which used silent speech-reading, showed activation in the cortical areas commonly associated with auditory speech (G. A. Calvert & Campbell, 2003). Correspondingly, some studies showed that AV speech activates motor circuits in the cortex (Hall, Fussell, & Summerfield, 2005; Skipper, Nusbaum, & Small, 2005). In addition, there are studies that show activation in multisensory brain regions (Bernstein & Liebenthal, 2014). Such studies show that congruent auditory and visual speech stimuli elicit super-additive activation (AV > A + V) whilst incongruent stimuli produces sub-additive activation (AV < AO + VO) or response suppression (AV < AO or VO) in some studies (Gemma A. Calvert, 1997; G. A. Calvert & Campbell, 2003). These studies lend support to a view that multisensory neurons in multisensory brain regions are activated because of amodal gestural information provided by the auditory and visual stimuli (L. D. Rosenblum et al., 2007).

In contrast to gesture-based model of AV speech integration, modality-specific processing accounts of AV speech integration have gained support. Studies show that unisensory coding operates not only prior to integration but remains operational during integration. There are studies that point to integration of modality-specific features rather than articulatory gestures. For example, Dodd and colleagues (Dodd et al., 1988) reported a study of children with phonological disorders who were given McGurk stimuli in AO, VO and AV conditions. The auditory and visual stimuli were presented congruently or incongruently and children were asked to point to a picture corresponding to an auditory stimulus with correct responses on AV trials needing perception of a McGurk fusion. The results showed that children were more likely to respond to the auditory component than VO or fused McGurk’s stimuli. Dodd et al.
reasoned that if knowledge about articulatory gestures were the only necessary information for AV integration, then a population of children with disordered phonology should not have been biased toward the auditory component of the stimuli. Further, support for modality-specific AV integration comes from a population of profoundly deaf children with cochlear implants. Bergeson et al. (Bergeson, Pisoni, & Davis, 2005) showed that children who receive cochlear implants early in life show enhanced auditory-only skills in contrast to children who receive implants later and show improved lip-reading skills. They reasoned that children’s ability to perceive stimuli in each sensory modality is affected by experience of that modality and the information is not amodal.

In sum, Summerfield’s theory of AV speech perception can account for AV benefits and other constraints. However, it does not address the perception of non-native speech. Another theory, The Reverse Hierarchy Theory (RHT), fills in the gap with the debate about the presence and size of an AV benefit effect and therefore is reviewed below.

1.4.8 The Reverse Hierarchy Theory (RHT)

The Reverse Hierarchy Theory (RHT) of perception (Ahissar, Nahum, Nelken, & Hochstein, 2009; Hochstein & Ahissar, 2002) assumes that perception occurs in a top-down manner. The theory holds that a novel perceptual task is performed on the basis of immediate high-level representations unless mapping of low-level to high-level representation is absent. RHT postulates that in such situations, re-mapping of lower-level input representations to a higher representational level occurs before a perceptual task is complete. Such remapping is referred to as “perception with scrutiny” and backward search that may result in dissimilar lower-level input representations being remapped to a common high-level representation or vice versa.

Bernstein and colleagues (Bernstein, Auer, Eberhardt, & Jiang, 2013) applied RHT to AV speech perception by assuming speech remapping depends on the natural correlation between auditory and visual input. Top-down processing from visual speech processing can lead to correlated auditory features generating a remapped speech transformation. Backward search can take the form of a cross-modal search, with perceptual distinctions of one modality guiding scrutiny of representation in another modality. They suggested that AV speech carries more information and therefore attention can be directed away from a novel auditory signal and furthermore this possibly impedes uni-sensory (AO) perceptual learning.

Bernstein and colleagues proposed two hypotheses: (1) Stimuli presented to a trainee’s primary perceptual pathway will impede learning via a lower-priority pathway; (2) Stimuli
presented to a trainee’s lower priority pathway will promote learning by a higher-priority pathway. These hypotheses predict a cost from an incorrect backward search because the path is highly activated compared to the less developed path. In relation to L2 learning, Bernstein and colleagues suggest it is possible to provide input through a lower priority pathway (e.g. somatosensory) for training sounds in a new language. Bernstein and colleagues (2013) offer some support for these predictions. They asked whether AV training provides benefit or hinders auditory-only perceptual learning of degraded speech. They trained English speakers to learn paired associations between vocoded spoken nonsense words and pictures after AV or AO-training. Results were that AV-trained participants learned significantly more than AO-trained participants and learning generalised to novel untrained words. Interestingly however, when the researchers alternated AV and AO conditions on a list-by-list basis (mixed training), they observed that AO stimuli were more effective than AV stimuli. These findings perhaps highlight a role for selective attention in language learning as assumed by Majerus et al. (Majerus, Poncelet, Van der Linden, & Weekes, 2008).

In follow-up studies, Bernstein and colleagues (2013) reported that concurrent AV speech impeded AO perceptual learning of paired associates in prelingually deaf adults with a cochlear implant. Eberhardt and colleagues (2014) trained participants to perform a paired associate task in visual-only (VO) mode and found they were able to perform the task with high accuracy. In the second experiment, they repeated AV training with the addition of synchronous vocoded acoustic speech. Results revealed that under this condition, VO learning was impeded. In the next experiment they demonstrated that vocoded AO stimuli were less informative than VO stimuli. They conducted another experiment with vibrotactile stimuli, which promoted VO perceptual learning. Finally, they carried out their final experiment with non-training controls and showed that training with VO stimuli generalised to new consonant stimuli but not to sentences.

Reverse Hierarchy Theory has also been applied to the perception of non-native speech. The study by Bidelman and Dexter (2015) examined the “cocktail party” phenomenon commonly exhibited by bilinguals whereby they show decreased L2 comprehension when they are in the environments with increased speech-in-noise (SIN) ratio. The examiners recorded mismatch potentials in native and non-native listeners while they listened to English contrasts in noise. Results showed that late non-native speakers benefited from lower SIN and that cortical responses were delayed in non-native speakers. They argued that bilinguals recruit top-down processing to compensate for degraded speech and hypothesised that speaking more than one language shapes brain function beneficially but that bilingualism might also impede speech perception. However, RHT has not been applied to perception of non-native speech in the
very early stages of encountering a new language. Furthermore, to the best of our knowledge no studies have applied RHT to the AV perception of non-native speech in naïve participants.

1.5 Summary of research questions

We reviewed the evidence for perceptual narrowing occurring in early periods of life and that the dominant linguistic environment continues to influence how a second-language is perceived. We reviewed studies which demonstrate that certain non-native contrasts are difficult to perceive for adult speakers, resulting in auditory perceptual assimilation into native categories. However, only a handful of studies considered the question of whether visual cues also constraint second-language perception or whether certain visual aspects of non-native speech are perceived as ‘illegal’ or unusual. While several studies looked at the ability to discriminate languages based on the visual input (lip-reading), the factor of language familiarity was not controlled whereas the current study used a relatively unfamiliar language, Russian. Furthermore, none of the previous studies specifically investigated whether non-native visual aspects of speech are assimilated into native visual speech. This gap is addressed in the current study. In addition, we reviewed the debate in the literate regarding the AV benefit effect, including its presence, size, and language-specificity. The current study aims to extend the investigation of the AV benefit effect into the domain of naïve participants who had no prior knowledge of Russian. Further issues of interest highlighted in the literature review were speech synchrony and the resulting ‘visual anticipatory’ effect, which is influenced by whether stimuli are presented in a native or a non-native language. The current study is interested in investigating whether naïve participants are able to detect a particular pattern of asynchrony inherited in the unknown Russian language. Another related issue emerged from the literature review, namely rate-dependent speech perception. The current study joins the debate regarding which speed is most beneficial to perception of L2, whether the presence of audio-visual speech benefits or hinders processing of rate-modified speech, as well as whether processing of time-compressed speech is language-dependent.

Further questions of interest which emerged from the literature review were theoretical models which can account for the processing of visual aspects of non-native speech in naïve participants. The questions were addressed by using the aforementioned cross-modal repetition priming paradigm and the use of an orthographic prime in the form of cognate words. Several theories, such as The Perceptual Assimilation Model (Best, 1995) and The Reverse Hierarchy Theory (Ahissar et al., 2009) were examined in their ability to account for the results and answer the question of mental representation of non-native speech.
Finally, the review of literature showed that there are a handful of studies that examine non-native AV speech processing at multiple levels – phonemes, syllables, words, and sentences. The current study fills in this gap by using multiple levels, which allows for tapping into different levels of processing – phonological, morphological, sub-lexical, lexical, and semantic.

Prior to deriving specific predictions for the study, it is important to have an overview of the non-native language used in the current study, Russian. Therefore, we now turn to a review of the levels of language processing, parameters of sound and articulation, and description of sound organs, followed by a background to the Russian language.
Chapter 2: Background to language, articulation, and Russian Language

The current thesis is concerned with multiple levels of language processing. Therefore, it is useful to briefly describe 5 levels of language processing: phonological, morphological, syntactic, semantic, and pragmatic (Baker & Hengeveld, 2012). It is worth noting that the semantic and pragmatic levels were not directly addressed in the study.

2.1 Levels of language processing

2.1.1 Phonological level

Phones are the basic sound units of languages whilst phonemes are specific sound and written units of a given language. Phonemes may consist of two or more phones or allophones. Depending on a phonetic context, phonemes can be realised by different allophones. For instance, the p in spot, port, spoon, and pat produce phonetically different acoustic pattern but all represent the same phoneme. Morphemes and words are comprised by a number of combinations of phonemes. In relation to the current inquiry, the phonological level refers to the structure and systematic patterns of phonology of the English and Russian languages.

2.1.2 Morphological Level

Phonemes are combined into units called morphemes, which in turn comprise words. For instance, “unbreakable” includes three morphemes: un-; -break-; -able. Therefore, the morphological level refers to these basic units of words and the rules that govern their formation. At this level of analysis, parts of words and structure of words, such as stems, root words, prefixes are analysed.

2.1.3 Syntactic level

The syntactic level in the area of linguistics refers to the rules and order of words in a sentence/phrase. These syntactic rules are different across languages and mainly refer to grammar and logic of a given language. At the syntactic level of analysis, the meaning that is derived from the change of word order, the addition/subtraction of words is the main point of inquiry.
2.1.4 Semantic Level

The semantic level refers to the analysis of the meaning of utterances, words, phrases, and sentences. This level is also concerned with the relationship between words not only within sentences but also between sentences. The main focus is on how the meaning is conveyed or changed through semantic rules.

2.1.5 Pragmatic Level

The pragmatic level is concerned with the context in which utterances occur. This level is focused on what meaning is intended and how it is interpreted. Pragmatics focuses on the social meaning behind the words, the inferred intent of the speaker, and the ability to understand the intended meaning, namely pragmatic competence.

The current study is concerned with multi-modal processing of language and it is useful to briefly define some of the basic terms referred to in the thesis, including parameters of sounds and organs of articulations (Hixon, Weismer, & Hoit, 2014).

2.2 Parameters of Sounds

There are three parameters of sounds: voicing, place of articulation, and manner of articulation.

*Voicing* refers to whether the vocal cords vibrate (voiced) or do not (voiceless). For example, /v/ is a voiced consonant while /f/ is a voiceless consonant.

*The place of articulation* refers to where sounds are made in the mouth. These include bilabial (sounds made with the lips), labio-dental (with lips and teeth), lingua-dental (the tongue tip touches the inside of central incisors), alveolar (sounds made with the tongue touching the alveolar ridge), palatal (behind alveolar), velar (soft palate is in front of the uvula), and glottal (most posterior place of articulation).
The manner of articulation is the way in which a sound is formed in relation to constriction of the mouth. These include plosives (vocal tract is blocked), fricatives (airstream through a narrow constriction), nasals (closed vocal tract and open nasal cavity), affricates (combination of a stop and a fricative), liquids (vowels produced with an open vocal tract), glides (a gliding motion from one vowel to another made by articulators).

Vowels are formed by the continuous flow of the air or vibration of the vocal folds with the airstream escaping from the mouth in an unobstructed way.

Vowels are not as easily defined in relation to their place and manner of articulation compared to consonants because of the unobstructed way they are produced. Furthermore, diphthongs are formed when one vowel glides into another without interruption.

The following features determine production of vowels: position of the lower jaw (high, mid, low), tongue site (front, central or back), place of lips (rounded or unrounded), vocal tract tension (lax or tense), and duration of the vowel articulation.

2.3 Speech Organs

The mandible, or lower jaw or jawbone is the lowest bone of the face. It holds the lower teeth in place. The mandible bone impacts on speech by modifying the site and shape of the tongue. This in turn, changes the volume of the oral cavity. The mandible can perform raising, lowering, protrusion, retraction, and oblique lateral movements. The raising is especially involved in production of front vowels.

Another speech organ, the teeth are used in orchestrated fashion with lips and tongue and instrumental in production of many sounds, such as dental and labio-dental.

Lips are instrumental in speech articulation. There are multiple actions that lips perform, including closing, raising, lowering, rounding, protruding, retracting, raising, and lowering the angles of the mouth.

The tongue is a very mobile organ which assists in the production of multiple sounds. There are two parts to the tongue: oral and pharyngeal. The oral part is instrumental in speech as it moves freely. The top of the tongue is influential in modifying places of articulation and is divided into the tip/apex, the blade, and the middle. The tongue is capable of producing multiple movements, including horizontal forward-backward movement (e.g., low back vowels such as [a]); upward-downward movement (e.g., vowels as [i]); forward-backward
movement (e.g., [l] and [r]); concave-convex cross-sectional movements (e.g., [s]); degree of central grooving (e.g., articulation of [s] or [y]); spread-tapered surface plan configuration (e.g., [t] or [s]).

2.4 Russian Language

This section introduces the Russian language used in the current study as an unfamiliar language for many English speakers. An overview of the phonetic system of the Russian language including specific vowels and consonants as well as unique accent and stress will be described and compared with English. The Chapter will conclude with a brief overview of comparative studies of Russian and English.

2.4.1 Background

Russian is classified as an East Slavic language and is an official language in Russia, Belarus, Kazakhstan, Kyrgyzstan and many minor and unrecognised territories. It is an unofficial but widely-spoken language in Ukraine, Latvia, Estonia, and to a lesser extent, the countries that were constituent republics of the former Soviet Union and the Eastern Bloc. It is the most geographically-widespread language of Eurasia and the most widely spoken of the Slavic languages. It is also the largest native language in Europe, with 144 million speakers in Russia, Ukraine and Belarus. Russian is the eighth most spoken language in the world by number of native speakers and the seventh by total number of speakers. Russian is one of the six official languages of the United Nations (Challenges of peace operations : into the 21st century : executive summary in the languages of the United Nations, 2002).

2.4.2 Phonetic System of the Russian Language

There are 33 letters in the Russian alphabet, which is also called Cyrillic, and these are graphemic depictions of 42 sounds including 10 vowels and 21 consonants with two letters that are neither vowels no consonants, defined as a hard sign and a soft sign. The Russian orthography is transparent for spelling and is considered to be phonetic because there is mostly a one-to-one correspondence between letters and sounds. The phonetic rules of intonation and stress that is unique to Russian present difficulties for non-Russian speakers learning the language. Below is an overview of the phonetic system of the Russian language, with a particular focus on the vowels and consonants used in the current study.
2.4.2.1 *Russian Vowels*

Unlike English, vowels in Russian have no diphthongs and are not divided into long and short (e.g. [i] and [i:] in *live* and *leave*). There are six vowel sounds in Russian (a, o, e, u, i, ɨ) but there are an unusual number of allophones. There are four members of phoneme ɨ, two members of e, four members of phoneme a, two members of phoneme o and two members of phoneme u. Use of allophones depends on phoneme position in a word, the nature of adjoining consonants and the degree of stress. In unstressed positions, all Russian vowels are reduced to a certain degree. The vowels ɨ and i are not reduced but the vowel a is reduced if unstressed or in the syllable immediately preceding the stress and before the stress location. Furthermore, in all other syllables a is pronounced weakly, thus creating a second level of vowel reduction. As a rule, the further a vowel falls from the point of maximum stress, the weaker is the vowel.

The most difficult Russian vowel for English speakers to pronounce is ɨ. The closest English sound is i, but this is different to the Russian vowel. To produce the Russian sound ɨ, English speakers have to place their tongue in the position between the sounds [i] and [u] and then stretch their mouth as for a smile. It is worth noting that the letter ɨ almost never starts a word in Russian, except for a few words borrowed from foreign languages.

The Russian i sound is similar to the English sound ee (*eg. seed*) in the pronunciation of those who do not use a diphthong for the English sound. The Russian ɨ is a pure vowel and must be distinguished from the common diphthongal pronunciation of the English ee, and as such also occurs in Russian. The tendency for native English speakers to diphthongise ɨ is particularly noticeable in final positions (Hamilton, 1980).

The Russian vowel ə is an advanced back vowel, close to the vowel in such English words as *cup, tuck* and presents no difficulty for English speakers. The sound ə never occurs in stressed syllables and is never preceded by a soft consonant in the same word. It occurs only before the stress (Hamilton, 1980).

2.4.2.2 *Russian Consonants*

There are 21 consonant letters in the Russian orthography, although linguists differ in their classification of 7 consonants (Hamilton, 1980). These consonants correspond to 36 sounds in writing because the Russian alphabet does not represent soft consonants with separate...
consonant letters. Therefore, differences between palatalised and non-palatalised consonants in Russian are not reflected orthographically, e.g., /bj/ and /b/ both correspond to the same grapheme. Russian consonant phonemes do not show the same allophonic variation as English consonants. Russian consonant allophones differ from each other slightly and are formed naturally if surrounding sounds are pronounced correctly. Unlike English, Russian consonants have no aspiration. Russian consonants are subdivided into ‘voiced-unvoiced’ and soft (palatalised) and hard (non-palatalised) and the softness of consonants has semantic value. The contrast of soft (palatalized) versus hard (non-palatalized) consonants is one of the most important features of Russian phonology. Soft consonants are produced with the tongue high in the mouth, near the palate while the hard consonants are produced with the tongue typically held lower than an English production of the same consonant sound.

The Russian sound s is one of eleven fricative Russian phonemes and is dental. In English, voiced fricatives have only partial voicing while in Russian the vocal cords are made to vibrate throughout the whole duration of the sound. The English s is formed similarly to the principle allophones of the Russian s, but the English s is normally apical and not laminal as in Russian. Moreover, in the English s, the blade of the tongue is raised slightly towards the hard palate, creating a somewhat higher pitch compared to the Russian s (Hamilton, 1980).

The Russian sound м is a member of 4 Russian nasal consonants (м, н, and мъ, нъ) and is a bilabial consonant. The principle allophone of Russian м presents no difficulty to English-speaking learners, being identical with the normal English м in man (Hamilton, 1980).

The Russian language has two rolled consonants. The р allophone is a voiced alveolar rolled consonant and the Russian allophone рь is a voiced palatalised (soft) alveolar rolled consonant. These are vibrant and dental. The relationship of рь to р is the same as that of s to z, or f to v. The majority of English speakers are unable to pronounce a rolled р without special practice and experience great difficulty holding a continuous rolled р or producing it in certain sound combinations. Since the р-sound occurs in English only before vowels and is not heard fully finally or before consonants, English speakers tend to omit the rolled р altogether in Russian words, or may insert a vowel when р occurs before a consonant (Hamilton, 1980).

2.4.2.3 Syllable structure

English has mainly single-syllable words but Russian words are primarily two to three syllable strings. It is not unusual for Russian words to have as many as 8 syllables and a
sequence of two, three, and four consonants is common in Russian. Within 3-phoneme sequences, a lateral, a trill or a labiodental fricative is commonly heard. Many multiple sequences involve consonants with more elaborate articulations compared to English (Hamilton, 1980).

2.4.2.4 Accent and stress

As in English, word-accent in Russian is ‘free’; that is, it does not fall on the same syllable in every word. The Russian word-accent is also free to move from syllable to syllable. The accent in Russian is a stress-accent and is formed by increasing the breath stream. It is possible to distinguish different degrees of stress within a word in Russian. Some syllables have secondary stress, such as a re-tonic syllable (i.e. syllables immediately in front of the stress) and are more strongly stressed in Russian than other unstressed syllables (Hamilton, 1980).

2.4.2.5 Length

Consonant length plays a more important role than vowel-length in Russian when compared to English. In some cases, consonant-length serves to distinguish one word from another word, but vowel-length never does. Since double consonants also occur in English writing, English speakers generally have no problem distinguishing between single and double consonants in spoken Russian. However, they are often unable to produce double r and r because no such sounds occur in English (Hamilton, 1980).

2.4.2.6 Intonation

Non-specific questions in Russian, which can be answered ‘yes’, ‘no’ or ‘maybe’ differ in intonation from English. In Russian there is a sudden rise followed by a sharp fall. This contour is very distinct from English. Tentatives or requests in which the verb is in the imperative in English have a rising pitch at the end of the sentence. This is very different from the melodic contour of the Russian tentatives (Hamilton, 1980).

In summary, Russian has a more complex consonant inventory and a smaller and simpler phonemic vowel inventory than English. On average, words in Russian have more syllables, variable stress patterns and complex consonant structure making them more challenging to articulate than English words. Such differences may be expected to influence phonological
acquisition of Russian as an L2 by native English speakers if such influences are observed. By extension, the hypothesis of AV benefits on non-native speech processing predicts that these are the speech features in Russian that are most likely to elicit any effects if they are observed.

2.4.3 Comparative studies of English and Russian languages

To the best of our knowledge, no study has examined cross-linguistic AV speech perception of Russian in naïve native English speakers. This represents a gap considering that the study of Russian language is becoming popular in several countries (including Asia and Australia). The majority of research on this topic falls in the realm of auditory speech perception only and the relevant literature is reviewed below.

Gilinchinskaya and Strange (Gilinchinskaya & Strange, 2010) tested PAM (Best, 1995) by presenting native Russian speakers with English vowels. They adopted a perceptual assimilation task wherein English vowels were presented to Russian listeners who were asked to rate the vowels’ perceived goodness on a 9-point Likert scale. The prediction of PAM that similarity of acoustic patterns between English and Russian will have an effect on assimilation was upheld for all but one English vowel [ɛ]. Unfortunately, the study did not include a contrast with English native speakers presented with Russian vowels. Nonetheless, the results motivate the set of experiments reported here testing the effect of language overlap on the perception of Russian vowels by native English speakers. According to PAM, overlap in vowel structure between the English and Russian languages should produce an effect of similarity on speech perception. The question remains open regarding AV speech and whether similar processes operate for visual aspects of speech.

Banzina and colleagues (Banzina, Dilley, & Hewitt, 2016) were interested in the perception of vowels produced by Russian-speakers of English. They first measured vowels, produced by the Russian-speakers in terms of F0, intensity, duration, F1 and F2. They demonstrated that the most significant factors were in the temporal and vowel quality dimensions whilst differences in intensity and fundamental frequency were less obvious. In fact, Russian-produced secondary-stressed (SS) and unstressed–unreduced (UU) syllables were almost half the duration compared to English. In addition, vowel quality in Russian-spoken UU and SS syllables show they were articulated with the jaw significantly less open and tongue higher in the mouth when compared to native speakers. Banzina et al. (2016) conducted a cross-modal phonological priming experiment, whereby English-speaking participants were presented with an auditory prime followed by a printed English word (target) asked whether the target word
was a real English word or not. The auditory primes consisted of native words, Russian-spoken words, Russian-spoken words with artificially “corrected” syllables, or native unrelated words. The results showed that over-reduced UU vowels triggered no priming effect, while artificially corrected stimuli produced priming levels similar to native stimuli. Russian-spoken words with SS syllables led to some priming effect, albeit reduced. They concluded that accuracy of UU syllables appear to be crucial for speech identification but SS syllable accuracy was less crucial. Given the differences between the two languages, described in the study, we would predict that English speakers may detect these differences via the visual and AV modes of presentation.

Evidence of cross-linguistic effects between English and Russian speech perception comes from bilingual participants (Gildersleeve-Neumann & Wright, 2010). Russian-dominant bilingual children residing in North America produced Russian-influenced phones in English speech (palatalized consonants and trills) and significantly higher rates of trill substitution, final devoicing and vowel errors. Shafiro and Kharkhurin (Shafiro & Kharkhurin, 2009) also included two groups of Russian-English bilinguals with variable English-language experience and an English-speaking control group. The results showed that Russian-English bilingual speakers have difficulty identifying English vowel contrasts that do not exist in Russian. For late-bilinguals, auditory identification was correlated with reduced visual word recognition of the difficult contrast.

Studies of AV speech synthesis in Russian are only restricted to the artificial speech. Specifically, a number of studies by Karpov and colleagues (Karpov et al., 2013), using an AV speech synthesis system for the Russian language tested the effects of speech context on modelling asynchrony between auditory and visual speech modalities. They conducted a corpus-based study and synchronised AV speech using a talking head. The rationale for their studies was the fact that languages have different degrees of time asynchrony between phonemes and visemes in free speech. They argued that time discrepancies between auditory and visual speech units are different for the English and Russian languages and predicted different asynchronies between speech and visual cues for each language with English characterised by a greater asynchrony than Russian. Their results suggest asynchrony between articulation and sound may be an AV cue for English-speaking listeners who are processing Russian speech. It is therefore expected that native English speakers will be sensitive to such information during the AV processing of Russian natural stimuli.
In sum, we can expect considerable differences between speech perception of English and Russian stimuli at different levels of processing—phonological, pre-lexical, lexical, prosodic, rhythmic, and so on. Furthermore, there is some indication from comparative studies of Russian and English (e.g. Karpov, 2013) that it is possible to descend the differences in asynchrony of speech production between the two languages. As a result, we can derive a set of hypotheses for the current study.
Chapter 3: Rationale for the current study

3.1 Visual perceptual assimilation

The theoretical underpinning of the current study is that native language exerts a powerful effect on the processing of non-native speech and this effect is well documented in the auditory L2 speech domain. The current study extends findings to the visual and audio-visual aspects of non-native speech processing for a language not previously investigated (Russian). The main hypothesis is that visual cues from non-native speech in Russian will be perceived by English-speaking adults according to the similarity of cues with the English language, as predicted by theories, such as PAM (Best, 1995) and NLM (P. K. Kuhl, 1991). The first prediction is that Russian stimuli that are similar to English stimuli will be assimilated according to the native English categories/prototypes. In addition, the study aims to replicate findings obtained with other languages by showing that differences between the language pair –Russian-English can be detected when processing L2 speech, including visual only speech. The hypothesis will be tested using a variety of experimental paradigms, including Language Identification, Same-Different Paradigm, Cross-Modal Match/Mismatch, and Real versus Non-Real (backwards) paradigm.

3.1.1 Native vs Non-native effect

The underlying assumption is that native stimuli (English) will be processed better (more accurately and faster) than non-native stimuli (Russian). This assumption is consistent with the majority of literature in the field of both auditory and AV speech processing.

3.1.2 AV benefit effect

There is a debate in the literature about the reliability of the AV benefit effect across languages and therefore a call to replicate cross-linguistic effects and AV speech. Therefore, the aim of this study is to test the hypothesis that the AV mode adds benefit to L2 speech perception beyond auditory speech alone. Within this premise, it is predicted that the AV benefit will be most pronounced for Russian stimuli that are most dissimilar from English. Consistent with prior findings, it is also predicted that auditory-only speech would be superior to visual-only speech.
3.1.3 Mental representation of non-native speech

An additional aspect of the study is to tap into the mental representation of native and non-native speech processing by evoking a prototype in one perceptual domain, e.g., auditory, to test whether it facilitates matching/discrimination in another perceptual domain, e.g., visual. The notion of prototypes is proposed by theories such as NLM (P. K. Kuhl, 1991) and FLMP (Massaro, 1987).

1. The prediction is that native prototypes will facilitate discrimination of L2 stimuli, as predicted by aforementioned theories.
2. Mapping from auditory to visual will be easier than mapping from visual to auditory, as predicted by RHT (Hochstein & Ahissar, 2002).
3. Cross-modal language continuous input will be processed more efficiently than cross-modal language discontinuous input, as predicted by RHT (Ahissar & Hochstein, 2004).
4. A written native cue will prime L2 speech perception and result in better identification/discrimination of L2 stimuli than when no cue is provided, as predicted by a vast body of data showing that reading impacts on spoken word recognition (e.g. Okano et al, 2016).
5. Written cognates (words that share meaning and form across languages) will have a facilitatory effect on processing of visual only non-native speech, similar to that demonstrated in the auditory domain (e.g. (Desmet & Duyck, 2007).

3.1.4 Level of Input

6. The hypothesis is that the longer the stimulus, the easier it will be to identify/discriminate it. Specifically, it is expected that sentences will be easier to process than words whilst the latter will be easier to process than syllables and vowels. This hypothesis was partially based on the results of Soto-Faraco (Soto-Faraco et al., 2007) and Ronquest et al (Ronquest et al., 2010) for words and sentences which showed easier discrimination of sentences compared to words. The current hypothesis extends the finding by proposing that syllables and vowels processing will be less efficient compared to words and sentences. The underlying assumption is that the differences reflect various levels of processing -phonological, morphological, lexical, etc. and that more cumulative information is available at the higher levels (e.g., RHT of Hochstein & Ahissar, 2002).
3.1.5 Temporal aspects of L2 processing

Other hypotheses to be tested are that dynamic aspects of speech facilitate processing of L2.

7. The first hypothesis is that the differences in asynchrony between auditory and visual speech modalities inherent in each language will be recognised by non-native speakers. The latter hypothesis aims to extend the finding from perception of synthesised Russian speech (e.g., (Karpov et al., 2013)) to ecologically valid natural Russian speech.

8. Furthermore, the study aims to investigate temporal correlates of L2 processing by manipulating speed of stimuli presentation. There is a debate in the literature whether slow presentation facilitates processing of non-native speech in the auditory domain. The working prediction is that slower presentation will facilitate processing of non-native speech in the AV and visual domains compared to playing stimuli quickly. The RHT (Hochstein & Ahissar, 2002) predicts perception in the top-down manner by default but that low-level properties are accessed first in non-native speech and then mapped backwards in a top-down manner.

3.2 Mapping of experiments to hypotheses

3.2.1 Visual perceptual assimilation

The hypothesis of visual perceptual assimilation is tested in experiments using a variety of experimental paradigms, including Language Identification, Same-Different Paradigm, Cross-Modal Match/Mismatch, and Real versus Non-Real (backwards) paradigm. In Experiments 1,2,4, and 5, visual perceptual assimilation is examined at the level of vowels and syllables. In Experiments 6-8, visual perceptual assimilation is examined at the level of words and sentences.

3.2.2 Native vs Non-native effect

The underlying assumption that native stimuli (English) will be processed better (more accurately and faster) than non-native stimuli (Russian) is examined through all 8 experiments.
3.2.3 AV benefit effect

The AV benefit effect is examined in all 8 experiments by including different modalities (e.g. AO, VO, AV). In Experiment 3 only AV and AO conditions are included. The particular emphasis is on comparing AV and AO performance for the Russian dissimilar items. Similarly, item rating paradigm conducted in Experiment 4, allowed for the comparisons of stimuli, especially non-native items, between AV and OA modalities.

Mental representation of non-native speech

9. Experiments 3 and 5 examine whether mapping from auditory to visual is easier than mapping from visual to auditory. The experiments include cross-modal language continuous input contrasted with cross-modal language discontinuous input. The effect of a written native cue is examined in Experiment 6 by comparing performance with a written cue with that without a cue. Experiment 6 uses written cognates to demonstrate a facilitatory effect of a written cue on processing of VO non-native speech. Item rating paradigm conducted in Experiment 4 in the AV and AO modalities also sheds lights on mental representation of non-native speech.

3.2.4 Level of Input

10. The first 5 experiments focus on perception of vowels and syllables whilst experiments 6-8 focus on perception of words and sentences. The underlying assumption is that the longer the input, the easier it is to make an accurate response because of the accumulated amount of information from different levels of processing.

3.2.5 Temporal aspects of L2 processing

11. Experiments 7 and 8 examine asynchrony between auditory and visual speech modalities inherent in each language and aim to investigate temporal correlates of L2 processing by manipulating speed of stimuli presentation.
Chapter 4: Perception of the visual aspects of vowels and syllables in non-native speech.

This Chapter summarises the first five experiments conducted in the study. The experiments described are all concerned with the perception of vowels and syllables in an unfamiliar non-native language in different sensory modalities – auditory only (AO), visual only (VO) and auditory-visual (AV) conditions. Our working assumption is that non-native AV and visual only speech will be perceived using native speech categories leading to the prediction that perceptually dissimilar items should be more salient and therefore easier to identify as non-native than perceptually similar items (cf., Hazan et al., 2005). In contrast to Hazan et al. (2005) however, who tested speakers from different ethnicities using language pairs from different language families, we tested Caucasian speakers using Indo-European languages. The specific question to be tested is whether the visual aspects of non-native speech (Russian) will be assimilated into native language (English) categories, given the overlap in ethnicity and language relative to studies reported by Hazan et al. (2005). The prediction that follows is that the visual aspects of Russian that are dissimilar to English will be identified more readily than visual aspects of Russian that are similar to English. In all experiments, English-speaking participants were presented with unfamiliar vowels and syllables in Russian ranked according to their similarity to English and tested using a speech perception paradigm (cf. Hazan).
4.1 **Experiment 1: Perception of Vowels and Syllables**

In Experiment 1 participants were presented with either English or Russian vowels/or syllables and requested to indicate whether the item is English or not. The stimuli were presented in different modalities - AO, VO, and AV. The key independent variables were error rates and response times. The dependent variables were language (English or Russian), Modality (AO, VO, and AV), and the degree of Similarity to the native prototype/categories for non-native Russian items (Similar, Identical, Dissimilar).

Experiment 1 Hypotheses:

Hypothesis 1 (Primary Hypothesis): Demonstration of Visual Perceptual Assimilation with the new pair of languages, namely English and Russian.

Hypothesis 1 states that in the visual only mode (VO) Russian items that are similar to English will be assimilated into the native categories/prototypes, resulting in high error rates due to false positive response (Yes, it’s English). In contrast, Russian items that are most dissimilar will result in low error rates due to correct rejection (No, it’s not English). Thus Hypothesis 1 predicts 3-way interaction between Modality, Language, and Similarity, with the effect of Similarity only evident for the most dissimilar Russian items. Whilst the primary hypothesis is demonstration of visual perceptual assimilation in VO mode, replication of perceptual assimilation in AO mode is also expected. This will be replicated with the new pair of languages, namely English and Russian.

Hypothesis 2: Modality Hierarchy: AV>AO>VO and AV benefit effect.

Hypothesis 2a: AV benefit effect will show that the overall performance in both English and Russian in the AV mode will be better (more accurate/quicker) than the performance in the AO mode. VO modality will be the least informative of all modalities.

Hypothesis 2b: The AV advantage is predicted to be the most salient for the Russian items that are most dissimilar to English compared to the items that are similar.

4.1.1 Method

*Participants:* Thirty-six undergraduate psychology students (30 females), from The University of Melbourne participated in the experiment in exchange for course credit. The
mean age of participants was 25 years. All were native speakers of English. None of the participants had any prior knowledge of the Russian language. All reported having no known hearing impairment and normal or corrected-to-normal vision.

**Stimuli and design:**

Russian and English stimuli

The male speaker was positioned at a fixed distance (1400 cm) from the camera and recorded against a blank background. He was instructed to keep his head movements stable during the recording and to maintain a neutral facial expression. The whole face of the speaker was presented in each video.

The items were recorded using a Sony digital video camera. The resulting video and audio files were imported onto a PC and a short program was written in the DMDX software (Forster & Forster, 2003) to present them in the experiments. The stimuli were presented on a 350 cm video monitor, in 1024 by 768 resolution, controlled by a Windows PC. The audio stimuli were presented binaurally to participants via a set of Sony headphones.

Stimuli were produced by recording speech of a 35 year-old male bilingual Russian-English speaker, who had learned English at the age of 8 years and had spoken both English and Russian since that time. Items were repeated four times to attain a sensitive measure of responses. For each modality, the total number of item presentations after repetition was 240. Half of the items (120) were English and the other half (120) were Russian. Within each language there were four types of items: 12 single vowels (3 vowels repeated 4 times), 36 vowel-consonant syllables (9 items repeated 4 times), 36 consonant-vowel syllables (9 items repeated 4 times), and 36 consonant-vowel-consonant syllables (9 items repeated 4 times). Items were classified into types according to similarity between English and Russian based on degree of rated sound resemblance between English and Russian (Hamilton, 1980). Dissimilar vowel pairs consisted of the Russian vowel ɨ, (which does not exist in English) and the closest English sound, the vowel æ. Similar items were Russian ɨ and English i:, which are similar syllables in both languages but are not identical (Hamilton, 1980). Identical pairs were made up of Russian a and English æ which are considered nearly identical in the two languages (Hamilton, 1980). In parallel, 3 types of consonants were used: Dissimilar, Similar and Identical. The Russian consonant m is identical to the English consonant m, whereas the Russian consonant s is similar but not identical to the English consonant s, and the Russian consonant r is dissimilar because it does not exist in English (Hamilton, 1980). A full stimuli list is included in the Appendix B. Auditory (A), Visual (V), Audio-Visual (AV) tokens
Visual perceptual assimilation of non-native speech

(instances) of all stimuli were created and recorded in AV mode. Auditory tokens were created using the AV stimuli without corresponding visual input. Visual tokens were created by disconnecting the speakers from the display computer monitor and playing them without the corresponding sound.

Procedure: Each participant was tested under 3 conditions: AO, VO, and AV. In each condition, 120 Russian and 120 English items appeared, with every participant responding to every item in every condition. Half way through each trial a statement appeared to indicate that half of the trial had been completed. A break of 15 mins was provided after each condition. Before the main experiment for each condition, participants were presented with 4 practice trials with feedback.

Fourteen participants received all conditions in a fixed order: AV, AO, and VO. Twenty-two participants received the visual condition first (presenting the AO and AV conditions first was deemed to result in a possible priming of the subsequent visual condition), and half of these had the AV condition next, while the other half had the auditory condition as the second condition.

Participants were tested individually in a sound attenuated room. In the VO and AV conditions, videos of the whole face of a speaker pronouncing an item appeared centred on a computer monitor (sized 16 x19 cm, visual angle 5 degrees) and were presented one by one in a random order. In the AO task, a static picture of a speaker was presented to minimise differences between AO and VO conditions. In each condition, participants read a set of instructions asking them to sort sounds into English and non-English items (please refer to Appendix for full list of instructions). First the word “ready” appeared in the centre of the monitor and was displayed for 472 ms, followed by a stimulus. Participants were required to press the “Yes” key if an item was an English item and the “No” key if it was a non-English item (response keys were counterbalanced across participants). If no response was made during 10s, the next items automatically appeared. Participants were instructed that each item would be presented several times but in random order. In the VO condition they were requested to sort stimuli on the basis of facial information provided by the speaker who was pronouncing the stimuli, without the corresponding sound. To minimise bias, the non-native language was not defined as Russian but referred to only as a foreign language. Before the main experiment for each condition, participants were presented with 4 practice trials. The experiment lasted approximately 45 minutes with breaks of 5 minutes before each set. No feedback was provided after each presentation.
4.1.2 Results

The results were analysed with a mixed ANOVA, with Modality (AV, AO, VO), Language (English and Russian) and Similarity (Similar, Dissimilar, Identical) as within participant variables and order of task presentation (1: AV, AO, VO; 2: VO, AV, AO; 3: VO, AO, AV) as a between participant variable. Response error rates and response times served as the dependent variables in separate analyses. In this analysis, order of task presentation was included as a between participant variable in an initial analysis to assess the potential effect of task order. The analysis for this experiment showed neither main effect of task order, nor any interaction between task order the variables of interests (all $F$ values <1), so analyses reported below were conducted without including this factor. This is the case for all subsequent analyses reported in this thesis, except where indicated otherwise.

Error Analyses:

The error data are reported in Table 1.

Table 1. Mean Percentage Error Rates for Experiment 1 as a Function of Language, Modality, and Similarity for Audio-Visual (AV), Audio-Only (AO), and Video Conditions

<table>
<thead>
<tr>
<th>Similarity</th>
<th>English</th>
<th>Mean</th>
<th>Russian</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AV</td>
<td>AO</td>
<td>Video</td>
<td>AV</td>
</tr>
<tr>
<td>Identical</td>
<td>27.56 (20.97)</td>
<td>23.99 (14.33)</td>
<td>34.81 (18.35)</td>
<td>28.79 (17.88)</td>
</tr>
<tr>
<td>Similar</td>
<td>35.53 (24.02)</td>
<td>28.55 (22.10)</td>
<td>51.27 (20.12)</td>
<td>38.45 (22.08)</td>
</tr>
<tr>
<td>Dissimilar</td>
<td>37.29 (22.14)</td>
<td>34.67 (19.33)</td>
<td>43.76 (16.01)</td>
<td>38.57 (19.16)</td>
</tr>
<tr>
<td>Mean</td>
<td>33.46 (22.38)</td>
<td>29.07 (18.59)</td>
<td>43.28 (22.94)</td>
<td>35.27 (19.71)</td>
</tr>
</tbody>
</table>

Note: Values in parentheses indicate standard deviations.

The Mauchly test of sphericity indicated that the assumption of sphericity had been violated ($p<0.05$). Therefore, the Greenhouse-Geisser corrections were applied.

The analysis showed a significant main effect of Modality $F(1.630, 57.036) = 19.38, p<0.01$, $\eta^2_p = 0.356$.

There was a significant main effect of Language $F(1.35) = 11.46, p=0.002, \eta^2_p = 0.247$. 

82
There was a significant main effect of Similarity \( F(1,364.47,726)=139.02, \ p<.001, \ \eta^2=0.799 \).

Interpretation of the main effects must be considered in the context of a significant three-way interaction, reported below.

The three way interaction between Modality, Language and Similarity was also significant \( F(2,989,104.627)=31.85, \ p<.01, \ \eta^2=0.476 \).

To investigate interaction a further 2 separate two-way ANOVAs were conducted for English and Russian. The first factor Modality had 3 levels (AV, AO, VO), and the second factor Similarity had 3 levels (Similar, Dissimilar, Identical).

The results for the English two-way ANOVA showed that The Mauchly test of sphericity indicated that the assumption of sphericity had been violated for Modality (\( p<0.05 \)). Therefore, the Greenhouse-Geisser corrections were applied for the factor of Modality.

There was a main effect of Modality \( F(1,654, 54.568)=10.453, \ p<.01, \ \eta^2=0.241 \).

There was a main effect of Similarity \( F(2,66)=15.916, \ p<.01, \ \eta^2=0.325 \).

There was a significant interaction between Modality and Similarity \( F(4,8)=96.000, \ p<.01, \ \eta^2=0.170 \).

To investigate this two-way interaction within the separate two-way ANOVA for English, a further one-way ANOVA was conducted for Modality (AV, AO, and VO).

The results showed a significant main effect of Similarity in the AV condition \( F(2,66)=7.237, \ p=0.001, \ \eta^2=0.180 \), in the AO condition \( F(2,66)=8.455, \ p=0.001, \ \eta^2=0.204 \), and the VO condition \( F(2,66)=19.982, \ p<.01, \ \eta^2=0.377 \). Planned paired comparisons showed that for English items in the AV conditions, participants’ error rates were better for Identical compared to Similar items (\( t(35)=3.061, \ p=0.004 \)) whilst no such significant difference was observed in the AU condition (\( p>0.05 \)). For both AV and AU conditions, Identical items resulted in significantly smaller error rates than Dissimilar items (all \( p<.001 \)). In contrast, for English items in the Video condition, Identical items resulted in significantly better error rates than Dissimilar items \( t(35)=3.114, \ p=0.004 \). Furthermore, for the English items in the VO condition, a one-way t-test comparing participants’ performance to the chance level (50%) showed that their performance was at chance for Similar items but better than chance for Identical and Dissimilar items (\( p<0.05 \)).
The results for the Russian two-way ANOVA showed that the Mauchly test of sphericity indicated that the assumption of sphericity had not been violated for Modality ($p > 0.05$). There was a main effect of Modality $F(2,66)=5.979$, $p=0.004$, $\eta^2=0.153$. There was a main effect of Similarity $F(2,4)=143.688$, $p<.01$, $\eta^2=0.813$. There was a significant interaction between Modality and Similarity $F(4,8)=43.449$, $p<.01$, $\eta^2=0.567$.

To investigate this two-way interaction within the separate two-way ANOVAs for Russian, a further one-way ANOVA was conducted for Modality (AV, AO, and VO). The results showed a significant main effect of Similarity in the AV condition $F(2,70)=66.882$, $p<.01$, $\eta^2=0.656$, in the AO condition $F(2,70)=102.898$, $p<.01$, $\eta^2=0.746$, and the VO condition $F(2,66)=31.842$, $p<.01$, $\eta^2=0.491$. Planned paired comparisons showed that for the Russian items in the AV and AO conditions, participants’ error rates were better for Dissimilar items compared to Identical items ($t(35)=18.606$, $p<.01$ and $t(35)=11.248$, $p<.01$ respectively) and compared to Similar items. The latter were no different than the chance level ($p>0.05$). For the Russian items in the Video condition, Dissimilar items resulted in significantly better error rates than Similar items $t(35)=3.941$, $p<.001$. Furthermore, for the Russian items in the VO condition, a one-way t-test comparing participants’ performance to the chance level (50%) showed that their performance was at chance for Similar and Identical items ($p>0.05$) but better than chance for Dissimilar items ($p<0.05$).

Further planned paired comparisons were performed to investigate the AV benefit. The overall performance in the AV mode will be more accurate than in the AO mode was not supported $t(35)=0.639$, $p=0.527$, thus showing no overall AV benefit effect across all items. However, consistent with prediction in Hypothesis 4b, the AV benefit effect was evident for the Russian Dissimilar items which were responded to significantly more accurately in the AV mode compared to the AO mode $t(35)=2.672$, $p=0.01$. No similar advantage was seen with the native English items ($t(35)=0.763$, $p=0.451$), against the prediction of overall AV effect across all items, including native ones.

Further planned paired comparisons were performed to see whether participants perceived Identical and Similar items differently and showed a significant difference between these two groups of items ($t(35)=8.095$, $p<.01$).
Analysis of Response Time:

The response rates are reported in Table 2.

Table 2. Mean Response Time for Experiment 1 as a functional of Language, Modality, and Similarity for Audio-Visual (AV), Audio-Only (AO), and Video Conditions

<table>
<thead>
<tr>
<th>Similarity</th>
<th>English</th>
<th>Mean</th>
<th>Russian</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AV</td>
<td>AO</td>
<td>Video</td>
<td>AV</td>
</tr>
<tr>
<td>Identical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2525(346)</td>
<td>2453(330)</td>
<td>2573(381)</td>
<td>2517(352)</td>
</tr>
<tr>
<td>Similar</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2791(422)</td>
<td>2709(340)</td>
<td>2887(400)</td>
<td>2796(387)</td>
</tr>
<tr>
<td>Dissimilar</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2738(383)</td>
<td>2785(384)</td>
<td>2833(442)</td>
<td>2785(403)</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2685(384)</td>
<td>2649(351)</td>
<td>2764(408)</td>
<td>2670(381)</td>
</tr>
</tbody>
</table>

Note: Values in parentheses indicate standard deviations.

The Mauchly test of sphericity indicated that the assumption of sphericity had not been violated \( (p > 0.05) \).

The analysis showed a significant main effect of Modality \( F(2,66) = 2.777, p = 0.069, \eta^2 = 0.078 \).

The analysis showed a significant main effect of Language \( F(1,33) = 84.09, p < .001, \eta^2 = 0.718 \).

There was a significant main effect of Similarity \( F(2,66) = 94.433, p < .001, \eta^2 = 0.741 \).

Interpretation of the main effects must be considered in the context of a significant three-way interaction, reported below.

The three way interaction between Modality (AV, VO, AO), Language (English, Russian) and Similarity (Similar, Dissimilar, Identical) was also significant \( F(4,132) = 31.85, p < .01, \eta^2 = 0.146 \).

To investigate the interaction further, 2 separate one-way ANOVAs were conducted for VO and AO each, looking at Language (English, Russian) by Similarity (Similar, Dissimilar, Identical) interaction. For the AO analysis, the interaction was significant \( F(2,66) = 69.496, p < .01, \eta^2 = 0.678 \). In the AO mode, Russian Dissimilar items were responded significantly quicker than Russian Similar \( t(33) = 10.796, p < .001 \). However, for the English items the
pattern was reversed, with English Similar items being quicker than Dissimilar $t(35)=2.149$, $p=0.039$, which was unpredicted.

A one-way ANOVA for the VO, looking at Language (English, Russian) by Similarity (Similar, Dissimilar, Identical) interaction, showed a significant result $F(2,7)=24.747$, $p<.01$, $\eta^2= 0.414$. In the VO mode, Russian Dissimilar Items were responded to quicker than Similar $t(35)=3.941$, $p<.001$. However, no such effect was seen for the English Dissimilar items compared to English Similar items in the VO mode $t(35)=1.010$, $p=0.319$.

Planned pairwise comparisons between item types indicated that there was a significant difference between the response times for Russian Dissimilar items in AO, which were significantly quicker than the responses for the Russian Dissimilar items in the VO ($t(35)=2.976$, $p=0.005$). No such difference in response rates was evident for the English Dissimilar Items between the Audio and Video condition ($t(35)=0.477$, $p=0.637$).

Further planned paired comparisons showed that response rates in the VO condition were significantly slower than those in AO condition $t(35)=2.829$, $p=0.008$. However, there was no AV effect evident in response time across all items compared to AO and there was no AV effect evident for Russian Dissimilar items compared to Similar items ($p>0.01$).

Further planned paired comparisons were conducted to investigate the perception of Identical and Similar items as separate categories. This was evident in quicker responses for the Identical items compared to the Similar items $t(35)=13.033$, $p<.001$.

4.1.3 Discussion

A detailed analysis of errors showed that Russian items that are similar or identical to English items resulted in poor discrimination while Russian items that are dissimilar (non-existent in English) were more accurately and quickly identified as non-native. This pattern was consistent with the idea of perceptual assimilation of non-native speech, demonstrated in the auditory domain. Critically for the main Hypothesis 1, which addresses a novel question of whether there is a similar process in the visual domain, there was some evidence of perceptual assimilation in VO mode for Russian dissimilar items. Indeed, Russian Dissimilar items were identified more accurately and responded to quicker compared to other items. The results are consistent with the expectation that the native language has an influence on the processing of non-native speech not only in the AO and AV modes but critically in VO mode. The prediction that visual aspects of non-native speech that are dissimilar to the native language
will be identified more accurately than visual aspects of non-native speech that are similar to the native language was upheld.

Several theoretical models can account for results in the auditory modality (Best, 1995; P. K. Kuhl, 1991). This study extends the findings to the visual modality, showing similar processes. Participants noticed characteristics of visual non-native speech that do not exist in the native language, thus suggesting that they compared foreign visual speech characteristics to their native prototypes/exemplars. Of greater interest was the sharp boundary between easy identification of non-native speech in the dissimilar condition compared to more challenging identification in the similar condition. This dissociation suggests that comparison with native prototype/exemplars is binary (similar or not). The results do not simply reflect response-bias toward accepting stimuli as native. If this was the case, we would not expect to see significant differences between similar and dissimilar items but rather acceptance of dissimilar Russian items as native as well. The finding that the response latencies were generally slow for items in the VO condition, even for the Russian dissimilar items suggests that ‘perception with scrutiny’ was potentially involved.

However, several results raised questions about the robustness of the observed perceptual assimilation effect in VO mode. Performance in VO mode for Russian items was at chance, with the exception of the most dissimilar Russian items. Moreover, high errors rates were present for English items in VO mode, such as chance performance for the English items which were deemed to be similar to Russian and were expected to be identified more accurately. The latter findings were not expected and were against the assumption that native English items would be identified better in VO mode than non-native Russian items. One possibility is that participants did not perceive stimuli using a prototype classification or comparison with exemplars. Similarity was defined here according to the cross-linguistic differences between Russian-English phonetics. Ranking of error rates found a wide variety in the pattern of errors not predicted a priori. A replication of results with a subset of items is therefore desirable before any firm conclusions can be drawn.

Another novel finding was differences in perception of non-native speech across modalities. As expected, participants made more errors in the VO condition compared to the AV and AO conditions. This suggests that VO input affords limited information during identification of a non-native language when compared to AV and AO modes, consistent with previous fund of knowledge. However, participants were able to extract some information from the visual input when processing the most dissimilar Russian items. This finding suggests that the similarities between native and non-native speech constrain performance not only in the AV and OA
Visual perceptual assimilation of non-native speech

modalities but also in the VO modalities and that there may be a similar process of perceptual assimilation which occurs not only in the OA and OV conditions but also in the visual modality. However, at chance performance for the similar English items cautions against strong inferences. It was also unexpected that English similar items were more accurate in the AO mode than dissimilar items.

Another interesting finding of the current study is that fact that the AV benefit effect was only evident for the Russian dissimilar items. The AV benefit effect (Hypothesis 2a) was expected to be seen across all items, with better performance in the AV mode compared to the AO mode. The fact that this was not the case for any of the English items suggests that there may be no additive (or beneficial) effect on discrimination of any but the most dissimilar aspects of non-native speech. Thus the results extend the finding that report limited AV benefit in non-native speech (e.g. (Harrison, 2016; Hazan, Sennema, & Faulkner, 2002) to a novel language.

To summarise, the experiment demonstrated effect of visual assimilation of non-native speech only for most dissimilar Russian items but there was high error rates for English items and no significant difference between performance of Similar and Identical Russian items in the VO mode, which was at chance. Furthermore, there was very limited AV benefit effect, evident only for the most dissimilar aspects of non-native speech. In contrast to showing limited AV benefit effect, the findings of the experiment were consistent with the idea of perceptual assimilation in the auditory domain. Thus, participants made significantly more errors classifying perceptually similar non-native items supporting the idea that unfamiliar non-native speech is assimilated into native speech categories. This effect was reversed for perceptually dissimilar items, with performance being significantly above chance. These results suggest that a "perceptual magnet" or assimilation process is used to process non-native speech whereby similar and identical items are subject to possible processing costs because of resemblance to native speech prototypes whereas dissimilar items that do not match a prototype or specific exemplar can be identified more readily (Best, 1995; P. K. Kuhl, 1991). Again, the fact that the English similar items were discriminated better and quicker than dissimilar was unexpected, suggesting that the process of auditory perceptual assimilation may need to be replicated when tested with a novel pair of languages.

It is important to comment about the limitations of the current study. Specifically, the classification of items was based on Russian-English comparative phonetics (Hamilton, 1980). Although the stimuli was reviewed by a bilingual qualified teacher of Russian and English, the stimuli was not validated with error data or pilot study before the experiment.
Visual perceptual assimilation of non-native speech

The results of participants’ error ranking allows for finer calibration and therefore derivation of a sub-selection of items that can be more precisely defined as similar or dissimilar for native English speakers. These rankings will serve as methodological controls in subsequent experiments.

4.2 Experiment 2: Validation study

Experiment 1 suggested that visual aspects provide restricted benefit for the discrimination of only most dissimilar non-native sounds. Moreover, performance in VO mode resulted in high error rates, including for English items. The classification of items for Experiment 1 was based on Russian-English comparative phonetics (Hamilton, 1980) and might have not reflected differences in visual correlates of speech. Thus, ranking of errors from Experiment 1 allows of derivation of a subset of items reflecting the range of identification performance (error rates) for the three modalities with native English speakers. The aim of Experiment 2 was to interrogate the results from Experiment 1 and re-test the primary hypothesis that visual aspects of non-native dissimilar speech are recognised better/quicker than those for similar non-native speech.

In Experiment 2 participants were presented with either English or Russian vowels/or syllables and requested to indicate whether the item is English or not. The stimuli were presented in different modalities-AO, VO, and AV. The key independent variables were error rates and response times. The dependent variables were language (English or Russian), Modality (AO, VO, and AV), and the degree of Similarity to the native prototype/categories for non-native Russian items (Similar, Identical, Dissimilar).

Experiment 2 Hypotheses:


Hypothesis 1 states that in the visual only mode (VO) Russian items that are similar to English will be assimilated into the native categories/prototypes. In contrast, Russian items that are most dissimilar will be discriminated as foreign in the VO mode. Thus, Hypothesis 1 predicts that the degree of assimilation into English is dependent on the similarity of the Russian item to the native English category.
Hypothesis 1a: Replication of Auditory Perceptual Assimilation with an experimentally derived set of stimuli.

Hypothesis 2: Modality Hierarchy: AO>VO

Hypothesis 2: Visual modality will be less informative (less accurate and results in slower response rates) compared to the auditory modality.

Analysis of Item Goodness: In order to analyse which items were the best and the worst discriminators for the participants, ranking of items according to item mean error rates was performed for each language.

Table 3 and Table 4 show the results of the ranking process.

**Table 3. The Best and the Worst Mean Percent Error for the Russian Items in the Audio Only, Video Only, and Audio Visual (AV) Conditions.**

<table>
<thead>
<tr>
<th>Russian 20 Worst Error Items</th>
<th>Russian 20 Best Error Items</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Items</strong></td>
<td><strong>Audio</strong></td>
</tr>
<tr>
<td>mam1a</td>
<td>100</td>
</tr>
<tr>
<td>as2</td>
<td>100</td>
</tr>
<tr>
<td>mam3</td>
<td>100</td>
</tr>
<tr>
<td>mam4</td>
<td>100</td>
</tr>
<tr>
<td>mam2</td>
<td>95.5</td>
</tr>
<tr>
<td>as4</td>
<td>95.5</td>
</tr>
<tr>
<td>a2</td>
<td>90.9</td>
</tr>
<tr>
<td>a3</td>
<td>90.9</td>
</tr>
<tr>
<td>Ma3</td>
<td>90.9</td>
</tr>
<tr>
<td>a4</td>
<td>90.9</td>
</tr>
<tr>
<td>sa4</td>
<td>90.9</td>
</tr>
<tr>
<td>as1</td>
<td>86.4</td>
</tr>
<tr>
<td>ma1</td>
<td>86.4</td>
</tr>
<tr>
<td>sa2</td>
<td>86.4</td>
</tr>
<tr>
<td>ma2</td>
<td>86.4</td>
</tr>
<tr>
<td>as3</td>
<td>86.4</td>
</tr>
<tr>
<td>a1</td>
<td>85.7</td>
</tr>
<tr>
<td>si1</td>
<td>81.8</td>
</tr>
</tbody>
</table>
Visual perceptual assimilation of non-native speech

<table>
<thead>
<tr>
<th>Russian 20 Worst Error Items</th>
<th>Russian 20 Best Error Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Items</td>
<td>Audio</td>
</tr>
<tr>
<td>sal</td>
<td>81.8</td>
</tr>
<tr>
<td>am1</td>
<td>81.8</td>
</tr>
</tbody>
</table>

Note: *Each of the 4 utterances of the same sound was numbered from 1 to 4.

Table 4. The Best and the Worst Mean Percent Error for the English Items in the Audio Only, Video Only, and Audio-Visual (AV) Conditions.

<table>
<thead>
<tr>
<th>English 20 Worst Error Items</th>
<th>English 20 Best Error Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Items</td>
<td>Audio</td>
</tr>
<tr>
<td>mi:m1</td>
<td>81.8</td>
</tr>
<tr>
<td>rir1</td>
<td>72.7</td>
</tr>
<tr>
<td>rir4</td>
<td>72.7</td>
</tr>
<tr>
<td>rir2</td>
<td>71.4</td>
</tr>
<tr>
<td>i:s1</td>
<td>68.2</td>
</tr>
<tr>
<td>ri:4</td>
<td>68.2</td>
</tr>
<tr>
<td>i:r1</td>
<td>63.6</td>
</tr>
<tr>
<td>i:r2</td>
<td>63.6</td>
</tr>
<tr>
<td>i:s2</td>
<td>63.6</td>
</tr>
<tr>
<td>rir3</td>
<td>63.6</td>
</tr>
<tr>
<td>i:s3</td>
<td>63.6</td>
</tr>
<tr>
<td>i:r4</td>
<td>63.6</td>
</tr>
<tr>
<td>i:s4</td>
<td>63.6</td>
</tr>
<tr>
<td>mi:m4</td>
<td>63.6</td>
</tr>
<tr>
<td>i:m3</td>
<td>61.9</td>
</tr>
<tr>
<td>si:s4</td>
<td>61.9</td>
</tr>
<tr>
<td>i:1</td>
<td>60</td>
</tr>
<tr>
<td>mi:2</td>
<td>59.1</td>
</tr>
<tr>
<td>mi:m3</td>
<td>57.1</td>
</tr>
<tr>
<td>sæs2</td>
<td>54.5</td>
</tr>
</tbody>
</table>

Note: *Each of the 4 utterances of the same sound was numbered from 1 to 4.
Table 5 shows descriptive statistics for the English and Russian languages and three modalities of presentation.

Table 5. The Best and the Worst Mean Percent Error in the Audio Only, Video Only, and Audio-Visual (AV) Conditions.

<table>
<thead>
<tr>
<th>Item type</th>
<th>Russian</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Audio</td>
<td>Video</td>
</tr>
<tr>
<td>Best a</td>
<td>7.53</td>
<td>41.24</td>
</tr>
<tr>
<td>Worst</td>
<td>90.43</td>
<td>62.81</td>
</tr>
<tr>
<td></td>
<td>(6.26)</td>
<td>(18.01)</td>
</tr>
</tbody>
</table>

Note: a 20 best and 20 worst items’ mean error rates with sd in parentheses.

Analysis of error rates from AO and AV conditions in Experiment 1 identified three types of items: Dissimilar (Russian items were correctly and consistently identified as non-native), English-As-Foreign (English items that were incorrectly identified as non-native and were not anticipated in Experiment 1) and Similar (items that were identified correctly as English and Russian items misidentified or assimilated into English). The rationale for using these groups of item was the fact that they were experimentally derived and should be more salient in terms of their visual correlates of speech compared to the stimuli from Experiment 1 which were based on comparative phonetics rather than differences in the visual characteristics of English and Russian. By using an experimentally derived set of stimuli, it was anticipated that visual perceptual assimilation could be demonstrated more robustly. Audio and Video conditions were used for direct comparison with the aim of minimising artefacts arising in the AV condition as the previous experiment showed no significant additive effect of AV presentation.

4.2.1 Method

Participants: A different group of participants to that used in Experiment 1 was used in Experiment 2. Twenty seven undergraduate psychology students from the University of Melbourne participated and received course credit for participation. There were 17 females. The mean age of participants was 24 years. All were native speakers of English and none had prior knowledge of the Russian language. All reported having no known hearing impairment and normal or corrected-to-normal vision.
**Materials:** The stimuli comprised a sub-selection of the stimuli used in Experiment 1. Eighteen items were used from the original stimuli set. They were in turn divided into 3 groups of 6 items each: *Dissimilar* (Russian items correctly and consistently identified as Russian), *English-As-Foreign* (English items incorrectly identified as Russian), and *Similar* (items that were identified correctly as English and the Russian items that were misidentified as English). Each item was repeated four times, resulting in 72 trials.

**Procedure:** Each participant was tested in two conditions: audio only (AO) and video only (VO). Order of stimuli presentation was randomised. Participants were tested individually in a sound attenuated room. In the VO condition, videos of the whole face of a speaker pronouncing an item appeared centred on a computer monitor (sized 16 x19 cm, visual angle 5 degrees) and were presented one by one in a random order. In the AO task, a static picture of a speaker was presented to minimise differences between visual and auditory conditions. In each condition, participants read a set of instructions asking them to sort sounds into native and non-native items. Participants were required to press the “Yes” key if an item was an English item and the “No” key if it was a non-English item (keys were counterbalanced). Participants were instructed that each item would be presented several times. In the VO condition they were asked to sort stimuli on the basis of facial information provided by the speaker who was pronouncing the stimuli, without corresponding sound. To minimise bias, the non-native language was not defined as Russian but referred to only as a foreign language. Before the main experiment for each condition, participants were presented with 4 practice trials. The experiment lasted approximately 45 minutes with breaks of 5 minutes before each set.

### 4.2.2 Results

**Error Analyses:**
A mixed two-way ANOVA was conducted with the within-participants factors Modality (VO, AO) and Similarity (*English-As-Foreign; Similar, Dissimilar*) and order of task presentation as between-participants variable. The latter was not significant ($F<1$) and the below analysis is reported without including that between-participant factor.

Table 6 shows descriptive statistics for each condition.
Table 6.  Mean Percentage Error Rates for Each Condition: Audio Only (AO) and Video

<table>
<thead>
<tr>
<th>Similarity</th>
<th>AO</th>
<th>Video</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-As-Foreign;</td>
<td>39.99 (21.67)</td>
<td>51.17 (7.80)</td>
<td>45.85 (14.74)</td>
</tr>
<tr>
<td>Dissimilar;</td>
<td>26.54 (15.38)</td>
<td>53.39 (12.23)</td>
<td>40.00 (13.81)</td>
</tr>
<tr>
<td>Similar</td>
<td>53.70 (13.04)</td>
<td>55.87 (10.80)</td>
<td>54.79 (11.92)</td>
</tr>
<tr>
<td>Mean</td>
<td>40.08 (16.70)</td>
<td>53.48 (10.28)</td>
<td>46.88 (13.49)</td>
</tr>
</tbody>
</table>

Note: standard deviation reported in parentheses.

The Mauchly test of sphericity indicated that the assumption of sphericity had not been violated ($p>0.05$).

There was a significant main effect of Modality $F(1,26)=30.539, p<.001, \eta^2=0.540$.

There was a significant main effect of Similarity $F(2,52)=16.002, p<.001, \eta^2=0.381$.

The interaction between Modality and Similarity was also significant $F(2,52)=11.589, p<.001, \eta^2=0.308$, as shown in Figure 1. To interrogate the results of the interaction, a one-way ANOVA was conducted for each Modality. The results were not significant for the VO modality $F(2,52)=1.691, p=0.194, \eta^2=0.061$. A one-sample $t$-test comparing the errors to the chance level (50%) showed that their performance was not different from chance $t(26)=2.547, p>0.01$.

A one-way ANOVA for the AO modality showed that Mauchly’s test of sphericity was significant ($p=0.003$) and the assumption of sphericity had been violated. The Greenhouse-Geisser’ corrections were applied and the results showed a significant effect of Similarity $F(1.448,37.660)=17.635, p<.001, \eta^2=0.404$. Planned paired comparisons showed that Dissimilar items were significantly more accurate than English-As-Foreign items $t(26)=2.318, p=0.029$ in the AO format. A one-sample $t$-test comparing errors to the chance level found errors for the Similar items did not differ from the chance level in the Audio condition $t(26)=1.474, p=0.153$. 

94
Figure 1. Interaction Between Modality and Similarity in the Identification Task (English versus Non-English).

Note: The points represent mean error rates for the Dissimilar, English-As-Foreign, and Similar item types. The whiskers depict standard error of the means.

Response Time analyses:
A mixed two-way ANOVA was conducted with the within-participants factors Modality (VO, AO), Similarity (English-As-Foreign; Dissimilar; Similar), and order of task presentation as between-participants variable. The latter was not significant (F<1) and the analysis is reported without including that between-participant factor.

Table 7 shows descriptive statistics for each condition.

Table 7. Mean Response Time in ms for Each Condition: Audio Only (AO) and Video.

<table>
<thead>
<tr>
<th>Similarity</th>
<th>AO</th>
<th>Video</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-As-Foreign</td>
<td>2314 (382)</td>
<td>2598 (441)</td>
<td>2456 (412)</td>
</tr>
<tr>
<td>Dissimilar</td>
<td>1918 (340)</td>
<td>2585 (446)</td>
<td>2251 (393)</td>
</tr>
<tr>
<td>Similar</td>
<td>2177 (461)</td>
<td>2524 (433)</td>
<td>2305 (447)</td>
</tr>
<tr>
<td>Mean</td>
<td>2136 (395)</td>
<td>2569 (440)</td>
<td>2337 (417)</td>
</tr>
</tbody>
</table>
Visual perceptual assimilation of non-native speech

Note: standard deviation reported in parentheses.

The Mauchly test of sphericity indicated that the assumption of sphericity had been violated ($p<0.01$) and Greenhouse-Giesser corrections were applied.

There was a significant main effect of Modality $F(1,25)=41.069, p<.001, \eta^2 = 0.622$. There was a significant main effect of Similarity $F(1.765, 44.115)=23.953, p<.001, \eta^2 = 0.489$. The interaction between Modality and Similarity was also significant $F(1.508, 37.688)=17.453, p<.001, \eta^2 = 0.411$. To interrogate the results of the interaction, a one-way ANOVA was conducted for each Modality. The results were not significant for the VO modality $F(2,52)=2.924, p=0.06, \eta^2 = 0.101$.

A one-way ANOVA for AO format showed that Mauchly’s test of sphericity had not been violated ($p>0.01$). There was a significant effect of Similarity $F(2,50)=30.776, p<.001, \eta^2 = 0.552$. Planned paired comparisons showed that Dissimilar items were responded to quicker than English-As-Foreign items $t(25)=8.589, p<.001$ in the AO format.

Discussion

The aim of this experiment was to validate the results of Experiment 1 with a sub-set of stimuli that were experimentally derived and more experimentally valid than that based on the comparative phonetics between Russian and English. However, the primary hypothesis that there is a visual perceptual assimilation process of non-native speech was not supported. The responses in the VO condition were at chance regardless of item type, indicating a relatively high degree of task difficulty. This was reflected in slow response times for participants in the VO mode. High error rates were also observed in the visual modality in the previous Experiment 1. The findings here show that even for the most discriminable dissimilar non-native items, VO input does not produce enough information to make an accurate decision even if visual characteristics of non-native speech are potentially identified, as evident in slower reaction times. Similar to Experiment 1, participants were more accurate and quick in the AO condition compared to the VO condition.

The results of Experiment 2 replicated results from the AO condition in Experiment 1. The Dissimilar items produced fewest errors. Moreover, responses in AO conditions for the Similar items were at chance. As those items correspond to items misidentified as English in Experiment 1, the present results replicate the previous findings. Finally, results with the
*English-As-Foreign* items replicate the findings from Experiment 1 by showing that participants misidentified certain English items as non-native in the AO condition.

To summarise, the aim of Experiment 2 was to validate the results of Experiment 1 with a sub-selection of experimentally derived stimuli. However, the results were not replicated, possibly because of the high degree of difficulty inherent in the language identification paradigm.

### 4.3 Experiment 3: Same-Different Paradigm

Results so far suggest that our participants gain little extra from visual aspects when discriminating unfamiliar sounds at the level of vowels/syllables in their non-native language. The previous experiment showed no reliable and robust evidence to support the primary hypothesis that there is visual perceptual assimilation. To test the reliability of these findings, a Same-Different paradigm was developed to present the same stimuli as that used in Experiment 2 but using different task demands. This paradigm also allow to test whether participants’ difficulties in making a language identification judgement were due to the low-level perceptual difficulties in discriminating between the two language’s contrasts. The prediction was that participants would have greater difficulties differentiating similar contrasts compared to dissimilar contrasts in native and non-native conditions. Two conditions were used -AV and AO, to allow for a direct comparison and closer examination of the AV benefit effect.

In Experiment 3 participants were presented with either English or Russian vowels/or syllables paired with the same stimuli (language congruous) or with different stimuli (language incongruous). They were requested to indicate whether two stimuli were the same or not. The stimuli were presented in 2 modalities-AO and AV. The key independent variables were error rates and response times. The dependent variables were Pair type (language congruous and language incongruous), Modality (AO and AV), and the degree of Similarity to the native prototype/categories for non-native Russian items (Similar, English-as-Foreign, Dissimilar).

Experiment 3 Hypotheses:

Hypothesis 1: Participants will be more accurate/quick for same pairs than different pairs of stimuli, regardless of whether they are native or not.
Hypothesis 2: Participants will be more accurate/quick with most Dissimilar items compared to the Similar items.

Hypothesis 3: AV benefit effect AV>AO- participants will be more accurate/quick in the AV mode compared to the AO mode.

4.3.1 Method

Participants: Thirty-four undergraduate psychology students from the University of Melbourne who had not participated in any of the previous experiments participated and received a course credit for participation. There were 20 females. The mean age of participants was 22 years. All were native speakers of English and none had prior knowledge of the Russian language. All reported having no known hearing impairment and normal or corrected-to-normal vision.

Materials: The stimuli were the same as Experiment 2 i.e. three categories of items: Dissimilar, English-As-Foreign, and Similar. A total of 36 items were used: 18 English and 18 Russian items. Within the 18 items for each language, there were 2 instances of the same sound produced by the same speaker on two different occasions. Eighteen items from each language were then paired either with the same item or with the corresponding item from the other language. For instance, in the AO condition there were 4 combinations: (1) an English sound followed by the another instance of the same sound in English; (2) an English sound followed by the corresponding sound in Russian; (3) a Russian sound followed by another instance of the same sound in Russian; (4) a Russian sound followed by the corresponding English sound.

Procedure: Before each item the word “ready” was flashed on screen, signalling the beginning of a trial. Each item was repeated four times during the course of the experiment: twice followed by the same stimulus and twice by a corresponding stimulus from the other language. If a no response was given, the next item was automatically presented. No feedback was provided. In the task, participants read instructions that explained that a bilingual person recorded sounds in English and another foreign language. It was also explained that audio files were extracted from AV recordings and that stimuli were paired so that some belong to the same language while others do not. For instance, the English sound “ra” can be followed by the speaker pronouncing the other instance of “ra” in English or by “ra” in the foreign language. Four examples were provided and participants practiced to make a response. They
were instructed to press the right “YES” button if the stimuli were the same and the left “NO” button if they were not. Participants were also instructed that each item may be presented several times over the course of the experiment. There were two conditions: AO and AV and the order of their presentation was randomised across participants.

4.3.2 Results

Analysis was a within-subjects three-way ANOVA. Within-participants factors were Modality, Similarity and Pair Type. The Modality factor had 2 levels: AO and AV. The Similarity factor had 3 levels: Dissimilar, English-As-Foreign, and Similar. The Pair Type factor had 2 levels: Same and Different. The between-subject factor Order of Presentation was not significant (F<1) and the analysis is reported without that factor. Both response times and error rates were collected and analysed.

Table 8 shows descriptive statistics for all the stimuli.

**Table 8. Mean Percentage Error Rates for Same-Different Experiment.**

<table>
<thead>
<tr>
<th>Modality</th>
<th>Pair Type</th>
<th>Similarity</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Similar</td>
<td>Dissimilar</td>
</tr>
<tr>
<td></td>
<td>Different</td>
<td>67.03 (26.87)</td>
<td>3.43 (7.57)</td>
</tr>
<tr>
<td>AV</td>
<td>Same</td>
<td>21.70 (10.94)</td>
<td>7.29 (9.32)</td>
</tr>
<tr>
<td></td>
<td>Different</td>
<td>72.54 (23.70)</td>
<td>5.03 (7.98)</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>46.88 (18.58)</td>
<td>6.34 (8.42)</td>
</tr>
</tbody>
</table>

Note: standard deviation reported in parentheses.

The Mauchly test of sphericity indicated that the assumption of sphericity had been violated ($p<0.05$). Therefore, the Greenhouse-Geisser corrections were applied. There was a significant main effect of Similarity $F(1.083,35.735)= 100.98$, $p<.001$, $\eta^2 = 0.754$. 

99
Visual perceptual assimilation of non-native speech

There was a significant main effect of Pair type $F(1,33)= 31.801, p<.001, \eta^2= 0.491$. These main effects will be interpreted in the context of a significant three-way interaction reported below.

There was a three-way interaction between Modality, Pair Type, and Similarity $F(1.390, 45.862)=83.164, p<.001, \eta^2= 0.716$. Two separate one-way ANOVAs for each Modality were conducted to investigate the interaction further. The results were significant for all factors ($p<.001$). In the AO modality, all items were worse when in Different language pairs compared to Same language pairs except for Dissimilar items in Different language pairs which were better than Dissimilar items in Same language pairs $t(33)=3.780, p=0.001$. The pattern was different in the AV format with Dissimilar items in same language pairs showing no significant difference in accuracy compared to the Dissimilar items in different language pairs.

Further planned pairwise comparison showed that participants made significantly more errors for the Similar items compared to the Dissimilar items $t(33)=14.72, p<.001$.

A planned pairwise comparison showed significantly more errors for Different language pair errors than Same language pairs $t(33)=5.71, p<.001$.

Planned paired comparisons showed that participants made fewer errors for the Same language pairs in the AV compared to the AO condition $t(33)= 4.63, p<.001$. This was consistent with Hypothesis 3 of AV benefit effect. However, the effect did not hold the different language pairs as they made more errors in the AV condition when items were from Different language pairs $t(33)=2.69, p=0.01$.

Response Times Analysis.

Table 9 shows descriptive statistics (response times) for all the stimuli.

### Table 9. Mean Response Times in ms in Same-Different Experiment for Audio Only (AO) and Audio-Visual (AV) Conditions

<table>
<thead>
<tr>
<th>Similarity</th>
<th>Same AO</th>
<th>Same AV</th>
<th>Same Mean</th>
<th>Different AO</th>
<th>Different AV</th>
<th>Different Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>English as Foreign</td>
<td>2165(453)</td>
<td>2015(507)</td>
<td>2090(480)</td>
<td>1817(431)</td>
<td>1830(386)</td>
<td>1823(408)</td>
</tr>
<tr>
<td>Dissimilar</td>
<td>1696(475)</td>
<td>1659(434)</td>
<td>1676(455)</td>
<td>1939(411)</td>
<td>1907(351)</td>
<td>1921(381)</td>
</tr>
<tr>
<td>Similar</td>
<td>1947(420)</td>
<td>1878(401)</td>
<td>1913(410)</td>
<td>1867(379)</td>
<td>1946(581)</td>
<td>1906(480)</td>
</tr>
</tbody>
</table>
The Mauchly test of sphericity indicated that the assumption of sphericity had not been violated ($p > 0.05$).

There was a significant main effect of Similarity $F(2,50)= 19.538, p<.001$, $\eta_p^2= 0.439$.

There was a significant interaction between Modality and Pair Type $F(1,25)=4.162, p<0.05$, $\eta_p^2= 0.143$.

There was a significant interaction between Pair Type and Similarity $F(2,50)=28.526, p<.001$, $\eta_p^2= 0.533$.

A one-way ANOVA for each Pair Type showed significant results (all $p<.001$) for all but Different language AV condition. Planned pairwise comparisons showed that there was no difference in AO for Same and Different Pair types (all $p>0.01$) but Same language pair resulted in quicker response in AV format compared to AO $t(33)=2.334, p=0.026$.

Planned paired comparisons showed that Dissimilar items were responded to quicker than Similar items $t(33)=2.32, p=0.03$.

4.3.3 Discussion

The purpose of this experiment was to interrogate sensitivity to native and non-native sounds using an alternative paradigm, i.e., Same-Different judgement. As expected in Hypothesis 1, participants made more errors for pairs from different languages compared to those from the same language. In the AO condition, a decision about whether two items were different also took longer than whether they were the same, suggesting a more conservative response bias as reflected in slower reaction times. The results suggest that identification of items from different languages was more demanding on resources before a same judgement could be made (with Caucasian face and languages from the same family), also generating more errors, most likely because non-native items did not conform to extant categories and positive responses required more information.

Consistent with our expectation of Hypothesis 2 that the most dissimilar items would be easier to discriminate, fewest errors were observed for items that are most different to English (Dissimilar). Conversely, more errors and slower responses were seen for Russian items that are similar to English (Similar). These results are in accord with multiple studies reporting that perceptual sensitivity to non-native sounds is affected by the similarity to a native
language. This is consistent with the view that similarity of an item to the native language will cause discrimination errors due to ambiguity in a category boundary between native and non-native languages (Best, 1995; Flege et al., 1997; P. K. Kuhl, 1991). Furthermore, sensitivity in discriminating sounds that are most different from English was diminished when they were not contrasted to native sounds. This was evident in the fact that the participants made significantly more errors when sounds that were most different from English (Dissimilar) were presented in the same pair compared to when they were presented with the contrasting English sound in the AO modality.

In line with the findings of Experiment 1 and Experiment 2 the present results show a limited facilitatory effect of AV speech compared to AO speech. As predicted in Hypothesis 3, participants found it easier to decide (including quicker response rates) that items belong to the same language in the AV modality than the AO condition, while they had greater difficulties judging items from different languages in the AV modality. Therefore, there is an advantage for AV presentation when the language is the same, suggesting visual aspects might give extra weight to confirmatory decisions (true positive) but this may lead to greater resistance when a negative decision (true negative) is required. Correct judgement of different pairs may be more difficult because more information is needed before a judgement is made, requiring a higher decision threshold and more cognitive resources to make a correct decision. In contrast, in the AO condition, participants were better in judging dissimilar Russian items when they were contrasted with the English ones, although this was not reflected in quicker response times.

The current study validates the classification of items into different groups by showing that some English items (English-As-Foreign) and some Russian items (Similar) were not identified correctly as expected based on rankings from Experiment 1. The results extend these intuitions, as Russian items that are similar to English items are not easily discriminated in the AV or AO conditions, suggesting again that reference might be made to extant category prototypes or exemplars during non-native speech perception. If no category were relevant i.e. during dissimilar item discrimination, judgements would be less ambiguous, easier and faster.

The experimental paradigms used thus far are limited for exposing mental processes involved in the perception of non-native speech. Due to the design of experiments, which again forced the participants to make a binary response, it remains unclear whether some Russian items were wrongly accepted as belonging to English but judged to be ‘poor’ exemplars of English. Item ratings are a more accurate estimate of the processes used during participants’ responses.
4.4 Experiment 4: Item Rating

This experiment investigated how participants perceived experimental stimuli in terms of similarity to their native English language. A ranking task is routinely used in evaluating similarity of items within a given category, especially in the domain of auditory speech perception (P. K. Kuhl, 2010; P. K. Kuhl et al., 1992). In the current study the question was related to between category discrimination and perception of items from different languages. The purpose was to see whether the categories of items which were derived from the previous experiments corresponded to participants’ judgement of these items.

Experiment 4 Hypotheses:
Hypothesis 1. Similar items would be ranked as native English items whilst Dissimilar items would be ranked as non-native.
Hypothesis 2. AV condition’s ranking would be more accurate than in the AO condition, thus showing the AV benefit effect.

4.4.1 Method

Participants: Twenty-four undergraduate psychology students from the University of Melbourne who had not participated in any of the previous experiments participated and received a course credit for participation. There were 19 females. The mean age of participants was 22 years. All were native speakers of English and none had prior knowledge of the Russian language. All reported having no known hearing impairment and normal or corrected-to-normal vision.

Materials: The stimuli for this experiment were the same as those used in Experiments 2 and 4. There were three groups of items: English-As-Foreign, Dissimilar, and Similar. A total of 36 items was used: 18 English and 18 Russian items.

Procedure: Each item was repeated four times. A 5-point likert scale was presented after each stimulus, reminding participants that 1 means a Non-English item, 2 means a Very Bad example of an English sound, 3 means Acceptable, 4 means a Good English item and 5 means Very Good. Participants were instructed to press corresponding numbers on a keyboard and told that an item may be presented several times over the course of the experiment. The order of AV and audio conditions were randomised across participants.
4.4.2 Results

Participants ranked stimuli under different conditions (AV and AO). Item ratings were averaged for each condition and two-way ANOVA was performed. The first factor Item type had 3 levels (English-As-Foreign, Dissimilar, and Similar) and the second factor Modality had two levels (AV and AO).

The between-subject factor Order of Presentation was not significant (F<1) and the analysis is reported without that factor.

Table 10 shows descriptive statistics for two conditions in Experiment 4.

<table>
<thead>
<tr>
<th></th>
<th>Audio</th>
<th>AV</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dissimilar</td>
<td>2.38</td>
<td>2.19</td>
<td>2.29</td>
</tr>
<tr>
<td>(0.34)</td>
<td>(0.38)</td>
<td>(0.36)</td>
<td></td>
</tr>
<tr>
<td>Similar</td>
<td>4.25</td>
<td>4.44</td>
<td>4.35</td>
</tr>
<tr>
<td>(0.21)</td>
<td>(0.25)</td>
<td>(0.23)</td>
<td></td>
</tr>
<tr>
<td>English-As-Foreign</td>
<td>3.71</td>
<td>3.65</td>
<td>3.68</td>
</tr>
<tr>
<td>(0.53)</td>
<td>(0.42)</td>
<td>(0.48)</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>3.45</td>
<td>3.43</td>
<td>3.44</td>
</tr>
<tr>
<td>(0.36)</td>
<td>(0.35)</td>
<td>(0.36)</td>
<td></td>
</tr>
</tbody>
</table>

Note: 1=Non-English; 2=Very Bad; 3=Acceptable; 4=Good; 5=Very Good.

The Mauchly test of sphericity indicated that the assumption of sphericity had not been violated (p>0.05).

There was a significant main effect of Item Type \(F(2,46)=229.672, p<.001, \eta^2=0.909\).

There was a significant two-way interaction between Item Type and Modality \(F(2,46)=14.303, p<.001, \eta^2=0.383\).

Planned pairwise comparison showed that rating for Dissimilar items were significantly lower (judged as more foreign) than for Similar items \(t(23)=27.993, p<0.001\) or English-as-Foreign items \(t(23)=12.958, p<0.001\). Further planned pairwise comparisons showed that Similar items resulted in higher rating (judged as better English) that English-as-Foreign items \(t(23)=6.097, p<0.001\).

The interaction between Item Type and Modality was further explored with planned pairwise comparisons. Dissimilar items were judged as more foreign in the AV mode than in the AO mode \((t(23)=4.811, p<0.001\). The latter was consistent with the AV benefit effect. However, for the Similar items the opposite was the case, with higher ranking (better English) in the AV mode than in the AO mode.
4.4.3 Discussion

The purpose of this experiment was to investigate how participants perceive experimental stimuli, with an emphasis on non-native stimuli. It was unclear whether participants made errors in previous experiments because they incorrectly classified English items as Russian but considered them as bad examples of English items. It was hypothesised that the Russian items that are similar to English would be accepted as English but it was unclear whether they would be perceived as bad examples. Given differences between AO and AV errors in some experiments but not others, it was important to trace any difference in ranking under these conditions.

The results showed that participants judged Russian items that are most dissimilar to English as non-English or very bad examples of English. Consistent with the derived category of Similar items, participants rated English items as very good exemplars of English and misclassified Russian items that were similar to English as very good examples of English items. Interestingly, English items that comprised the English-As-Foreign category were considered by participants as acceptable English items, suggesting participants accepted them as English, although not as well as Similar items’ members.

One interesting result was the difference between AO and AV conditions, with the participants giving a higher ranking (judging items as better English items) in the AO condition for Similar items but ranking Dissimilar items as more foreign in the AV condition. One explanation for this pattern is that seeing a speaker talk made the stimuli look less English. Perhaps a bilingual face affords paralinguistic characteristics that make stimuli appear less English. This is unlikely though, as participants gave higher ratings (judging items as better English items) for Similar items in AV condition compared to the AO condition. Again, the current experiment demonstrated a limited AV benefit effect, which is dependent on the item’s degree of similarity to the native language. In sum, English-speaking participants possibly perceive Russian auditory and AV stimuli categorically, as evidenced by differences in ratings in AO and AV conditions. The cause of these differences is unclear and further study of mental representations of non-native sounds is required to understand the processes involved in perception of foreign speech, especially its visual characteristics.
4.5 **Experiment 5: Match-Mismatch**

The purpose of this experiment was to examine visual representations of speech that can be inferred from cross-modal matching. It was hypothesised that when participants see the face of a speaker pronouncing a sound without the corresponding auditory input, this would result in translation of the visual information into sounds. It was further hypothesised that visual information may not be sufficient to generate the exact corresponding auditory sound but it will be sufficient to generate several possibilities or perhaps a fuzzy auditory representation. The mechanism of this generation/translation is assumed to be different in relevant theories. For instance, Prototype theory assumes matching or mapping of visual information into visual prototypes and corresponding auditory prototypes. In contrast, Motor theory proposes a direct translation from facial motor movement in the visual modality into the corresponding sounds.

Another hypothesis tested here was that when participants see a face pronouncing non-native sounds, responses would differ depending on the similarity of the sound to English sounds (prototypes). It was predicted that participants would correctly match Video and Sound with Russian sounds that are different to English but not with the ones that are similar to English.

The second purpose of this experiment was to test differences between Video-Sound and Sound-Video conditions. It was predicted that matching would be easier in the Sound-Video condition since more information is available in the auditory modality, resulting in a better generation/translation process. It was predicted that when participants hear a non-native sound that is sufficiently different from their native language, they would be able to match it correctly to the corresponding visual image but not when the sound is similar and is perceived as English. Finally, it was anticipated that there would be fewer errors when two stimuli are matched e.g. visual image and sound match (cross-modality continuous) compared to when they are not matched (cross-modality discontinuous). Therefore, the design of the experiment was to present participants with a stimulus in one modality followed by either a matched or mismatched stimulus in the other modality. The independent variables were again error rates and response time rates whilst dependent variables were Format (Sound-Video, Video-Sound), Similarity (Dissimilar, Similar, English-as-Foreign), and Pair Type (Match and Mismatch).

**Experiment 5 Hypotheses:**

Hypothesis 1: Performance in the Sound-Video condition will be more accurate and will results in quicker response times than that in the Video-Sound condition;
Hypothesis 2: Performance for the *Dissimilar* items will be more accurate and will result in quicker response times than that for the *Similar* items;

Hypothesis 3. Performance for the Matched pairs (cross-modality continuous) will be more accurate and will result in quicker response times than that for the Mismatched pairs (cross-modality discontinuous).

4.5.1 Method

*Participants:* A different group of participants to that used in previous experiments was used in Experiment 5. Twenty seven undergraduate psychology students from the University of Melbourne participated and received course credit for participation. There were 17 females. The mean age of participants was 24 years. All were native speakers of English and none had prior knowledge of the Russian language. All reported having no known hearing impairment and normal or corrected-to-normal vision.

*Materials:* The stimuli were the same as those used in Experiment 2. They were again three groups with 6 items each: *Dissimilar* (the Russian items that were correctly and consistently identified as Russian), *English-As-Foreign* (the English items that were incorrectly identified as Russian), and *Similar* (the items that were identified correctly as English and the Russian items that were misidentified as English). A total of 36 items were used: 18 English and 18 Russian items.

*Procedure:* Within 18 items for each language, there were 2 instances of the same sound produced by the same speaker on two different occasions. 18 items from each language were then paired either with the same item or with the corresponding item from the other language. The original AV files were separated into audio and video files. In the next step, the files were paired in four possible combinations. For instance, for the audio-visual condition there were two matched pairs and two mismatched:

1. An audio English item followed by a corresponding matched video image in English;
2. An audio English items followed by a mismatched video image in Russian;
3. An audio Russian items followed by corresponding matched video image in Russian;
4. An audio Russian word followed by the mismatched video image in English

Before each presentation the word “ready” was briefly flashed on the screen, signalling the beginning of each presentation. Each item from the two languages was repeated 4 times over the course of the experiment: twice followed by a matching stimulus, and twice followed by a non-matching stimulus from the other language. If no response was made during 10s, the next
items automatically appeared. No feedback was provided, except during the practice trials. In the sound-video matching task, participants read a set of instructions that explained that a bilingual person recorded sounds in English and another foreign language. In order to minimise any bias, the Russian language was not identified and was referred to as a “foreign language”. It was also explained that the video images and sounds were separated and paired so that some of them match while others do not. For instance, the English sound “ra” could be followed by the face of the speaker pronouncing “ra” in the other language. Four examples were provided for practice. They were instructed to press the right “YES” button if sound and video matched, and the left “NO” button if they did not match. The participants were also instructed that each item will be presented several times over the course of the experiment. In the “Video-Sound Match” condition, identical instructions were used except that the participants were informed that a video image would appear first, followed by the sound. The order of presentation was counterbalanced across the participants. Prior to the beginning of the experiment 4 practice trials were conducted with feedback provided. The experiment lasted approximately 45 mins with a 5-min break provided between the conditions.

4.5.2 Results

A mixed three-way ANOVA was conducted, where within-subjects factors were Format, Similarity, and Pair Type. The first within-subjects factor, Format, had 2 levels: Sound-Video and Video-Sound. The second within-subjects factor, Similarity had 3 levels: Dissimilar, Similar, English-as-Foreign; and the third factor Pair Type had 2 levels: Match and Mismatch. The between-subject factor was Condition Order, which was not significant (F<1) and the results are reported without including that factor in the analysis.

Table 11 shows descriptive statistics for all factors.

Table 11. Mean Percentage Error Rates in Match-Mismatch Experiment.

<table>
<thead>
<tr>
<th>Format</th>
<th>Pair Type</th>
<th>Similarity</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Similar</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dissimilar</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>English-as-Foreign</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td></td>
</tr>
<tr>
<td>SV</td>
<td>Match</td>
<td>38.00 (10.10)</td>
<td>41.46 (13.63)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>46.00 (14.00)</td>
<td>47.81 (13.84)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40.41 (16.80)</td>
<td></td>
</tr>
<tr>
<td>SV</td>
<td>Mismatch</td>
<td>56.00 (10.08)</td>
<td>54.16 (14.06)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>46.77 (19.61)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>60.14 (12.48)</td>
<td></td>
</tr>
</tbody>
</table>
Visual perceptual assimilation of non-native speech

<table>
<thead>
<tr>
<th></th>
<th>Match</th>
<th>VS</th>
<th>Mismatch</th>
<th>VS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VS</td>
<td>40.20</td>
<td>(12.25)</td>
<td>50.00</td>
<td>(19.68)</td>
</tr>
<tr>
<td>Mismatch</td>
<td>59.26</td>
<td>(12.57)</td>
<td>45.84</td>
<td>(19.06)</td>
</tr>
<tr>
<td>Mean</td>
<td>48.37</td>
<td>(11.18)</td>
<td>47.15</td>
<td>(18.09)</td>
</tr>
</tbody>
</table>

Note: standard deviations reported in parentheses.

The Mauchly’s test of sphericity showed that the assumption of sphericity had not been violated ($p>0.05$).

There was a significant main effect of Format $F(1,26)=18.31$, $p<.001$, $\eta^2=0.413$.

There was a significant effect of Similarity $F(2,52)=11.449$, $p<.001$, $\eta^2=0.306$.

There was a significant effect of Pair Type $F(1,26)=10.843$, $p=0.023$, $\eta^2=0.294$.

There were significant two-way interactions between Format and Similarity $F(2,52)=3.293$, $p=0.044$, $\eta^2=0.112$, as well as between Similarity and Pair Type $F(2,52)=7.766$, $p=0.001$, $\eta^2=0.230$.

To investigate the interactions further, a separate one-way ANOVA was conducted for each Format (Video-Sound and Sound-Video) with the within-subject factor of Similarity (Similar, Dissimilar, English-as-Foreign).

For the Sound-Video one-way ANOVA, there was a significant main effect of Pair Type $F(1,26)=14.940$, $p=0.001$, $\eta^2=0.365$ and there was a significant interaction between Similarity and Pair Type $F(2,52)=5.939$, $p=0.005$, $\eta^2=0.186$. Planned paired comparisons showed that mismatched pairs resulted in higher error rates for Similar and English-as-Foreign items compared to matched pairs (all $p<.001$) but no such difference was evident for the Dissimilar items in the Sound-Video mode ($p>0.01$).

In contrast, a one-way ANOVA for the Video-Sound Pair type only showed a significant main effect of Pair Type $F(1,26)=30.588$, $p<.001$, $\eta^2=0.541$. Planned pairwise comparisons showed that in the Video-Sound condition, responses to the matched pairs were significantly more accurate for Similar items compared to mismatched Similar items $t(26)=4.943$, $p<.001$ but no such difference was evident for the Dissimilar and English-as Foreign items. These results showed limitation for the Hypothesis 3 that matched pairs are more accurate than matched pairs.
Further planned pairwise comparison between the Sound-Video and Video-Sound conditions showed that the participants made significantly more errors for the Video-Sound condition compared to the Sound-Video condition \( t(26)=4.27, p<.001 \), consistent with Hypothesis 1. Planned paired comparisons showed that participants made significantly more errors with the Similar items compared to Dissimilar \( t(26)=3.96, p<.001 \), consistent with Hypothesis 2.

**Response rate analysis:**
A mixed three-way ANOVA was conducted on response rates, where within-subjects factors were Format, Similarity, and Pair Type. The first within-subjects factor, Format, had 2 levels: Sound-Video and Video-Sound. The second within-subjects factor, Similarity had 3 levels: Dissimilar, Similar, English-as-Foreign; and the third factor Pair Type had 2 levels: Match and Mismatch. The between-subject factor was Condition Order, which was not significant \( (F<1) \) and the results are reported without including that factor in the analysis.

Table 12 shows descriptive statistics for response rates for all variables.

**Table 12. Response Rates for Match-Mismatch Experiment.**

<table>
<thead>
<tr>
<th>Format</th>
<th>Pair Type</th>
<th>Item Type</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Similar</td>
<td>Dissimilar</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SV</td>
<td>Match</td>
<td>2432 (378)</td>
<td>2219 (373)</td>
</tr>
<tr>
<td></td>
<td>Mismatch</td>
<td>2378 (346)</td>
<td>2435 (345)</td>
</tr>
<tr>
<td>VS</td>
<td>Match</td>
<td>2234 (320)</td>
<td>1950 (388)</td>
</tr>
<tr>
<td></td>
<td>Mismatch</td>
<td>2218 (342)</td>
<td>1996 (344)</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>2316 (347)</td>
<td>2150 (362)</td>
</tr>
</tbody>
</table>

Note: standard deviations reported in parentheses.

The Mauchly’s test of sphericity showed that the assumption of sphericity was not violated \( (p>0.05) \)

There was a significant main effect of Format \( F(1,26)=26.076, p<.001, \eta^2=0.501 \).

There was a significant main effect of Similarity \( F(2,52)=28.303, p<.001, \eta^2=0.521 \).

There were significant interactions between Format and Similarity \( F(2,52)=16.848, p<.001, \eta^2=0.393 \), as well as Similarity and Item Pair \( F(2,52)=5.747, p=0.006, \eta^2=0.181 \).
To investigate the interactions further, a separate one-way ANOVA was conducted for each Pair Type (Video-Sound and Sound-Video) with the within-subject factor of Item Type (Similar, Dissimilar, and English-as-Foreign). For the Sound-Video one way ANOVA, there was a significant main interaction between Similarity and Pair Type $F(2,52)=8.440$, $p=0.001$, $\eta^2= 0.245$. Planned paired comparisons showed that Dissimilar matched pairs resulted in quicker response rates compared to mismatched pairs $t(26)=4.199$, $p<.001$ but no such interaction was observed for other item types (all $p>0.01$).

In contrast, a one way ANOVA for the Video-Sound Pair type only showed a significant main effect of Similarity $F(2,52)=39.432$, $p<.001$, $\eta^2= 0.603$. Planned pairwise comparisons showed that in the Video-Sound condition, Dissimilar items were responded quicker compared to other types of items $t(26)=4.909$, $p=0.01$, regardless whether items were matched or mismatched. No such interaction was evident for other item types (all $p>0.01$).

4.5.3 Discussion

The results showed a high degree of difficulty when participants are shown a speaker pronouncing a sound followed by hearing the sound (Video-Sound). One is that when participants see the face of the speaker, there is insufficient information for generation/translation of an image into a corresponding sound. Indeed, previous findings in the area of lip-reading show that discrimination between native and non-native sentences is easier when sentences are longer (Soto-Faraco et al., 2007). The current experiment employed single vowels/syllables, which do not carry other sources of information (e.g. supra-segmental, rhythmic, semantic, etc.). Ronquest and colleagues (Ronquest et al., 2010) reported that monolingual and bilingual speakers can discriminate similar Spanish-Catalan language pairs on the basis of visual information alone but they presented single words and sentences that contain far more linguistic information than a single syllable.

Other reasons for the unexpected results are that the quality of visual stimuli was suboptimal. It is also possible that the generation/translation process was weak due to a lack of linguistic information and that the auditory input disrupted the matching process, perhaps due to poor quality. This seems unlikely, though, given that previous studies (Soto-Faraco et al., 2007, Ronquest et al, 2010) report that lip-reading is a challenging task and it is therefore likely that the present task had carried a high degree of difficulty.
Finally, it is possible that the presented visual image afforded alternative auditory sounds as potential candidates for response. Subsequently, when only one sound was heard, possibly a non-native sound, it was necessary to translate that sound back to the visual image and in turn map it back to the auditory sound. Such a process may result in reliance on immediate visual memory and this is demanding of cognitive resources and susceptible to disruption. From a theoretical perspective, the latter account would be consistent with the Reverse Hierarchy Theory (RHT) of perceptual learning (Ahissar & Hochstein, 2004; Ahissar et al., 2009; Hochstein & Ahissar, 2002). This theory assumes that non-native speech perception arises on the basis of immediate high-level representations unless mapping of low-level to high-level representation is absent, which conceivably could be the case with non-native speech. RHT assumes that in such situations, re-mapping of lower-level input representations to a higher representational level occurs before a perceptual task is complete. The remapping is referred to as perception with scrutiny or backwards search, which would result in dissimilar lower-level input representations being remapped to a common high-level representation or vice versa.

One interesting aspect of performance in the Video-Sound condition was that Dissimilar Russian items were responded to quicker in comparison to other types of items, suggesting that some visual aspects of non-native speech were extracted but that information was not sufficient for an accurate response.

While performance in the Video-Sound condition was at chance, it is informative to inspect the pattern of responses to different types of stimuli. Unexpectedly, the sounds that are similar between English and Russian resulted in better performance than for dissimilar items. This pattern of responses to the similar items is a direct challenge to Hypothesis 2 of the study. These results contrast with previous studies, such as Kim and Davis (Kim & Davis, 2014) who argued that providing a visual cue re-sets the auditory system for the processing of auditory stimuli. They argued that this re-set increases sensitivity to auditory input. The present findings suggest that resetting may not be as efficient for the processing of non-native speech.

The results do support the hypothesis 1 that the Sound-Video condition would be easier than the Video-Sound condition. However, the overall error rates for the Sound-Video condition were still high, although better than chance. This suggests that when participants hear a sound they form a visual representation of that sound or alternatively generate an abstract category, although the process may not always be accurate, especially for non-native speech that is similar to English. This suggestion resonates with the conclusion of Kaganovich et al (Kaganovich et al., 2016). They also proposed that the matching process recruits mental
Visual perceptual assimilation of non-native speech

visualisation of articulatory organs mapped to silent input. They further suggested that auditory working memory would have to be actively recruited in order to match the auditory (or sub-vocal) input with visual articulation. It is also conceivable that there are differences in allocation of cognitive resources between competing input streams depending on whether the input is native or not. Indeed, Russian dissimilar sounds resulted in quicker responses, consistent with Hypothesis 3. Differential allocation of higher-order cognitive resources reflects top-down processing involved in the task. The results of the current experiment suggest that processing of the top-down stream may also be active when processing a non-native language.

In sum, the results of the current experiment highlight a degree of task difficulty in matching paradigms, evident in relatively high error rates. One explanation for such a high number of errors is insufficient information afforded by vowels and syllables, which rarely ever occur in isolation in natural speech. Therefore, it is necessary to investigate whether the effects found so far are also observed in more ecologically valid contexts. The key assumption is that more information will be available for processing speech presented in words and sentences. The experiments reported in the following chapter were designed to test this assumption.

4.6 Summary of Chapter 4
Chapter 4 presented the results from Experiments 1-5. The following is a brief summary of the experiments’ results. Experiment 1 showed limited evidence of visual perceptual assimilation of non-native speech, only present for the most dissimilar Russian items in the VO mode. However, there were high error rates for English items, not predicted a priori. There was limited AV benefit effect in Experiment 1. Experiment 2 utilised experimentally derived stimuli sub-set and failed to show a visual perceptual assimilation effect. Responses in the VO mode were at chance regardless of item type in Experiment 2. Experiment 3 utilised a same-different paradigm. The results showed that the Russian items which were the most dissimilar to English resulted in better performance, especially when they were contrasted with the corresponding English sounds. There was a limited AV benefit effect, with participants making a more accurate decision for whether items were the same in the AV mode compared to AO. Experiment 4 addressed the question of how non-native items are perceived in terms of their similarity to the native English language. Participants performed item rating and the results showed that participants judged items differently in AO and AV modes, specifically they judged items as better English items in the AO condition. To investigate the cause of these differences, Experiment 5 utilised a match-mismatched paradigm. The results showed a high degree of difficulty for the Video-Sound condition.
whilst the Sound-Video condition was better than chance. The following Chapter 5 examines whether similar processing of non-native speech occurs at the level of words and sentences.
Chapter 5: Perception of words and sentences.

This Chapter summarises the final three experiments conducted in the study. The experiments described are all concerned with AV perception of words and sentences in an unfamiliar non-native language. The rationale, method, results and discussion of each experiment can be found in this Chapter. Speech perception is tested in auditory only (AO), visual only (VO) and auditory-visual (AV) conditions. The expectation is that non-native AV speech is perceived using native speech categories leading to the prediction that perceptually dissimilar items should be more salient and therefore will be easier to identify than perceptually similar items. In contrast to the first six experiments, which tested minimal phonetic features, the goal of these experiments is to provide a more ecologically valid test of the hypothesis. The specific hypothesis is that the visual aspects of non-native speech (Russian) will be assimilated into native language (English) categories during word and sentence processing. The prediction that follows is that the visual aspects of Russian that are dissimilar to English will be identified more readily than visual aspects of Russian that are similar to English during speech perception.
5.1 **Experiment 6: Language Identification of Words and Sentences**

One inference from results so far is that visual aspects of non-native speech, even for the most dissimilar items are not recognised reliably during visual only presentation contrary to expectation. Moreover, there is no reliable AV benefit effect at the single vowel/syllable level. However, it is possible that visual information is not sufficient for a benefit to be realised at the level of a single vowel or syllable using the utilised experimental paradigms. One reason may be that there is insufficient time and too little information for participants to notice distinctive visual features and to evoke native speech category prototypes. Prior studies demonstrate that increasing stimulus length facilitates native and non-native speech identification for VO speech (Ronquest et al., 2010) and it is possible to identify the language of presented sentences based on VO input (Soto-Faraco et al., 2007). The goal of the following experiments is to interrogate the effects of visual input on speech perception at the word and sentence levels.

The first experiment uses written words as a visual prime. The underlying assumption is that written words would evoke native prototypes, which could in turn be mapped onto native and non-native target stimuli. Assuming that written words evoke articulatory and visual features, the experimental design therefore also allows a test of the Motor theory of speech perception (Liberman & Mattingly, 1989). One prediction that can be derived from Motor theory is that identifying a spoken target word as native or non-native in the AV paradigm should be better when a written word cue is provided compared to when it is not. This prediction was derived from studies of cross-modal priming, which show that watching a speaker articulating words silently or seeing a printed prime word facilitates subsequent identification of the target word presented in auditory modes (Buchwald et al., 2009; Kim, Davis, & Krins, 2004; van der Zande et al., 2014). Similar results were obtained at the sentence level for native speech perception (Lansing & McConkie, 2003). Moreover, if our speculation above about insufficient information constraining AV effects is correct, it would be expected that AV and VO performance with sentences would be superior to performance with written words alone. Speech perception in the present context is not limited to morphological or phonological information and might include multiple inputs such as lexical, segmental, temporal, rhythmic cues etc. that are assumed to facilitate speech perception. Furthermore, ecological validity at the word and sentence level will be increased. Several researchers claim the language discrimination paradigm is a more ecologically valid task and argue for more indirect measures instead of direct measures (Soto-Faraco et al., 2007). The experiment also uses English-Russian cognate words, which are assured to give access to the lexicon during the
processing of non-native speech, even in the initial stages of vocabulary acquisition (e.g. (Desmet & Duyck, 2007). In Experiment 6 participants are presented with stimuli (words and sentences with or without primes) in either English or Russian and asked whether it is English or not. The dependent variables are again error rates and response times whilst independent variables are modality (Video Only Word, Video Written Word prime, Audio-Visual Word, Video Only Sentence, Video Written Sentence prime) and Language (English and Russian).

Experiment 6 Hypotheses:
Hypothesis 1: Performance with sentences will be better (more accurate and quicker) than that with words;
Hypothesis 2: Performance with written primes will be better than that without the primes;

5.1.1 Method.

Participants: Nineteen undergraduate psychology students from The University of Melbourne who received course credit. There were 13 females. Mean age of participants was 19 years. All were native speakers of English and none had any knowledge of Russian language. All reported having no known hearing impairment and normal or corrected-to-normal vision.

All participants completed 6 conditions as described below in the following order:

Condition 1a: Video-Only word presentation
Materials and Procedure: Stimuli were Russian and English spoken cognates. Cognates were defined as spoken words with the same linguistic derivation and meaning but not necessarily identical in pronunciation in the two languages. There were 40 cognates for each language, with 36 experimental stimuli and 4 example stimuli. Before each cognate presentation the word “ready” was briefly flashed on screen, signalling the beginning of a presentation. Each word was repeated twice over the course of the experiment. If no response was given, the subsequent item was automatically presented. No feedback was provided. Participants read a set of instructions that explained how a bilingual person would speak cognate words in English and another foreign language. Russian language was not identified. It was further explained that all video images and sounds would be separated and that in this condition only video images would be presented without the corresponding sound. Verbal clarification of instructions was given when necessary. Four examples were provided and participants practiced making a response. They were instructed to press the “YES” button if they thought the speaker was saying a word in English or “NO” button if they thought the speaker was
saying a word in a foreign language. The participants were also instructed that each item would be presented several times over the course of the experiment.

**Condition 1b: Video-Written word presentation**

*Materials and Procedure:* As in Condition 1a except that a printed word appeared on screen prior to participants seeing the face of a speaker. The instructions were that the written word could be followed by an image of the speaker saying the word in English or the foreign language.

**Condition 1c: Audio-Visual word presentation**

*Materials and Procedure:* As in Condition 1a except in the AV mode.

**Condition 2a: Video-Only sentence presentation**

*Materials and Procedure:* Stimuli were Russian and English sentences. The English sentences are reported in the Appendix A. Russian sentences included idiomatic translation of English sentences. The number of words was kept constant across languages and mean exposure time for English and Russian languages did not differ. There were 12 sentences in each language (10 experimental and 2 practice sentences). Each sentence was repeated three times. The procedure was identical to Condition 1a except participants were told that they would see a speaker saying sentences (rather than words). They were instructed to press a “YES” button if they thought that the speaker was saying a sentence in English, and the “NO” button if they thought that a speaker was saying a sentence in a foreign language.

**Condition 2b: Video-Written sentence presentation**

*Materials and Procedure:* As in Condition 2a except that a printed sentence appeared on screen prior to participants seeing the face of a speaker. The instructions explained that the sentence could be followed by an image of the speaker saying this sentence in English or in the foreign language.

**Condition 2c: Audio-Visual sentence presentation**

*Materials and Procedure:* As in Condition 2a, except the sentence was presented in the audio-visual mode. Note that this condition was subsequently eliminated from analysis because of a ceiling effect with participants showing near perfect performance.
5.1.2 Results

A two-way within subjects ANOVA was performed. The first factor Modality had five levels: Video Only Word, Video Written Word prime, Audio-Visual Word, Video Only Sentence, Video Written Sentence prime. The second factor Language had two levels: English and Russian.

Table 13 shows descriptive statistics for all VO conditions.

Table 13. Mean Percentage of Error Rates for Visual Only (VO) With and Without Written Primes Conditions in Experiment 6.

<table>
<thead>
<tr>
<th>Modality</th>
<th>Russian Mean (sd)</th>
<th>English Mean (sd)</th>
<th>Mean (sd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VO Word</td>
<td>50.95 (8.98)</td>
<td>50.43 (4.90)</td>
<td>50.69 (6.94)</td>
</tr>
<tr>
<td>VO Word with written word</td>
<td>42.98 (9.73)</td>
<td>26.60 (9.45)</td>
<td>34.79 (9.59)</td>
</tr>
<tr>
<td>VO Sentence</td>
<td>69.95 (6.53)</td>
<td>51.92 (14.50)</td>
<td>60.94 (10.52)</td>
</tr>
<tr>
<td>VO Sentence with Printed sentence</td>
<td>21.04 (11.12)</td>
<td>4.20 (3.12)</td>
<td>12.62 (7.12)</td>
</tr>
</tbody>
</table>

Note: standard deviations reported in parentheses.

Table 14 shows descriptive statistics for AV conditions:

Table 14. Mean Percentage of Error Rates for Audio Visual(AV) Conditions in Experiment 6.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Mean (sd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AV word English</td>
<td>2.13 (3.01)</td>
</tr>
<tr>
<td>AV word Russian</td>
<td>8.26 (6.55)</td>
</tr>
<tr>
<td></td>
<td>5.19(4.78)</td>
</tr>
</tbody>
</table>

Note: standard deviations reported in parentheses.

There was a significant main effect of Modality $F(4,72) = 318.385, p < 0.01, \eta^2 = 0.946$

There was a significant main effect of Language $F(1,18) = 85.078, p < .001, \eta^2 = 0.825$.

The analysis showed a significant interaction between Modality and Language $F(4,72) = 8.99, p < .001, \eta^2 = 0.333$. 
To investigate interaction further, a separate one-way ANOVA was conducted for each language, with within-subject factors being 5 types of Modality.

A one-way ANOVA for the English language with the Mauchly test of sphericity indicated that the assumption of sphericity had been violated ($p<0.05$). Therefore, the Greenhouse-Geisser corrections were applied.

There was a main effect of Modality $F(2.109, 37.965)=205.296, p<.001, \eta^2_p=0.919$. Planned paired comparisons showed that consistent with Hypothesis 2, performance for English conditions with the printed prime was better than that without the prime. Specifically, participants were significantly more accurate for the VO with printed word condition than VO word without the prime condition $t(18)=12.76, p<.001$. Similarly, performance at the VO sentence level, was more accurate when a written sentence prime was presented first compared to VO sentence without the prime $t(18)=15.393, p<.001$. For the English stimuli, performance in the AV word condition was significantly better than that in VO $t(18)=39.70, p<.001$. For the English stimuli, there was no difference between participants’ performance for VO sentence and VO word conditions, contrary to Hypothesis 1 (length of input). Moreover, both conditions without a written prime (VO sentence and VO word) were not significantly different from 50% chance level ($p>0.05$).

A one-way ANOVA for the Russian language with Modality as within-subject factors was also conducted. The Mauchly test of sphericity indicated that the assumption of sphericity had not been violated ($p>0.05$).

There was a significant main effect of Modality $F(4,72)=143.004, p<.001, \eta^2_p=0.888$. Planned paired comparisons showed that consistent with Hypothesis 2 performance with the printed prime was better than that without the prime. Specifically, performance with the Russian words presented with a prime were significantly more accurate that those without the prime $t(18)=2.743, p=0.013$. Similarly, responses to the Russian sentences with the printed sentence as a prime were significantly more accurate that those without the primes $t(18)=17.393, p<.001$. Again, performance in the AV word condition was significantly more accurate compared to that in the VO word condition $t(18)=15.480, p<.001$.

Against the Hypothesis 1, performance for the Russian VO sentences was at chance rather than being significantly better than VO word.

Comparisons between English and Russian VO sentences showed that both were at chance (one-sample t-test $p>0.05$) whilst planned paired comparison between languages showed that participants made more errors for the Russian items compared to the English items $t(18)=9.224, p<.001$. 


Response time analysis:
A two-way within subjects ANOVA was performed. The first factor Modality had five levels: Video Only Word, Video Written Word prime, Audio-Visual Word, Video Only Sentence, Video Written Sentence prime. The second factor Language had two levels: English and Russian.

Table 15 shows descriptive statistics for all VO conditions.

<table>
<thead>
<tr>
<th>Modality</th>
<th>Russian</th>
<th>English</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>VO Word</td>
<td>2581 (550)</td>
<td>2581 (366)</td>
<td>2581 (458)</td>
</tr>
<tr>
<td>VO Word with written word</td>
<td>2264 (319)</td>
<td>2222 (361)</td>
<td>2243 (340)</td>
</tr>
<tr>
<td>VO Sentence</td>
<td>4236 (219)</td>
<td>4134 (208)</td>
<td>4185 (213)</td>
</tr>
<tr>
<td>VO Sentence with Printed sentence</td>
<td>4141 (337)</td>
<td>3616 (194)</td>
<td>3878 (266)</td>
</tr>
<tr>
<td></td>
<td>3306 (356)</td>
<td>3138 (282)</td>
<td>3222 (319)</td>
</tr>
</tbody>
</table>

Note: standard deviations reported in parentheses.

Table 16 shows descriptive statistics for AV conditions:

<table>
<thead>
<tr>
<th>Condition</th>
<th>Mean (sd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AV word English</td>
<td>2043 (378)</td>
</tr>
<tr>
<td>AV word Russian</td>
<td>2044 (248)</td>
</tr>
<tr>
<td>AV word Russian</td>
<td>2044 (313)</td>
</tr>
</tbody>
</table>

Note: standard deviations reported in parentheses.

There was a significant main effect of Modality $F(4,72) = 314.194, p < 0.01, \eta^2 = 0.946$
There was a significant main effect of Language $F(1,18) = 23.552, p < .001, \eta^2 = 0.567$.
The analysis showed a significant interaction between Modality and Language $F(4,72) = 16.756, p < .001, \eta^2 = 0.482$.
To investigate interaction further, a separate one-way ANOVA was conducted for each language, with within-subject factors being 5 types of Modality.
A one-way ANOVA for the English language with the Mauchly test of sphericity indicated that the assumption of sphericity had not been violated ($p > 0.05$). There was a main effect of Modality $F(1,18)=290.022$, $p<.001$, $\eta^2_p= 0.942$. Planned paired comparisons showed that consistent with Hypothesis 2, response times for conditions with the printed prime were quicker than without the prime. Specifically, participants were significantly quicker for the VO with printed word condition than VO word condition $t(18)=4.214$, $p=0.001$. Similarly, response times at the VO sentence level were significantly quicker when a written sentence prime was presented first compared to VO sentence without the prime $t(18)=11.381$, $p<.001$. For the English stimuli, participants’ response times for VO sentence was significantly slower than in the VO word conditions $t(18)=21.446$, $p<.001$, contrary to Hypothesis 1 (length of input).

A one-way ANOVA for the Russian language with Modality as within-subject factors was also conducted. There was a significant main effect of Modality $F(1,18)=103.136$, $p<.001$, $\eta^2_p= 0.851$. Planned paired comparisons showed that consistent with Hypothesis 2 response times for the conditions with the printed prime were quicker than those without the prime. Specifically, performance with the Russian words presented with a prime were significantly quicker than that those without the prime $t(18)=2.993$, $p=0.008$ but there was no difference in response times at the level of sentences ($p>0.05$).

5.1.3 Discussion

The goal of the experiment was to investigate ability to process native and non-native visual speech at word and sentence levels rather than at the level of isolated vowels/syllables. The underlying logic of the current experiment was greater ecological validity of natural speech at the word and sentence levels, which allows for extraction of multiple aspects of speech. It was hypothesised that performance in the written word prime conditions would be better than that in the conditions without a written cue. The results supported this hypothesis by showing that performance with written primes was superior (both more accurate and quicker) to VO without primes presentation. Moreover, performance in the VO conditions without a written prime was no better than chance, showing participants could not discriminate visually presented words and sentences as native or not native.

The results contrast with findings of Soto-Faraco et al (Soto-Faraco et al., 2007) and Ronquest et al (Ronquest et al., 2010). In the study of Soto-Faraco et al (2007), the researchers observed discrimination of native and non-native languages based on visual-only (speech-reading) aspects alone using a same/different paradigm whereby they presented sentences in Catalan.
and/or Spanish and asked bilingual Catalan/Spanish participants whether they were the same or not. Critically, participants were bilingual speakers exposed to both languages on a daily basis. Participants without knowledge of the two languages were unable to perform the task. However, Spanish monolingual speakers who were not familiar with Catalan were able to do so, albeit not as accurately as Catalan/Spanish bilinguals. We found that monolingual English speakers who were not familiar with Russian language could not perform the task better than chance. One difference between studies is that the languages used in the study of Soto-Faraco et al (2007), namely Spanish and Catalan, are similar in terms of their phonological properties, rhythmic pattern, and lexicon. It is therefore likely that monolingual speakers were able to discriminate native and non-native languages because of high language familiarity and saliency of features between languages, such as subtle rhythmic differences. Our monolingual participants had no such advantages. Therefore, we contend that language exposure and language type are equally likely to deliver a benefit in non-native speech perception and that our results stem from lack of exposure to Russian in monolingual Australian English speakers. It is possible that the types of language (Russian is a Slavic language and English is a Germanic language) also had an effect on the present findings. However, we note from Chapter 3 that English and Russian have many phonological features in common (albeit with some differences).

The issue of language type effects on non-native speech perception was address in a study by Ronquest, Levi, and Pisoni (2010), which replicated the results of Soto-Faraco et al (2007) with English and Spanish bilinguals and monolinguals. Unlike the study of Soto-Faraco et al (2007) however, they used a different paradigm - language identification - asking participants if stimuli in the visual-only stimuli were Spanish or English. All the participants, including monolingual English speakers performed significantly better than chance. One explanation for differences with the present findings is that monolingual English-speaking US participants would have had greater knowledge of what Spanish (second most spoken language in US) sentences sound like compared to our Australian English speaking sample with virtually no exposure to Russian i.e. Russian is absent from the frequently spoken languages in Australia.

An additional difference between the present study and Soto-Faraco et al (2007) is that they used long and short (16-syllable utterances) sentences and single words as opposed to single cognate words and sentences used in the current study. It is likely that use of single cognate words is more challenging than that of 16-syllable short sentences or non-cognate words. This highlights the need for replication of the results of Soto-Faraco et al (2007) and Ronquest et al (2010) with other languages and experimental stimuli. It also potentially adds to research in
Visual perceptual assimilation of non-native speech

the domain of Russian second language use showing a reverse cognate effect (e.g. (Kouačević, 2012; Temnikova & Nagel, 2015).

Consistent with previous findings, the results showed that participants made more errors with non-native words and sentences compared to native stimuli. This result is consistent with the assumption that native language categories constrain non-native speech processing, resulting in misidentification of non-native items as native. Moreover, participants made more false positive errors with non-native items when no written primes were available, suggesting a possible bias to respond “English” when presented with Russian visual stimuli. The latter possibility is less likely when information is less ambiguous, such as with the benefit from a native written word cue, even when processing an unknown language. One explanation is that presenting a native printed word evokes a visual prototype or set of potential exemplars of that word, in turn forcing matching by comparing the non-native visual prototype to the native evoked template/exemplar. Slower reaction times for English sentences in VO condition compared to processing of cognates words in VO also points out to the fact that ‘perception with scrutiny’ was possibly used. No such difference in the speed of responses was observed for the Russian sentences and words, suggesting that the decision was possibly based upon a simpler algorithm.

One interesting result from the pilot study was that participants performed at ceiling when presented with sentences in the AV mode. This extends the findings of Soto-Faraco (2007) and Ronquest et al (2010) that did not include an AV component and suggests that hearing a non-native sentence disambiguates visual information and provides sufficient entropy to make an accurate identification. It is likely that the cumulative effects of hearing non-native words in a sentence allows time to build a cohort when compared to a more limited input for single word processing, due to more intonation, rhythm, lexical content at the sentence than word level.

To summarise, the results show that participants have difficulty identifying the language of native and non-native cognate words and sentences when they are presented in the VO mode. One possibility is that using a normal rate of presentation, participants ignored or assimilated visual aspects of non-native speech readily. One way of testing this possibility would be to focus attention to salient aspects of non-native visual speech, through temporal manipulation. Since visual representation of non-native speech is the focus of the study, the cross-modality matching paradigm was used again in Experiment 7.
5.2 Experiment 7: Matching

So far is there is no additional evidence that visual aspects enhance the perception of non-native speech when amount of information is increased. Prior studies demonstrate that audio-visual speech enhances perception of speeded-up speech (eg.(Massaro et al., 1998) but time reversed stimuli do not. Indeed, several studies use time-reversed speech as a control condition as it retains all but the lexical properties of speech (Conrey & Pisoni, 2006; Kim & Davis, 2004; L. Lachs & Pisoni, 2004). The goal of the present experiment was to investigate whether rate of speech presentation impacts matching of auditory and visual stimuli when processing native and non-native speech. Speed of presentation was manipulated to calibrate the effects of task difficulty on performance. The first hypothesis was that fast presentation would result in more errors whilst slow presentation will allow more time for the identification of visual aspects of non-native speech. Time reversed stimuli were used as a control condition, with native and non-native stimuli generating an equivalent rate of errors.

We note however that when Ronquest et al. (Ronquest et al., 2010) presented native English speakers with time reversed non-native Spanish words and sentences, participants could identify non-native items in the VO mode. The researchers argued that this was possible because participants could discriminate languages from rhythmic information only. Therefore, we may see similar effects when native Australian English speakers see (unfamiliar) Russian. Similar to the previous results in the cross-matching paradigm, it was hypothesised that matching sound-to-video will be easier than video-to-sound. In addition, matching native stimuli will be easier than foreign. Finally, performance for congruent within-language matching stimuli (e.g. English-with English and Russian-with-Russian) will be easier than for the mismatched stimuli.

The dependent variables in Experiment 7 were again response error rates and response times. The dependent variables were speed of stimuli presentation (quick, slow, back, and normal), modality (sound-video and video-sound), match type (match and mismatch), language match (eg. English with English; Russian with Russian), or language Mismatch (English with Russian, Russian with English).

Experiment 7 Hypotheses:
1. Slow speed of presentation will result in better performance for Russian stimuli than normal speed. In turn, quick speed will be worse than normal. Backwards will serve as a control and will result in the worst performance. (Slow>Normal>Quick>Backwards)
2. Performance of Sound-Video condition will be better than Video-Sound.
3. Performance for matched pairs (English-English, Russian-Russian) will be better than mismatched pairs (English-Russian, Russian-English).
5.2.1 Method

**Participants:** A different group of participants to that used previous experiments was used in Experiment 7. Seventeen undergraduate psychology students from the University of Melbourne participated and received course credit for participation. There were 13 females. The mean age of participants was 19 years. All were native speakers of English and none had prior knowledge of the Russian language. All reported having no known hearing impairment and normal or corrected-to-normal vision.

Each participant participated in all of the following conditions:

- **Condition 1a:** Matching Sound-Video Quick
- **Condition 1b:** Matching Video-Sound Quick
- **Condition 2a:** Matching Sound-Video Slow
- **Condition 2b:** Matching Video-Sound Slow
- **Condition 3a:** Matching Sound-Video Back
- **Condition 3b:** Matching Video-Sound Back
- **Condition 4a:** Matching Sound-Video Normal
- **Condition 4b:** Matching Video-Sound Normal

Condition 1a: Matching Sound-Video Quick and Condition 1b. Matching Video-Sound Quick:

**Materials and Procedure:** Stimuli consisted of cognate words used in Experiment 6. These files were processed in Adobe Premier so that speed of presentation was twice the original condition (200%). The resulting AVI files were then transformed into MPG files to reduce size. The files were also separated into audio and video files. The procedure was identical to Experiment 5 except that the participants were notified that the stimuli would be played quickly.

Condition 2a: Matching Sound-Video Slow and Condition 2b: Matching Video-Sound Slow.

**Materials and Procedure:** Stimuli (cognate words) were the same as those used in Experiment 6, but files were modified in Adobe Premier so that the speed of presentation was slower (70% of original speed). Again, the files were separated into the audio and visual files. The procedure was identical to Experiment 5 except that the participants were notified that stimuli would be played slowly.
Condition 3a: Matching Sound-Video Back and
Condition 3b: Matching Video-Sound Back.

Materials and Procedure: The stimuli from Experiment 6 were used but original files were modified in Adobe Premier so they would be played backwards. AVI files were separated into audio and visual files. The procedure was identical to Experiment 5 except that participants were notified that the stimuli would be played backward.

Condition 4a: Matching Sound-Video Normal and
Condition 4b: Matching Video-Sound Normal.

Materials and Procedure: The stimuli from Experiment 6 were used without any speed modification. AVI files were separated into audio and visual files. The procedure was identical to Experiment 5. The order of conditions 1, 2, 3 was randomised across participants, while condition 4 was administered last. This was done because the stimuli in the first 3 conditions were modified (quicker, slower or played backwards) while stimuli in Condition 4 were not modified. It was reasoned that presenting the stimuli at the original/normal speed might help performance in other conditions. Within each condition, the sound-video and video-sound presentations (versions a and b) were randomised across participants.

5.2.2 Results
A four-way within-subjects ANOVA analysis was employed. The first factor Speed had 4 levels (quick, slow, back, and normal). The second factor Modality had 2 levels (sound-video and video-sound). The third factor Match also had 2 levels (Match and Mismatch). Finally, the fourth factor Language had 2 levels: Language Match (eg. English with English; Russian with Russian), or Language Mismatch (English with Russian, Russian with English). The between-subject factor of Order of presentation was not significant and the analysis below is reported without including that factor.

Table 17 and Table 18 show descriptive statistics for errors in the Video-Sound and Sound-Video conditions respectively.

Table 17. Error Rates for Sound-Video Condition in Experiment 7.

<table>
<thead>
<tr>
<th>Match</th>
<th>Sound-Video</th>
<th>Mean</th>
<th>SD</th>
<th>Back</th>
<th>Quick</th>
<th>Slow</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-English</td>
<td>Back</td>
<td>39.16 (8.27)</td>
<td>32.04 (16.49)</td>
<td>37.74 (13.54)</td>
<td>21.08 (8.40)</td>
<td>32.51 (11.67)</td>
<td></td>
</tr>
<tr>
<td>Match</td>
<td>Back</td>
<td>31.38</td>
<td>41.65</td>
<td>39.71</td>
<td>22.38</td>
<td>33.78</td>
<td></td>
</tr>
</tbody>
</table>
The Mauchly test of sphericity indicated that the assumption of sphericity had not been violated \((p > 0.05)\).

There was a significant main effect of Speed \(F(3,48)=16.932, \quad p<.001, \quad \eta_p^2=0.514\).

There was a significant main effect of Match \(F(1,16)=11,445, \quad p<0.004, \quad \eta_p^2=0.417\), with participants being more accurate when stimuli were matched compared to when they did not match \(t(16)=3.38, \quad p<.001\) There was a significant main effect of Language \(F(1,16)=7.029, \quad p<.001\), \(\eta_p^2=0.305\), with participants being significantly more accurate with English stimuli compared to the Russian stimuli \(t(16)=3.057, \quad p=0.008\)

There was a significant three-way interaction between Modality, Speed, and Match \(F(3,48)=10.325, \quad p<.001, \quad \eta_p^2=0.392\), and between Modality, Speed, and Language \(F(3,48)=16.015, \quad p<.001, \quad \eta_p^2=0.500\).
To investigate three-way interactions further, separate 3-way ANOVAs were conducted for each Modality (Sound-Video and Video-Sound). The first factor Speed had 4 levels (back, quick, slow, and normal), the second factor Match had 2 levels (match, mismatch), and the third factor Language had 2 levels (English and Russian).

The results for the separate Sound-Video (SV) ANOVA showed that there was a main effect of Speed $F(3,48)=8.992$, $p<.001$, $\eta^2=0.360$ and Match $F(1,16)=15.465$, $p=0.01$, $\eta^2=0.491$.

There was a three-way interaction between Speed, Match, and Language $F(3,48)=3.025$, $p=0.038$, $\eta^2=0.159$. To investigate this interaction further a separate two-way ANOVA was performed for English and Russian separately, with Match having 2 levels (match mismatch) and Speed having four levels (back, quick, slow, and normal). For the SV English ANOVA there was a significant main effect of Speed $F(3,48)=10.116$, $p<.001$, $\eta^2=0.387$. There was a significant main effect of Match $F(1,16)=14.881$, $p=0.001$, $\eta^2=0.482$.

There was an interaction between Speed and Match $F(3,48)=10.348$, $p<.001$, $\eta^2=0.393$ Planned paired comparisons showed that participants were significantly more accurate for the SV matched English items (English-English) compared to the mismatched items (English-Russian) but only when the speed was normal ($t(16)=8.167$, $p<.001$) or backwards ($t(16)=4.022$, $p=0.001$).

A separate ANOVA for the SV in Russian showed that there was a significant main effect of Speed $F(3,48)=12.936$, $p<.001$, $\eta^2=0.447$. There was also a significant main effect of Match $F(1,16)=13.108$, $p=0.002$, $\eta^2=0.450$. Planned paired comparisons showed that participants were significantly more accurate for the SV Russian matched stimuli when it was played backwards $t(16)=2.331$, $p=0.033$, quick $t(16)=2.689$, $p=0.016$, and in normal speed $t(16)=4.179$, $p=0.001$ but not for slow speed ($p>0.05$).

The results for the separate Video-Sound (VS) ANOVA showed that there was a main effect of Speed $F(3,48)=7.353$, $p<.001$, $\eta^2=0.315$ and Language $F(1,16)=8.188$, $p=0.011$, $\eta^2=0.315$.

For the separate VS ANOVA, there was a two-way interaction of Speed and Match $F(3,48)=4.745$, $p=0.006$, $\eta^2=0.229$ and Speed and Language $F(3,48)=4.902$, $p=0.005$, $\eta^2=0.235$. To investigate these interactions further a separate two-way ANOVA was conducted for each language (English and Russian) with the first factor Speed having four levels (back, quick, slow, and normal) and the second factor Match having 2 levels (Match and Mismatch).
A separate VS ANOVA for English showed a significant main effect of Speed $F(3,48)=3.842$, $p=0.015$, $\eta^2_p=0.194$. There was a significant interaction between Speed and Match $F(3,48)=3.193$, $p=0.032$, $\eta^2_p=0.166$. Planned paired comparisons showed that participants were significantly more accurate for VS Match (English-English) stimuli compared to mismatched (English-Russian) when it was played in the slow $t(16)=2.104$, $p=0.052$ and normal speed $t(16)=2.076$, $p=0.054$ but not in the backwards and quick speed (all $p>0.05$).

A separate VS ANOVA for Russian showed a significant main effect of Speed $F(3,48)=4.585$, $p=0.007$, $\eta^2_p=0.223$. There was a significant interaction between Speed and Match $F(3,48)=4.154$, $p=0.011$, $\eta^2_p=0.206$. Planned paired comparisons showed that participants were significantly more accurate for VS Match (Russian-Russian) stimuli compared to mismatched (Russian-English) when it was played in the normal speed $t(16)=2.839$, $p=0.012$ but not in the slow, backwards, or quick speed (all $p>0.05$). Moreover, a one-sample test comparing responses to the VS Russian stimuli in the slow speed of presentation to the chance level (50%) showed that they were no different to chance ($p>0.05$).

To summarise, to investigate two significant 3 way interactions (Modality, Speed, and Match; Modality, Speed, and Language), separate 3-way ANOVAs were conducted for each Modality (Sound-Video and Video-Sound). The first SV ANOVA showed a three-way interaction between Speed, Match, and Language. The second VS ANOVA showed a two-way interaction of Speed and Match, as well as and Speed and Language.

Further planned paired comparisons were conducted to address the proposed hypotheses. Specifically, performance in the Sound-Video modality was overall more accurate than that in the Video-Sound modality $t(16)=4.605$, $p<.001$.

In addition, planned paired comparisons showed that the overall performance for the stimuli played in the normal speed was more accurate than that for the slow $t(16)=4.953$, $p<.001$ and quick speed $t(16)=6.097$, $p<.001$.

Response Time Analysis

A four-way within-subjects ANOVA analysis was employed. The first factor Speed had 4 levels (quick, slow, back, and normal). The second factor Modality had 2 levels (sound-video and video-sound). The third factor Match also had 2 levels (Match and Mismatch). Finally, the fourth factor Language had 2 levels: Language Match (eg. English with English; Russian with Russian), or Language Mismatch (English with Russian, Russian with English).
Table 19 and Table 20 show descriptive statistics for all Response Times in the Sound-Video and Video-Sound conditions respectively.

**Table 19. Response Times for Sound-Video Condition in Experiment 7.**

<table>
<thead>
<tr>
<th>Match</th>
<th>Back</th>
<th>Quick</th>
<th>Slow</th>
<th>Normal</th>
<th>Mean (sd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-English</td>
<td>2980</td>
<td>1649</td>
<td>2885</td>
<td>2207</td>
<td>2430 (361)</td>
</tr>
<tr>
<td>(327)</td>
<td>(549)</td>
<td>(321)</td>
<td>(245)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Match</td>
<td>2818</td>
<td>1574</td>
<td>2915</td>
<td>2210</td>
<td>2379 (377)</td>
</tr>
<tr>
<td>Russian-Russian</td>
<td>(318)</td>
<td>(559)</td>
<td>(400)</td>
<td>(232)</td>
<td></td>
</tr>
<tr>
<td>Mismatch</td>
<td>3064</td>
<td>1584</td>
<td>2811</td>
<td>2183</td>
<td>2410 (341)</td>
</tr>
<tr>
<td>English-Russian</td>
<td>(348)</td>
<td>(370)</td>
<td>(341)</td>
<td>(306)</td>
<td></td>
</tr>
<tr>
<td>Mismatch</td>
<td>3037</td>
<td>1878</td>
<td>3056</td>
<td>2321</td>
<td>2573 (353)</td>
</tr>
<tr>
<td>Russian-English</td>
<td>(267)</td>
<td>(545)</td>
<td>(329)</td>
<td>(272)</td>
<td></td>
</tr>
<tr>
<td><strong>Mean (sd)</strong></td>
<td>2975</td>
<td>1671</td>
<td>1167</td>
<td>2230</td>
<td>2448 (358)</td>
</tr>
<tr>
<td></td>
<td>(315)</td>
<td>(506)</td>
<td>(347)</td>
<td>(264)</td>
<td></td>
</tr>
</tbody>
</table>

Note: standard deviation reported in parentheses.

**Table 20. Response Times for Video-Sound Condition in Experiment 7.**

<table>
<thead>
<tr>
<th>Match</th>
<th>Back</th>
<th>Quick</th>
<th>Slow</th>
<th>Normal</th>
<th>Mean (sd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-English</td>
<td>3042</td>
<td>1787</td>
<td>3001</td>
<td>2365</td>
<td>2549 (385)</td>
</tr>
<tr>
<td>(360)</td>
<td>(457)</td>
<td>(433)</td>
<td>(290)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Match</td>
<td>2779</td>
<td>1788</td>
<td>3197</td>
<td>2443</td>
<td>2552 (387)</td>
</tr>
<tr>
<td>Russian-Russian</td>
<td>(335)</td>
<td>(507)</td>
<td>(378)</td>
<td>(329)</td>
<td></td>
</tr>
<tr>
<td>Mismatch</td>
<td>2998</td>
<td>1721</td>
<td>2904</td>
<td>2445</td>
<td>2517 (322)</td>
</tr>
<tr>
<td>English-Russian</td>
<td>(401)</td>
<td>(359)</td>
<td>(342)</td>
<td>(186)</td>
<td></td>
</tr>
<tr>
<td>Mismatch</td>
<td>3029</td>
<td>1746</td>
<td>3219</td>
<td>2525</td>
<td>2630 (363)</td>
</tr>
<tr>
<td>Russian-English</td>
<td>(405)</td>
<td>(506)</td>
<td>(332)</td>
<td>(211)</td>
<td></td>
</tr>
</tbody>
</table>
Visual perceptual assimilation of non-native speech

The Mauchly test of sphericity indicated that the assumption of sphericity had not been violated ($p > 0.05$).

There was a significant main effect of Modality $F(3,45)=10.914$, $p=0.005$, $\eta^2_p=0.421$, with participants being quicker for the SV modality than VS modality $t(16)=2.031$, $p=0.007$.

There was a significant main effect of Speed $F(3,45)=137.456$, $p<.001$, $\eta^2_p=0.902$.

There was a significant main effect of Match $F(1,15)=4.470$, $p=0.052$, $\eta^2_p=0.230$, with participants being quicker for the matched stimuli compared the mismatched $t(16)=2.045$, $p=0.008$.

There was a significant main effect of Language $F(1,15)=12.388$, $p=0.003$, $\eta^2_p=0.452$, with participants responding quicker to the native English stimuli compared to the Russian stimuli $t(16)=3.023$, $p<.001$.

There was a two-way interaction between Modality and Speed $F(3,45)=3.098$, $p=0.036$, $\eta^2_p=0.171$.

There was a two-way interaction between Speed and Language $F(3,45)=12.858$, $p<.001$, $\eta^2_p=0.462$.

There was a two-way interaction between Match and Language $F(1,15)=16.447$, $p=0.001$, $\eta^2_p=0.523$.

To investigate two-way interactions further, separate 3-way ANOVAs were conducted for each Modality (Sound-Video and Video-Sound). The first factor Speed had 4 levels (back, quick, slow, and normal), the second factor Match had 2 levels (match, mismatch), and the third factor Language had 2 levels (English and Russian).

The results for the separate Sound-Video (SV) ANOVA showed that there were significant main effects of Speed $F(3,48)=122.997$, $p<.001$, $\eta^2_p=0.885$, Match $F(1,16)=5.612$, $p=0.03$, $\eta^2_p=0.260$, and Language $F(1,16)=8.405$, $p=0.010$, $\eta^2_p=0.344$.

There was a two-way interaction between Match and Language $F(1,16)=27.798$, $p<0.01$, $\eta^2_p=0.635$. Planned paired comparisons showed that participants were slower with English matched stimuli compared to mismatched $t(16)=2.034$, $p=0.052$ but quicker with the matched Russian stimuli compared to the mismatched Russian stimuli when matching English stimuli compared that mismatched $t(16)=3.041$, $p=0.007$.

There was a two-way interaction between Speed and Language $F(3,48)=5.470$, $p=0.003$, $\eta^2_p=0.201$. Planned paired comparisons showed that participants were quicker with normal speed compared to quick speed $t(16)=2.034$, $p=0.052$ but slower with back speed compared to normal speed $t(16)=3.041$, $p=0.007$. There was a significant main effect of Language $F(1,16)=5.612$, $p=0.03$, $\eta^2_p=0.260$, with participants responding quicker to the English stimuli compared to the Russian stimuli $t(16)=3.023$, $p<.001$.

There was a two-way interaction between Modality and Speed $F(3,45)=3.098$, $p=0.036$, $\eta^2_p=0.171$.

| Mean (sd) | 2962 (375) | 1761 (454) | 3081 (371) | 2445 (254) | 2562 (364) |

Note: standard deviation reported in parentheses.
To investigate this interaction further a separate one-way ANOVA was performed for Match and Mismatch separately with Speed having four levels (back, quick, slow, and normal). For the SV Match ANOVA there was a significant main effect of Speed $F(3,48)=91.811$, $p<.001$, $\eta^2=0.852$ and Language $F(1,16)=6.969$, $p=0.018$, $\eta^2=0.303$.

There was a two-way interaction between Speed and Language $F(3,48)=2.880$, $p=0.045$, $\eta^2=0.153$.

Planned paired comparisons showed quicker responses to Matched Russian-Russian pairs of stimuli compared to English-English stimuli when they were played Back ($t(16)=2.938$, $p=0.010$) and Quick ($t(16)=2.333$, $p=0.033$) but no such difference was seen for the Normal and Slow speed (all $p>0.05$).

For the SV Mismatch ANOVA there was a significant main effect of Speed $F(3,48)=101.990$, $p<.001$, $\eta^2=0.864$ and Language $F(1,16)=22.238$, $p<.001$, $\eta^2=0.582$.

There was a two-way interaction between Speed and Language $F(3,48)=2.778$, $p=0.051$, $\eta^2=0.148$.

Planned paired comparisons showed quicker responses to Mismatched English-Russian pairs of stimuli compared to Russian-English stimuli when they were played Quick ($t(16)=3.539$, $p=0.003$), Slow ($t(16)=2.937$, $p=0.010$), or Normal ($t(16)=3.498$, $p=0.003$) but no such difference was seen for the Backward speed ($p>0.05$).

The results for the separate Video-Sound (VS) ANOVA showed that there was a main effect of Speed $F(3,45)=106.008$, $p<.001$, $\eta^2=0.876$.

There was two-way interaction for VS separate ANOVA for Speed and Language $F(3,45)=9.690$, $p<.001$, $\eta^2=0.392$. To investigate these interactions further a separate one-way ANOVA was conducted for Match and Mismatch, with Speed having four levels (back, quick, slow, and normal).

The results for the VS Match ANOVA showed a significant main effect of Speed $F(3,45)=75.6981$, $p<.001$, $\eta^2=0.835$. There was interaction between Speed and Language $F(3,45)=5.814$, $p=0.002$, $\eta^2=0.279$. Planned pair comparisons showed that for Match VS pairs participants were significantly slower for English-English pairs when the pairs were played in the Back speed compared to Quick speed $t(16)=10.962$, $p<.001$ and Normal speed $t(16)=8.956$, $p<.001$, but not slow ($p>0.05$).

In contrast, in the VS mode, Russian-Russian pairs played in the Slow speed resulted in the slowest response times compared to all other conditions (all $p<.001$).
The results for the VS Mismatch ANOVA showed a significant main effect of Speed $F(3,45)=103.562, \ p<.001, \ \eta^2= 0.873$ and Language $F(3,45)=6.684, \ p=0.021, \ \eta^2= 0.308$. Planned pair comparisons showed that for Mismatch VS pairs participants responded significantly quicker when stimuli was presented in the Quick speed compared to the Back speed $t(16)=16.565, \ p=0.007$, the Slow speed $t(16)=11.251, \ p<.001$, and the Normal speed $t(16)=6.039, \ p<.001$. For the VS Mismatched pairs, participants were significantly quicker for the English-Russian pairs than the Russian-English pairs $t(16)=2.083, \ p=0.054$.

5.2.3 Discussion

The results did not support the prediction that slowing speed of presentation would result in better performance for non-native Russian stimuli compared to the normal speed of presentation. Whilst slow speed of presentation seemed to help when an English image was followed by a matched English sound, no such effect was seen when the Russian image was followed by an English image. Generally, participants performed better with the normal speed of presentation. This suggests that the non-modified condition used in previous experiments is not artefactual and reflects typical (real time) speech processing.

Consistent with prior experiments, performance was easier and quicker when participants made a decision after hearing a target sound (Sound-Video condition). One interpretation of such findings is that hearing sounds primes corresponding visual images or prototypes prior to presentation. However, other competing explanations for these findings are that sounds contain necessary and sufficient information for matching performance and visual aspects are redundant. The fact that performance was better than chance in the condition when a visual native image was presented before hearing a non-native sound, suggests that a visual image does prime the corresponding sound, even though it may not be necessary. In any case, the most parsimonious explanation of the results puts sound at a premium for task performance and finds no significant support for any facilitation from visual only aspects of speech.

Mismatched pairs resulted in more errors than matched pairs, confirming the validity of the paradigm. Moreover, error rates were generally high for mismatched pairs, suggesting task demands were stringent. This suggests that when participants see an image that evokes a corresponding sound, the image is “fuzzy” or biased to the native language, resulting in poor performance for the mismatched pair.

One interesting result was the fact that with normal speed of presentation and matched pair of stimuli, there was no significant difference in performance between English and Russian
Visual perceptual assimilation of non-native speech

stimuli. This result is interesting in light of the expectation that non-native items would be more difficult because participants had no previous knowledge of Russian.

The three-way interaction between modality, language and speed of presentation shows that participants are more adept to processing normal stimuli in their native language compared to slow or quick native speech when they are matching sound with the corresponding visual image, and that they still extract some information (most likely rhythmic) from backward speech. However, the effects are dependent on whether the stimulus is matched or not.

In sum, the present results once again showed that little information is extracted from visual-only speech during non-native speech perception, even with cognate words played slowly. The next question is whether at the sentence level, native Australian English speakers can discriminate VO items played backwards (reversed) from non-native speech. Experiment 8 addresses this question.

5.3 Experiment 8: Sentence identification

The results of previous experiments showed that participants had difficulties identifying the language of cognates and sentences when presented in the VO modality as well as difficulty generating mental representations of speech from visual aspects only. If they cannot distinguish native and non-native cognates words and sentences in the VO modality, what conditions would be sufficient for veridical visual speech perception? Cross-linguistic AV studies suggest that it is possible to distinguish languages when words and sentences are reversed (Ronquest et al., 2010). In the study of Ronquest and colleagues (2010), English and Spanish single words and sentences were temporally reversed (played backwards) and participants were requested to state whether the items presented were English or Spanish. The two languages were identified as a part of the instructions but in one condition the participants were not informed that some stimuli were reversed and in another they were informed (observer’s condition). The reasoning behind the reversal of stimuli was to generate visemes that do not contain ‘legal’ phonotactic features but preserve coarse-grained segmental information, such as labial closure, as well as duration of vowel and consonants. This duration of visemes was expected to provide cues for temporal/rhythmic differences between Spanish and English but not for lexical retrieval. The results showed that monolingual US English-speakers performed above chance and could identify stimuli in all of the conditions, including isolated words and reversed sentences. As expected, they performed better with forward played stimuli than backward stimuli. Furthermore, they were better with sentences than
words when stimuli were played forward but not backwards. The results also showed that monolingual US English-speaking participants were more likely to identify words and sentences as Spanish when it was played backward compared to English suggesting a response bias to saying English for forward played stimuli. When monolingual participants were explicitly informed about stimuli played backward, there was an effect of length of utterance in the backward condition compared to when they were not informed. The effect of length was interpreted as evidence that participants relied on non-lexical cues (e.g. rhythmic differences) to identify languages. Moreover, if participants were informed about stimuli played backward, they were less likely to respond to the backward stimuli as Spanish (more conservative). This was interpreted as evidence that if participants knew that a stimulus looked unusual, they no longer relied on how natural it looked to make a judgement between English and Spanish language, thus biasing their response strategy toward false negatives.

The aim of the current study was to replicate the results of Ronquest, Levi, and Pisoni (2010) with Russian and English sentences played backward and forward/normal. Cognate words were found to be particularly difficult to identify visually in Experiment 7 and therefore were omitted in this experiment. Furthermore, to avoid any response bias inadvertently created by identifying specific languages or disambiguating temporally modified stimuli, a novel design was used. Specifically, participants were informed that they would see a speaker pronouncing either real sentences (in English or a foreign language) or nonsense sentences. In line with the previous experiments, the primary aim was to compare VO and AV modalities directly, thus expanding previous reports of VO identification of non-native speech using temporally modified input.

The dependent variables were again response error rates and times. The independent variables were Modality (VO and AV), Language (English and Russian), Sentence (Real and Reversed).

Experiment 8 Hypotheses:
Hypothesis 1a: Participants will be able to discriminate between real and reversed sentences in VO modality and will perform better with real vs reversed sentences.

5.3.1 Method

Participants: A new sample of participants to that used in the previous experiments was used in Experiment 7. Nineteen undergraduate psychology students from The University of Melbourne participated for course credit. There were 13 females. The mean age of participants was 19 years. All were native English speakers and none had knowledge of
Russian. All reported no known hearing impairment and normal or corrected-to-normal vision. Each participant completed both conditions 1a and 1b in the following order:

**Condition 1a: Real vs Reversed Sentence Video**

*Materials and Procedure:* Stimuli were Russian and English sentences used in Experiment 6. The English sentences are reported in Appendix A. The Russian sentences included idiomatic translation of the English sentences. The number of words was kept constant across languages and mean exposure time for English and Russian languages did not differ significantly. The nonsense sentences were created by auditorally transforming the original sentence files into backward played sentences using Adobe Premier resulting in 12 nonsense sentences derived from English and 12 backward played sentences derived from Russian, in addition to the original sentences. A total of 44 sentences was used, 40 experimental stimuli and 4 practice sentences. Within experimental stimuli, there were 20 English sentences (10 normal and 10 reversed) and 20 Russian sentences (10 normal and 10 reversed). The resulting stimuli were then separated into audio and video files using MPG program and only the video files were used in this condition. In the course of the experiment, each sentence was repeated twice. The participants were instructed that they would see a speaker saying either real sentences or nonsense sentences. They were instructed to press a “YES” button if the speaker was saying a real sentence, regardless of whether the sentence was English or in another language, or the “NO” button if they thought that the speaker was saying a nonsense sentence (the location of response buttons were counterbalanced across participants). They were instructed that that each item would be presented several times over the course of the experiment.

**Condition 1b: Real vs Reversed Sentence Audio-Visual**

*Materials and Procedure:* Stimuli were taken from Condition 1a prior to separation into video and audio tracks. The procedure was the same as in Condition 1a, with the exception of the stimuli presented in AV mode. The participants were provided with headphones, which were individually adjusted to fit each participant prior to practice trials. The same instructions for pressing the buttons were provided.

### 5.3.2 Results

A 3-way within-subject ANOVA was conducted for error rates. The first factor Modality had two levels: VO and AV, the second factor Language had two levels: English and Russian, and the third factor Sentence had two levels: Real and Reversed.
Table 21 below provides descriptive statistics for all factors.

**Table 21. Mean Error Rates for Video and Audio-Visual (AV) Conditions in Experiment 8.**

<table>
<thead>
<tr>
<th></th>
<th>Video</th>
<th></th>
<th>AV</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Real</td>
<td>Reversed</td>
<td>Real</td>
<td>Reversed</td>
</tr>
<tr>
<td>English</td>
<td>35.26</td>
<td>55.53</td>
<td>1.31</td>
<td>9.79</td>
</tr>
<tr>
<td></td>
<td>(20.20)</td>
<td>(19.36)</td>
<td>(2.26)</td>
<td>(14.00)</td>
</tr>
<tr>
<td>Russian</td>
<td>49.57</td>
<td>79.94</td>
<td>17.69</td>
<td>25.79</td>
</tr>
<tr>
<td></td>
<td>(11.89)</td>
<td>(13.60)</td>
<td>(14.15)</td>
<td>(22.56)</td>
</tr>
<tr>
<td>Mean</td>
<td>42.41</td>
<td>67.73</td>
<td>9.5</td>
<td>17.79</td>
</tr>
<tr>
<td></td>
<td>(16.04)</td>
<td>(16.88)</td>
<td>(8.20)</td>
<td>(18.28)</td>
</tr>
<tr>
<td></td>
<td>54.98</td>
<td></td>
<td>13.64</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(16.46)</td>
<td></td>
<td>(13.24)</td>
<td></td>
</tr>
</tbody>
</table>

Note: mean standard deviation reported in parentheses

The Mauchy test of sphericity indicated that the assumption of sphericity had not been violated ($p > 0.05$).

The analysis showed a significant main effect of Modality $F(1,18)=195.962$, $p<.001$. $\eta^2=0.916$.

There was a significant effect of Sentence Type $F(1,18)=104.477$, $p<0.01$. $\eta^2=0.853$.

The analysis found a significant main effect of Language $F(1,18)=90.991$, $p<0.01$, $\eta^2=0.835$.

There was a significant interaction between Modality and Sentence $F(1,18)=17.891$, $p=0.001$, $\eta^2=0.498$. When a separate one-way ANOVA was conducted for each modality, no significant interaction was evident in the AV modality ($p>0.05$) and it failed to reach significance in the VO modality ($p=0.09$). However, examining the error rates in the VO modality with a one-tail 50% t-test showed that participants were better than chance with real English sentences ($p<0.05$) but there were at chance for both Real and Reversed Russian sentences and Reversed English sentences (all $p>0.05$).

Planned pairwise comparisons between the modalities showed that participants made significantly more errors in the VO compared to AV conditions $t(18)=13.99$, $p<.001$. Furthermore, overall performance in the VO condition was no better than chance $t(18)=1.16$, $p=0.11$. The latter result was not consistent with the main Hypothesis 1a that participants would be able to differentiate real from reversed sentences in the VO mode.
Planned paired comparisons showed that participants made significantly more errors for Reversed Sentences than Real Sentences $t(18)=9.53$, $p<0.01$. This difference was seen across both the VO condition $t(18)=9.53$, $p<0.01$ and the AV condition $t(18)=3.06$, $p<0.01$.

Planned pairwise comparisons showed that participants made significantly more errors for Russian items compared to English items $t(18)=9.22$, $p<0.001$.

**Response Time analysis:**

A 3-way within-subject ANOVA was conducted for response times. The first factor Modality had two levels: VO and AV, the second factor Language had two levels: English and Russian, and the third factor Sentence had two levels: Real and Reversed.

Table 22 below provides descriptive statistics for all factors.

**Table 22. Mean Response Times for Video and Audio-Visual (AV) Conditions in Experiment 8.**

<table>
<thead>
<tr>
<th></th>
<th>Video</th>
<th>AV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Real</td>
<td>Reversed</td>
</tr>
<tr>
<td>English</td>
<td>4057 (242)</td>
<td>4415 (308)</td>
</tr>
<tr>
<td>Russian</td>
<td>4183 (252)</td>
<td>4447 (351)</td>
</tr>
<tr>
<td>Mean</td>
<td>4120 (247)</td>
<td>4431 (329)</td>
</tr>
</tbody>
</table>

Note: mean standard deviation reported in parentheses.

The Mauchy test of sphericity indicated that the assumption of sphericity had not been violated ($p> 0.05$).

The analysis showed a significant main effect of Modality $F(1,18)=27.925$, $p<.001$, $\eta^2=0.608$.

There was a significant effect of Sentence Type $F(1,18)=36.102$, $p<.001$, $\eta^2=0.667$.

The analysis found a significant main effect of Language $F(1,18)=58.187$, $p<.001$, $\eta^2=0.764$.

There was a 3-way interaction between Modality, Language, and Sentence $F(1,18)=6.634$, $p=0.019$, $\eta^2=0.269$.

A separate one-way ANOVA was conducted for each modality with within-subject factors of Sentence (Real and Reversed) and Language (English, Russian).

The results of the VO modality one-way ANOVA showed a significant main effect of Language $F(1,18)=18.141$, $p<.001$, $\eta^2=0.502$. Planned paired comparisons showed that the
English stimuli were responded to significantly quicker than the Russian stimuli \( t(18)=12.51, p<.001 \). The results of the AV one-way ANOVA showed the main effect of Sentence \( F(1,18)=39.245, p<.001, \eta^2= 0.686 \), with Real sentences resulting in quicker response latencies than Reversed sentences \( t(18)=3.060, p<.001 \). There was also a main effect of Language \( F(1,18)=30.760, p<.001, \eta^2= 0.631 \). Planned paired comparisons showed that the English stimuli were responded to significantly quicker than the Russian stimuli in the AV mode \( t(18)=9.530, p<.001 \). For the AV mode, there was a significant interaction between Sentence and Language \( F(1,18)=18.170, p<.001, \eta^2= 0.502 \). English Real sentences were quicker than Reversed \( t(18)=9.349, p<.001 \) but in the AV Russian Real and Reversed did not differ significantly (p>0.05).

Planned pairwise comparisons between the modalities showed that participants were significantly slower in the VO compared to AV conditions \( t(18)=1.837, p<.001 \).

Planned paired comparisons showed that participants were significantly slower for Reversed Sentences than Real Sentences \( t(18)=3.076, p<.001 \). This difference was seen across both the VO condition \( t(18)=4.432, p <0.01 \) and the AV condition \( t(18)=6.111, p=<0.01 \).

Planned pairwise comparisons showed that participants were significantly slower for the Russian items compared to English items \( t(18)=9.349, p<.001 \).

5.3.3 Discussion

The main question addressed in this experiment was whether participants discriminate real versus temporally reversed (backward) sentences. It was hypothesised that participants would differentiate real English sentences from backward sentences in the VO mode, based on the results of Ronquest, Levi, and Pisoni (Ronquest et al., 2010). In fact, participants here performed no better than chance for reversed sentences in the VO condition. There were important differences between the current study and Ronquest et al. (2010). The present study did not identify the type of foreign language (Russian), unlike Ronquest et al (2010) who did identify the foreign language as Spanish. It is possible that their design created an advantage by evoking general characteristic of Spanish (e.g. rhythm, intonation), in US participants. No such expectations were generated for the participants in the present study. Furthermore, the current study did not provide notice about the nature of stimuli manipulation (reversed presentation). It is possible that informing participants about the experimental manipulation results in a response strategy whereby participants do not rely on the naturalness of visual stimuli and use other cues, such as rhythmic differences instead.

In contrast to the results in the VO condition, participants were relatively accurate and quick for the reversed sentences in the AV condition, including Russian sentences, suggesting AV
presentation can disambiguate input and facilitate discrimination. There was no evidence of any additional benefit from visual aspects of speech however. Again, the current study highlights a need for the replication of previous results with different languages and experimental designs, including direct comparison of AV, VO, and AO condition. The latter was not included in the current experiment.

Finally, the results provide a novel insight into the process involved in making a judgement about sentence veridicality in an AV paradigm. It is still unclear, however, whether informing participants that the experimental stimuli will be reversed (as in Ronquest et al. 2010) would change response assuming this can evoke prototypes of how natural (forward) speech appears.
Chapter 6: General Discussion

This chapter presents a summary of all the results, followed by an evaluation of each experiment within the theoretical framework of auditory speech perception and the hypothesis of facilitation from visual cues. The chapter also presents implications of the study, notwithstanding limitations. Finally, future research directions and recommendations are considered.

Returning to the main question of the thesis we found very little evidence that the visual aspects of non-native speech are assimilated in the audio-visual (AV) perception of speech when native Australian-English monolingual speaking participants were presented with Russian language visual exemplars. We found that the visual aspects of non-native Russian speech that are dissimilar to the native English language were not identified any more readily than visual aspects of non-native speech that are similar to the native language during AV perception. In the eight experiments conducted, the results showed no consistent evidence of visual only perceptual assimilation in AV speech perception. Indeed, results showed that participants were unable to discriminate native and non-native languages based on visual-only input in contrast to the results reported by Hazan and colleagues (Hazan et al., 2006; Hazan et al., 2005) who used participants from rather different ethnic and language groups. Contrary to previous studies of Hazan and colleagues (2006), participants were unable to discriminate veridical from reversed sentences in a visual-only mode presentation. Furthermore, in cross-modal matching experiments, participants were unable to translate a visual image of the speaker pronouncing stimuli into the corresponding auditory signal, unless a written cue was provided or an auditory signal was presented. Slowing the presentation of stimuli during visual speech perception did not facilitate processing of visual aspects of non-native speech as expected.

Overall, an exhaustive test of the visible speech hypothesis showed a limited influence of visual aspects of speech on non-native perception when compared to auditory-only or audio-visual input. Results of the experiments comparing AV to other modalities show for the first time that naïve participants are more likely to falsely accept non-native stimuli as native in the multi-modal AV condition when compared to both auditory-only and visual only conditions. The results therefore constrain reports about the influence of visual aspects of speech on perception of sounds in a non-native language and highlight the role of language exposure and task demands on the putative effects of visual cues on non-native speech perception and, more
surprisingly, possible attention bottlenecks during AV perception of non-native speech. The findings will be discussed in terms of models of native and non-native speech perception.

6.1 Summary of Experiments

The first 5 experiments were concerned with vowels and syllables, whilst Experiments 6-8 explored perception of words and sentences. A language identification paradigm was used in Experiments 1, 2, 6, 7 and 8. In this paradigm, participants are presented with native and non-native speech stimuli (isolated English and Russian vowels and syllables in Experiment 1, 2; words and sentences (with or without written prime) in Experiment 6-8, and were asked whether a stimulus was English or not. Russian was not identified as the foreign language. This paradigm had been used in previous studies to explore how participants perceive native and non-native stimuli in different modalities - auditory, visual or AV (e.g. Hazan et al, 2005).

The results of Experiment 1 (vowel and syllable identification) showed that when participants were presented with dissimilar (non-existent in English) phoneme and syllable stimuli, they were able to discriminate between English and non-English stimuli. By contrast, for similar or identical Russian items, they tended to misclassify them as English. In Experiment 1 it was revealed that this pattern of responses was consistent across all three modalities - VO, AO, and AV. However, participants were more error prone in the VO modality compared to AO or AV. Unlike the results reported by Hazan and colleagues (2005), there was no beneficial additive effect of AV presentation when compared to the auditory-only condition for any items, including the most dissimilar ones. These results contrast with previous studies but highlight the possible influence of language background (i.e. naïve Australian English speakers with no exposure to Russian in their language environment), language type (i.e. English and Russian are both Indo-European languages), and presenter (i.e. the speaker had the same Caucasian ethnicity as the participants).

Experiment 1 also found high error rates for English items and ranking of error rates produced high variability, not predicted a priori. Therefore, Experiment 2 focused on replicating the findings with a validated subset of items, empirically defined as similar or dissimilar. Results from Experiment 2 replicated the findings from Experiment 1 in the AO condition by showing that unfamiliar Russian dissimilar items were discriminated best. In contrast, performance in the VO condition in Experiment 2 was no better than chance (even for dissimilar items), thus failing to replicate the findings from Experiment 1 in the visual-only condition.
In sum, initial results found no convincing evidence that naïve Australian English-speaking participants use the visual aspects of non-native speech in identification tasks. Furthermore, high variability in performance was noted and, most strikingly, there was no reliable AV benefit effect. Given discrepancies between the present findings and results from studies that do show an AV benefit, Experiment 3 was devised to interrogate the results using a different paradigm, a same-different task, whereby participants were presented with combinations of vowels and syllables from native and non-native languages and asked whether they were the same or not. Languages were not identified in the experiment to reduce possible response bias and avoid use of strategies to identify differences due to rhythmic patterns and intonation that were possible artefacts in similar studies (Soto-Faraco et al., 2007).

The results from Experiment 3 which utilised Same-Different paradigm, showed that participants made more errors with stimulus pairs that came from different languages when compared to those from the same language. For example, when they heard the English sound ‘ma’ followed by the speaker pronouncing ‘ma’ in Russian, they made more errors than when it was followed by the same English sound. However, fewer errors were observed for non-native items that were most distinctive from English whereas more errors were produced for non-native items that were similar to English, thus validating classification of items. Participants were more likely to detect the difference between pairs when the English items were contrasted with the Russian items that were classified as most dissimilar to English.

In line with the findings of Experiment 1 and Experiment 2, the results from Experiment 3 showed no additive benefit of AV presentation compared to the AO condition. An AV benefit was observed only when participants were presented with same language pairs compared to different language pairs. This suggests that the presence of audio-visual information aided the affirmative decision when same language pairs were presented but not when there was uncertainty, such as with different language pairs.

Experiment 4 investigated how participants perceived experimental stimuli using goodness of fit ratings on a visual 5-point scale ranging from Non-English item to Very Good English item. The results confirmed that participants mostly rated Russian items that were dissimilar to English as ‘non-English’ or ‘very bad examples of English’. Consistent with the category of Similar participants rated English items as very good exemplars of English and misclassified Russian items that were similar to English as very good exemplars of English.
Interestingly, English items that comprised the English-As-Foreign category saw participants rate these items relatively highly, suggesting that participants nevertheless accepted them as English exemplars, although not as well as Similar items. This finding weakened the assumption that English-As-Foreign items were perceived as non-native items. There were differences between items when rated in AO and AV conditions, with higher ratings (judging items as better English exemplars) in the AO condition than the AV condition for the Similar items. In contrast, they gave significantly lower ratings (less English) to the Dissimilar items in the AV condition than the AO condition, suggesting there was dissociation in the adoption of visual cues according to how similar items are to the native exemplars.

Experiments 5 and 7 used a cross-modal repetition priming and Match-Mismatch paradigm. In Experiment 5 participants were presented with combinations of matched or mismatched video and sound stimuli (vowels and syllables) in the native and non-native languages. As in earlier experiments, languages themselves were not identified to participants. In Experiment 7 participants were presented with cognate words played at different speeds. Experiment 5 and 7 used two conditions (Video-Sound and Sound-Video) and participants were asked whether stimuli matched. This paradigm is commonly used in cross-modal studies in order to tap into the mental representation of speech (Sánchez-García et al., 2013).

The results from Experiment 5 showed that when participants were shown a bilingual speaker pronouncing a sound in native and non-native speech, followed by hearing the same sound (Video-Sound) responses were at chance. Against the predictions, the responses for dissimilar non-native sounds were at chance, even for the matched pairs. Unexpectedly, similar sounds in native and non-native speech matched pairs resulted in performance that was significantly better than chance in the VO condition, although error rates were relatively high. This pattern of responses was a direct contradiction of the working assumption motivating the study. The prediction that the Sound-Video condition would be easier than the Video-Sound condition was supported. However, overall error rates for the Sound-Video condition were surprisingly high, although better than chance. These results suggest that if participants hear a sound they form the corresponding visual representation of the sound, although the process is not always accurate, especially for non-native speech. The reverse does not seem to be true however. If participants view images they do not form the corresponding sounds for native and non-native speech. The results from Experiment 5 highlight the effect of task difficulty in the paradigm, evident in high error rates. One reason for task difficulty may be insufficient entropy from the vowels and syllables. Therefore, Experiment 6-8 used words and sentences to afford more information at various levels of speech processing (e.g. segmental, suprasegmental, lexical).
Experiment 6 included several conditions: In **Condition 1a: Video Word** participants viewed a face of the speaker pronouncing native and non-native words in the VO mode and were asked to say whether the word was English or not. In **Condition 1b: Video Written Word** a printed word was presented prior to seeing the face. In **Condition 1c: Audio-Visual Word** participants heard a word and saw the face of a speaker simultaneously. In **Condition 2a: Video Sentence** participants viewed a face of the speaker pronouncing native and non-native words in the VO mode and were asked to say whether the word was English or not. In **Condition 2b: Video Sentence Written** a printed sentence was presented prior to seeing the face. In **Condition 2c: Audio-Visual Sentence** participants viewed a face of the speaker pronouncing native and non-native sentences in the AV mode and were asked to say whether the sentence was English or not. However, this condition was eliminated from further analysis because of a ceiling effect.

Results from Experiment 6 supported the prediction that performance in printed word primed conditions would be better than in conditions without a printed word cue. Moreover, performance in the VO conditions without a printed cue was at chance, suggesting participants could not discriminate whether visually presented words and sentences were native or not from visual aspects alone contrary to the main hypothesis. Although performance in printed cue conditions was better than performance without a cue, participants were at chance when presented with a printed native English cognate word followed by a non-native visual image, incorrectly identifying the image as native thus showing an increased tendency toward false positive responding. In contrast, participants performed above chance levels when discriminating between native and non-native sentences when presented they were presented with a written English-sentence cue. Also consistent with the hypotheses, performance with sentences was superior to words suggesting that the visual aspects of speech have an impact when stimulus entropy is greater. Results from Experiment 6 also replicated the finding that participants made more errors with non-native words and sentences compared to the native stimuli. Moreover, participants made more errors with the non-native items when no visual cues were available, suggesting benefit from providing a native cue/prime even when it was followed by a non-native stimulus/target.

In Experiment 7 participants were presented with the cognate words used in Experiment 6 but varying in duration: twice as fast, slower, reversed and non-modified. There were two types of presentation: sound-video and video-sound. Against prediction, significantly more errors were observed in the slow compared to the non-modified condition. Whilst slow speed of presentation seemed to help when an English image was followed by a mismatched Russian sound, no such effect was seen when the Russian image was followed by an English image. Consistent with other findings in this study, performance was better when participants made a
Visual perceptual assimilation of non-native speech

decision after hearing a target sound (sound-video) than in the video-sound condition and when pairs were matched compared to mismatched.

The goal of Experiment 8 was to determine what information is sufficient to make accurate judgement about non-native speech in the visual-only modality. In Experiment 8 participants were presented with the native and non-native sentences used in Experiment 6 plus nonsense sentences (reversed). The latter manipulation allowed for retention of rhythmic but not lexical properties of speech. There were two conditions: visual-only and audio-visual. The results showed that participants were at chance in the VO condition. In contrast, they were able to differentiate nonsense sentences from real sentences in the AV condition, even for non-native sentences, thus demonstrating an AV benefit compared to VO presentation. The results confirm the view that emerges across studies: visual aspects of speech generate no additional benefit unless they are accompanied by audio information. In other words, there is no indication that visual aspects of speech play a unique role in native and non-native speech.

The following is a comparison of the experimental findings, testing against the hypotheses and extant knowledge about native and non-native speech processing in different modalities. Specifically, auditory processing of native and non-native speech will be compared with AV speech processing and VO speech processing. The main findings will then be interrogated in light of the theoretical models of speech processing reviewed in Chapter 1.

6.2 Visual only speech

The primary motivation for the study was to test the hypothesis that non-native visual aspects of speech are assimilated in native speech perception; specifically that some visual aspects of non-native speech will be perceived as native when presented in the VO domain whilst other visual aspects of speech will be perceived as non-native speech. There are no studies examining this hypothesis at the level of vowels, syllables, words, and sentences. The motivation for this hypothesis is that some cross-language studies report a lack of facilitation from VO cues when training phonological contrasts in a non-native language (Hazan et al., 2005; Ortega, 2009).

The results show no convincing support for the hypothesis at the level of syllable/vowels. In Experiment 1 there was evidence of perceptual assimilation from the visual modality for the most dissimilar non-native items. However, Experiment 2 failed to replicate results of Experiment 1 and there was no evidence that dissimilar non-native items were discriminated better than similar non-native contrasts in the visual modality. It is important to note however
that errors were diagnostic in Experiment 1, i.e., high error rates were observed in the VO condition, whereas performance in Experiment 2 was at chance in the VO condition regardless of the type of item. Results in Experiment 2 are broadly consistent with the view that lip-reading in general, and specifically for vowel/syllables is not a natural task for most participants (P. K. Kuhl et al., 1988); (Bernstein & Auer, 1996). Results in the VO mode were informative in relation to the question of whether languages can be discriminated on the basis of visual cues alone. This question has previously been addressed at the level of words and sentences only (Soto-Faraco et al., 2007). One of the main contributions of the present results is using a novel language pair, Russian-English, to address this question at the level of vowels and syllables. Studies reporting differentiation of native and non-native speech based on visual information only used languages that are familiar to English speaking participants (Ronquest et al., 2010; Soto-Faraco et al., 2007; Weikum et al., 2013) or faces that are from different ethnicities. Russian spoken by a native Caucasian talker removes these artefacts. The current results show that VO differentiation is difficult for native Australian English speakers when the language is not at all familiar and the talker is the same ethnicity as the participant.

Given the results of the aforementioned studies, another hypothesis was that more information would be afforded by words and sentences when compared to vowels and syllables. Against this hypothesis, the current results showed that without a written cue, participants are unable to discriminate native and non-native speech based on the VO information. It is also unclear how to reconcile the present results with some previous reports. Therefore, we need to examine differences between the current study and those showing that language discrimination is possible based on visual-only information.

The first issue concerns language familiarity. The study of Soto-Faraco (Soto-Faraco et al., 2007) used Catalan and Spanish, which are linguistically closer languages than English and Russian and critically, often co-occur in the language environment of Spanish speakers. On both counts it is probable that the non-native language is more familiar for Spanish speakers than for Australian English speakers. Another difference concerns possible strategies adopted by participants in the native language judgments. Ronquest and colleagues (Ronquest et al., 2010) tested English and Spanish but also identified the languages and directly required participants to indicate if the stimuli were English or Spanish. It is possible that identification assisted US English-speakers as they might be exposed to Spanish, given that it is the second most common language spoken in the US. By contrast, familiarity with the Russian language in Australia is low and it is not even included in the top dozen languages spoken in Australia.
The second issue is cross-linguistic rhythmic differences. Cross-linguistic studies show it is possible to distinguish languages if words and sentences are temporally reversed. In the study of Ronquest and colleagues (Ronquest et al., 2010) participants were required to state whether a words or sentences presented normally or reversed was in English or Spanish. Reversed stimuli were assumed to preserve coarse-grained segmental information, such as labial closure, as well as the duration of vowel and consonants. The latter was considered to provide cues to rhythmic differences between English and Spanish. The results showed that monolingual US English-speakers were able to identify reversed words and sentences above chance. The results also showed that US participants were more likely to identify words and sentences as Spanish when they were reversed and showed a response bias of saying English for non-modified stimuli.

The results failed to replicate the findings of Ronquest and colleagues (Ronquest et al., 2010) which showed with their participants that they could differentiate language based on the visual only input. Methodological differences between the present study and Ronquest and colleagues study (2010) may explain this outcome. The current study did not explicitly identify a ‘foreign language’ as Russian. As previously argued, disambiguating the identity of the non-native language, may have evoked characteristics of the Spanish language, such as its rhythmic pattern, allowing participants to tune into unique aspects of the stimuli. It is important to note however that there are commonalities between the two studies. Similar to Ronquest and colleagues (2010), the results show that participants reveal a bias towards misclassifying reversed stimuli as non-native as well as more errors for non-native stimuli. In relation to reversed speech, it is also worth noting that methodological factors are likely to be important. In Experiment 7, which used a cross-modal matching design, participants were requested to match backward played cognate words cross-modally, with language being either continuous or discontinuous within the trial. The results showed that performance was poor for reversed non-native sounds presented auditorily then followed by a visual image - but above chance. The fact that participants were not able to perform the same task with reversed native words suggests that discriminatory aspects of reversed non-native speech were more obvious making identification easier than for native English sounds.

Another hypothesis was that slowing the speed of stimuli presentation would identify visual aspects of non-native speech. This hypothesis was tested in Experiment 7 by presenting words at a slower than normal rate. The rationale for the hypothesis was evidence from the auditory domain that slow speech is easier to process due to increased pauses (Bradlow & Bent, 2002; Picheny et al., 1986; (Smiljanic & Bradlow, 2008)) syllable duration and changes in segmental cues (J. Miller & Volaitis, 1989), (Joanne L. Miller & Dexter, 1988; J. L. Miller et
Moreover, slow presentation facilitates non-native listening comprehension (Chaudron, 1982) (Derwing, 1990). Several theories also assume that matching stimuli to prototypes is more difficult when the input is fast as compared to slow (Goldinger et al., 1991) (Nygaard et al., 1994) Contrary to expectation however, there was no evidence to that slowing speed of presentation facilitates identification of non-native visual aspects of speech. The finding that slower speech does not necessarily improve non-native speech identification adds to the existing body of literature. Derwing (1990) failed to find support the hypothesis that slower rate results in better comprehension for non-native speakers. The current study used 70% reduction to retain natural aspects of speech, which may explain why no benefit was found and it is possible that less reduction may show benefit. It is worthwhile noting that when participants heard a slow English word, followed by a video image, they made more errors than when English words were presented at a non-modified rate but no such trend was observed for Russian words. These results are relevant to the theoretical models of speech adaptation, discussed shortly.

In contrast to the results with slowed stimuli, participants adopted better to faster speech than slow speech. This result is consistent with literature showing that native listeners adapt to variation in speech rate, ranging from 140-180 words per minute (Lane, Grosjean, Le Berre, & Lewin, 1973) (Stine, Wingfield, & Myers, 1990; Wingfield, 1996) and are able to adapt to a rate which is twice as high as a typical rate (Foulke & Louisville Univ, 1971).(Stine et al., 1990). Prior research suggests that adaptation is different for non-native speech. For example rapid adaptation to compressed speech is less effective in a non-native language(Golomb et al., 2007) and increases in speech rate hinder perception of non-native speech compared to native speech in bilingual speakers (Rosenhouse et al., 2006). In the current study, performance for Russian time-compressed stimuli was no better than chance whereas performance to native time-compressed stimuli was significantly better than chance. These findings have potential implications for theoretical models of adaptation to time modified speech. Theoretical models explaining processing of time-modified speech, will be explained below keeping in mind that these theories are psychoacoustic in their origin.

Abstractionist models (D. Norris, 1994) (Dennis Norris & McQueen, 2008) propose that listeners store one representation – a canonical form of each word. Any variation in speaking rate has to be abstracted pre-lexically to access word meaning. These models assume that any signal variability is resolved before lexical meaning can be accessed. Signal degradation, such as due to segmental deletion in fast speech for example would increase cognitive load due to more challenging mapping between signal and representation. The theories assume that naïve
listeners, who do not have representations of non-native speech, will be less able to map non-native signals into existing representations.

Exemplar models (e.g., Goldinger et al., 1991) assume that storage of a sound pattern including exemplars produced by different speakers, with different rates, etc., form fuzzy categories for perception. Acoustic input is compared to sound representations. Variability due to speaking rate, speaker idiosyncrasies is not abstracted. Rather, spoken input is directly mapped onto the exemplars of fast or slow speech, for example, and compared to the best matching exemplars. The underlying assumption is that linguistic categories include a highly detailed collection of instances/exemplars and listeners can draw on knowledge of how fast or slow speech occurs. Indeed, in the current study participants performed better than chance with time-compressed native stimuli. This class of model generates predictions about the processing of non-native speech because it allows higher-order assumptions to play a role. For instance, in the study by Bosker and Reinisch (2015) listeners used pre-existing knowledge that non-native speech is typically slower than native speech, therefore presenting native and non-native speech at the same duration makes non-native input sound faster. However, the question remains if higher order assumptions about visual aspects of non-native speech. The results of the study failed to support for this idea, as participants were at chance for Russian time-compressed speech.

Another class of models - probabilistic models of speech perception, attempt to explain how listeners adapt to variability of a speech signal. These models, such as the Belief Updating Model of perceptual adaptation (Kleinschmidt & Jaeger, 2015) assume that listeners keep up to date with the consistencies of the acoustic signal, such as idiosyncratic pronunciations of certain sounds, speaking rate. Listeners then form adapted models of cue distributions that can be reapplied when encountering the same speaker, thus facilitating perceptual adaptation. An extension of the model to visual aspects of speech, would postulate a parallel process (or an integrated bimodal process) for the visual aspects of speech. The model is appealing because of its flexibility but there is little evidence from the current study that listeners were able to adapt their models of visual cue distribution to non-native speech.

The Reverse Hierarchy Theory (RHT) of perceptual learning (Hochstein & Ahissar, 2002); (Ahissar & Hochstein, 2004; Ahissar et al., 2009) makes the assumption that perception and learning occurs in a top-down manner by default. Since lexical and semantic context is rarely accessible in non-native speech (with the exception of cognates), the theory would predict that low-level (sub-lexical) acoustic properties would be accessed first then mapped backwards in a top-down (lexical) manner. Similarly, the theory predicts that if listeners are presented with
time-compressed native speech, lexical cues would not be available and sub-lexical cues will be accessed in a top-down manner. This theory also predicts a cost for backwards search and can account for discrepancy between performance for native and non-native time-compressed speech. Therefore, the finding of such differences here can be accommodated in part by RHT.

Theoretical models make different assumptions about the perception and production of time-modified speech occurs. Dupoux and Green (Dupoux & Green, 1997) for example argue that speech processing occurs at a relatively abstract level of perception. Therefore errors may be constrained by top-down cognitive processes in speech comprehension and biased by native speech categories. Others, such as Cutler and Mehler (1993) argue that adaptation to time-compressed speech can occur for non-native speech but will depend on whether a language pair shares rhythmic properties (among others). In support of the latter argument, Sebastian–Galles and colleagues (Sebastián-Gallés et al., 2000; Sebastian-Galles & Soto-Faraco, 1999) manipulated different language pairs, such as Catalan-Spanish (similar in sound, lexicon, morphological system, and syntax), French-English (dissimilar in sound, lexicon, morphological system, and syntax), and Dutch-English (mixed). They showed that adaptation to time-compressed speech was language dependent and argued that speech representation is pre-lexical (intermediate between lexical and acoustic/phonetic levels) and speech adaptation depends on the contrast between language pairs at a pre-lexical level, including regularities in syllable stress. The findings here support the view that processing of time-compressed speech is language dependent. Moreover, the results suggest that naïve English participants do not extract visual features of an unknown language at the pre-lexical level.

Other findings from this study contribute to understanding of non-native speech perception. Specifically, the hypothesis derived from previous studies of second-language lip-reading that sentences would be processed more readily in the visual domain compared to less informative units (words or single vowel/syllables) was supported. However, the results also contrast with findings from other studies. For example, Soto-Faraco and colleagues (Soto-Faraco et al., 2007) and Ronquest and colleagues (Ronquest et al., 2010) reported that participants could discriminate languages based on visual signals alone and discrimination was better for long sentences than for shorter ones. Ronquest and colleagues (2010) also showed that discrimination of languages based on visual input was better for sentences than for words. The current study failed to identify the same effects and participants could not discriminate English and Russian sentences based on visual-only input. The present results suggest that whereas Soto-Faraco and colleagues (2007) and Ronquest and colleagues (2010) found that rhythmic information facilitates discrimination, their results may depend on language familiarity or similarity. We assume that Australian English speakers do not know the sub-
lexical features (e.g. rhythmic, stress and vowel information) of Russian and so they rely on lexical information to make decisions about whether speech is English or not. Several studies, including Soto-Faraco and colleagues (2007) and Ronquest and colleagues (2010) showed that this information is scarce, with the former study reporting that only 3% of lexical information (words) could be recalled from sentences and the latter showed that only a quarter of lexical content was extracted from native VO speech stimuli. The difficulties with visual lip-reading of words and sentences displayed by the participants here are also in line with results from studies highlighting high levels of task difficulty in lip reading for hearing participants (Bernstein, Demorest, & Tucker, 2000; Kim & Davis, 2004).

Given that the present study failed to support the primary hypothesis, what might facilitate the process of native and non-native speech discrimination using visual-only aspects of speech? The current study hypothesised that the provision of a written cue/prime would result in better performance. Indeed, this hypothesis received some support from the results of Experiment 6 showing that performance for words and sentences was above chance if written cues/primes (cognates) were provided prior to VO presentation. Importantly, provision of a native cognate cue followed by non-native visual input resulted in a bias whereby participants incorrectly labelled the VO stimulus as native. The fact that discrimination was also above chance when native and non-native sentences were presented with a written English sentence also suggests that such stimuli provided sufficient information for discrimination. What, therefore, are the processing differences underlying cued and non-cued conditions?

One model designed to account for an effect of printed prime cues is the bi-modal interactive-activation (BIAM) model (Diependaele, Ziegler, & Grainger, 2010; Grainger & Ferrand, 1994; Graigner & Holcomb, 2009; Kiyonaga, Grainger, Midgley, & Holcomb, 2007, cited in Okano, Grainger, Holcomb, 2016). In this model a printed prime activates representations of sound at different level of speech processing including the level of phonemes and syllables (sub-lexical) and whole-words (lexical and semantic levels). This activation is bidirectional insomuch as hearing a word auditorily may also activate an orthographic word representation. Evidence for the BIAM model comes from a variety of cross-modal priming experiments (e.g. Grainger, Diependaele, Spinelli, Ferrand, & Farioli, 2003; Kiyonaga et al., 2007, cited in Okano et al, 2016) showing that related primes (repetitions) result in better identification. The facilitatory effect varies according to type of stimuli and prime duration. Whilst the BIAM model does not address VO speech, we can conjecture that orthographic representations activate auditory representation and perhaps also articulatory and visual representations. Indeed, Lansing and Helgeson (1995) demonstrated that when a written word (semantically related or not) was followed by VO speech presentation, speech reading was improved for
semantically-related primes. Lasting and Helgeson (1995) suggested that presenting a prime, activates a conceptual model of a word and lip reading is facilitated by matching the observed articulation to the model. They argued that when there is no match between prime and VO stimulus (semantically unrelated primes), lip reading is more difficult and error prone. This explanation resonates with to Motor Theory (Liberman and Mattingly, 1989) and Visual Place Articulatory Manner theory (Summerfield, 1987). The present study extends this account to non-native speech processing, by showing that if written word and sentence cues match visual input, performance is better. However, perceptual bias towards the native language remains, as could be seen by a tendency to misclassify (false positives) non-native stimuli as native.

The impact of print on processing of non-native language is not well researched. However, some insights come from eye tracking experiments. An eye-tracking paradigm involves presenting auditory and lexical information at the same time and measuring eye tracking as processing unfolds. For instance, Cutler and colleagues (2006; 2004) used this paradigm to study how certain difficult contrasts are perceived by L2 listeners. Weber and Cutler (2004) used Dutch learners of English who confuse the vowels sounds in *had* vs. *head* whilst Cutler and colleagues (2006) used Japanese learners of English who confuse the initial consonants of *write* vs. *light*. In these experiments, the Dutch participants were presented with pictures of a panda and a pencil. Correspondingly, the Japanese listeners were presented with pictures of a rocker and a locker. They had to click on the right object as they were listening to the auditory input. Their eye tracking was monitored. The results showed that the Dutch participants showed a bias towards the /ε/ sound, and their eyes were focused on the pencil regardless whether they heard *pen-* or *pan-* . A similar pattern was shown with the Japanese participants who showed a bias towards the /l/ sound, focusing their eyes on the locker whether they heard *lock-* or *rock-* . Therefore, they showed a bias towards their L1 category and assimilated L2 sounds into their L1 category. However, no written/orthographical input was presented.

Cutler (2015) argued that orthographical representations of difficult non-native contrasts do not assist with discriminating them auditorily or producing them correctly. She argued that unlike children, adult L2 learners receive explicit instructions regarding difficult contrasts and are well aware of their difficulties with certain contrasts. However, this does not in itself assist them with correct auditory discrimination. Moreover, she argued that it may even have a detrimental effect when orthographical representations increase competition.

Broersma and Cutler (2011) demonstrated this detrimental effect using a number of experiments within the cross-modal priming paradigm. They asked participants to perform a lexical decision task on a string of letters whilst they also had the auditory input. When the written input and the auditory prime were matched, participants’ response time was quicker.
than when it didn’t match. Most relevant to the current thesis, is the fact that this priming effect was present for L2 listeners. Moreover, it was present for Dutch listeners who showed confusion with the vowels used in the stimuli; specifically they responded more quickly to a written input DEAF following hearing an auditory prime daff. In contrast, daff did not serve as an effective auditory prime for native English participants. Furthermore, for both English and Dutch participants there was no priming effect when the auditory prime and a written input were not concordant, such as when participants heard definite as an auditory prime and read DEAF on the screen. Broersma and Cutler (2011) reasoned that even though deaf was embedded in definite and its lexical representation was activated briefly, the subsequent context –inite mismatched deaf. This absence of a priming effect was present for both L1 and L2. In contrast to the deaf/definite pair, when Dutch participants heard daffodil as an auditory prime, they showed a priming effect to the written word DEAF. As a result, Cutler (2015) argued that when an auditory distinctions between L2 contrasts cannot be perceived and they are incorporated in the lexicon, they hinder rather than facilitate L2 participants’ performance because of increased competition of embedded forms such as deaf in daffodil. Cutler (2015) concluded that unless lexical encoding of a contrast is accompanied by the corresponding auditory discrimination of the contrast, the value of orthographic input is limited. She argued that learning L2 produces interference from L1 at all levels, including orthography and the latter is least helpful with contrasts that are not discriminated auditorily.

Another study demonstrated that the presence of misleading orthography can hinder the acquisition of difficult L2 contrasts (Bassetti, 2006). This study used naïve English speakers who were in the early stages of learning Chinese and looked at the impact of pinyin – the writing system representing Mandarin Chinese. Using phoneme counting and segmentation tasks, Bassetti demonstrated that decipherable written input can hinder performance by resulting in the formation of non-target-like representation of L2 contrasts. The study of Escudero and colleagues (2010) looked at the impact of orthography on auditory perception of difficult contrasts in Dutch and Spanish participants. Their results also showed that orthographic representation impacted negatively on the auditory processing of L2 difficult contrasts when it was different to L1 orthographic representations.

In the recent studies of Escudero (2015) the role of orthography on a novel auditory L2 (Dutch) word-learning task was examined. The study compared easy and difficult to discriminate phonemic pairs. They included participants with transparent and opaque L1 orthographies (Australian English and Spanish participants). The results showed that independently from L1 and native orthography, the availability of orthography during word-learning tasks had a positive effect only on two contrasts, namely the ones with the highest or
intermediate accuracy levels within the perceptually difficult pairs. The researchers argued that orthography provides an extra cue to enhance differences that have already been mastered in the auditory domain. Participants’ performance on the difficult contrasts was at chance and not facilitated by the provision of the orthographic input.

In contrast, there are studies that show that orthographic input can be useful in L2 acquisition. Escudero and colleagues’ (2008) experiment used a difficult phonemic contrast for Dutch participants - i.e. /ɛ/–/æ/ and utilised an eye-tracking paradigm. The results showed that the participants who were only hearing input without orthographic input could not distinguish the first syllable of newly learned words with the difficult contrast. In contrast, when they also had orthographic input in addition to the auditory input, Dutch listeners looked at both images when hearing the first syllable of the new words in English. The researchers concluded that grapheme–phoneme mappings can facilitate phonological representations of new L2 contrasts.

Another study showed a positive effect of the presence of orthographic input. Showalter and colleagues (2013) investigated the acquisition of contrastive tonal contours in Mandarin using tone marks. Their results demonstrated that participants who had a tone-marked script outperformed the control group in acquiring the new phonological shape of items. They concluded that the provision of orthographic input, including in the form of unfamiliar diacritics, can facilitated foreign phonology acquisition.

One of the few studies examining orthographic input which also included AV input is the study of Erdener and Burnham (2005). They tested monolingual native speakers of Australian English and Turkish participants, using Spanish and Irish speech stimuli. English and Irish have opaque orthographies whilst Turkish and Spanish have transparent orthographies. There were 4 conditions: AO, AV, auditory–orthographic (participants heard and read the stimuli) and auditory–visual–orthographic (participants heard the stimuli, saw the speaker’s face and saw the stimuli presented in written form in a caption). Participants’ task was to repeat the input and their errors were rated by native speakers of Spanish and Irish. The results showed that with the AV condition, production performance was better, thus showing an AV benefit effect. When the written input was present, regardless of whether AV was also present, Turkish participants relied on the written input more than the Australian participants. This strategy proved fruitful especially with the Spanish stimuli with transparent orthography, but not with Irish stimuli. In contrast, Australian participants ignored the orthographic input and relied mostly on the auditory and AV sources, showing no difference between Spanish and Irish stimuli across different orthographic conditions.
The current study adds to the body of research which shows positive effect of orthographic input. Specifically, it facilitates visual perception of non-native words and sentences, even when primes are presented in the native language. These findings would need to be replicated with other orthographic scripts and different language between firm conclusions could be reached. However, the preliminary findings from the current study suggest that there is some value for presenting native orthographic input in the very early stages of L2 acquisition when processing sentences and cognates. Importantly, the latter may not generalise to other words.

Indeed, one of the questions of interest from the current results is what printed prime to use when assessing L2 speech processing in the very early stages of encountering a foreign language? Cognates may lead to better access to mental representation of non-native speech, as seen in many psychoacoustic studies. We now briefly review the relevant results. We note here that the paradigm used in the present study is a novel one and so comparisons in the field of L2 AV and speech-reading areas are not available.

Cognates can be defined as words that share form and meaning across languages. Importantly for the topic of the current study, cognate words can be pronounced differently, although the degree of difference is variable between language pairs. Cognate words are acquired in early in second-language acquisition (Desmet & Duyck, 2007). A cognate effect has been reported in studies of word recognition (Lemhofer & Dijkstra, 2004), translation (Boada et al., 2013) (Groot, 1992) and priming (Lalor & Kirsner, 2001). However, no study has examined visual speech aspects of cognate processing in naïve participants. The hypothesis tested was that cognate words have privileged access in the initial stages of language identification, including the processing of visual aspects of speech. The results show some support for this hypothesis. We will now examine the theoretical models underling this privileged access, keeping in mind that these theories are not addressing the visual aspects of speech. It must also be stressed that cognates do not always lead to facilitation in word recognition and production. Indeed, many studies show an inhibitory effect of cognates on speech production (Costa, Santesteban, & Caño, 2005) and false cognates (words with the same orthography and pronunciation but different meaning e.g. ‘room’ in Dutch and English) also produce effects of interference on visual word recognition (Brenders, van Hell, & Dijkstra, 2011).

One account of cognate facilitation effects comes from the Distributed Feature Model (Groot, 1992) which assumes that cognates can be processed faster and more accurately because they overlap in lexical and semantic levels of processing. This account is supported by the classic reports of semantic priming effects for cognates in fluent Dutch-English bilinguals (Groot &
Nas, 1991) and highly proficient Basque–Spanish bilinguals (Perea et al., 2006). Voga and Grainger (2007) later proposed that multiple activation of cognates results in stronger activation of semantic representations. When cognates share meaning (typically high imageability nouns) facilitation may be observed in perception (though not necessarily production). Increased activation of cognate semantic representations then results in top-down feedback into mapping of word form representations. Notably, top-down feedback could be facilitatory (as in priming) or inhibitory (as in production and with false cognates) and the direction of these effects varies according to the target language i.e. greater for L1 than L2 or vice versa. Midgley, Holcomb, Grainger (2011) proposed a hybrid model, incorporating the parallel activation model (Voga & Grainger, 2007) an Revised Hierarchical model (RHM; Kroll & Stewart, 1994). RHM assumes that connections are stronger from L2 to L1 than vice versa. This added connectivity contributes to cognate advantage for L1 words but has little influence on the cognate advantage for L2 words in priming, which continues to be driven by semantic feedback. By contrast, Sánchez-Casas and García-Albea (Sánchez-García et al., 2013) assume that cognates share a common abstract morphological representation. In their model, words from the same morphological family (e.g., rich, richer, richness) are connected by an abstract morphemic representation. In cognate words e.g., rich, richer, the supra-lexical morphological representations are therefore modality independent (e.g. orthographic versus auditory) and also language independent. The present results are compatible with the assumption of amodal representations given that written cognates facilitate VO speech processing.

According to all theories we may expect that when participants read a written word (cognate) and form a mental representation of the word (including visual or articulatory aspects), they should be better equipped to detect a discrepancy between native and non-native VO speech. However, the current study did not support this prediction completely, as participants tended to misclassify visual images of non-native cognates as native (false positive bias). It is likely that the visual correlates of articulation for unfamiliar non-native cognates were too subtle to detect. Alternatively, the more liberal response bias may be related to accepting a high degree of variability in the visual input. Based on the concept of immediate visual memory span(Sperling, 1960) visual image of cognates are expected to decay faster than that of a sentence. Indeed, sentence primes resulted in better performance for native and non-native sentences compared to cognate words.

Evidence of ‘translation’ processes in AV speech perception is seen from the results in cross-modal repetition priming and matching experiments. Experiments 5 and 7 used this paradigm. In this paradigm cross-modality matching was manipulated according to continuity of context.
Experiment 5 used vowels and syllables whilst Experiment 7 used cognates played at different speeds. Results from Experiment 5 showed that when participants are shown a speaking face pronouncing a vowel or a syllable, subsequent audio presentation of a sound (Video-Sound condition) produces matching responses that are no better than chance. Against predictions, responses for non-native sounds were at chance, even in matched (language continuous) pairs. Unexpectedly, both native and non-native sounds that were rated experimentally as ‘similar’ generated fewer error rates for matched pairs in the VO condition.

The present results contrast with the findings from Buchwald and colleagues (2009) who presented visual only primes followed by auditory targets and found that words presented auditorily in noise were better identified if they were preceded by a video clip of the matching prime. This effect was maintained when the same stimuli was presented by different speakers. Dodd and colleagues (1988) report that VO primes facilitate performance on a semantic categorisation task with auditory targets. These researchers interpreted their findings as evidence that visual primes and auditory targets activate the same amodal representation. We note, however that stimuli here were presented in clear format (no noise) and single vowels/syllables were used in Experiment 5. It is possible that insufficient entropy was afforded for task performance. It is also possible that one visual image generated many auditory stimuli as potential candidates. Subsequently, when a single sound was heard, possibly a non-native sound, it was necessary to translate that sound to a visual non-native articulation. This would require greater cognitive resources and the process would be more effortful and this may explain the liberal response bias. RHT specifically assumes such a cost (Ahissar & Hochstein, 2004; Ahissar et al., 2009; Hochstein & Ahissar, 2002). The primary assumption is that input is mapped to high-level representations by default unless this process is disrupted (e.g., when non-native speech does not map to any higher level representation). As a result, re-mapping of lower-level input to higher levels of representation should occur before a percept is complete. Such remapping is referred to as “perception with scrutiny” or backward search.

The present results also highlight the fact that different mechanisms may operate during the task of cross-modal matching depending on the linguistic setting and experimental design. For example, Kim and Davis (2014) argue that a visual cue allows the auditory processing system to re-set for processing of auditory stimuli. They argued that re-setting increases sensitivity to incoming auditory stimuli. The apparent lack of any benefit from visual cues when followed by an auditory signal here suggests that this mechanism may not operate as efficiently when processing non-native speech. One caveat to this claim is that unlike previous studies, both the ethnicity and the language of the experimental talker were closely matched in the current
study. It may be that benefits from AV speech and VO speech in particular are minimized in conditions where the talker is the same (Caucasian) ethnicity as the participant who speaking a related (Indo-European) language. The issue of match/mismatch between talker ethnicity and language merits closer scrutiny. The level of experimental control used here minimized differences between participant and speaker rather than testing the hypothesis of visual cuing effects on native and non-native speech perception. We conjecture that these effects may be a consequence of heightened perception of speech non-prototypical representations in talkers who are from different cultural, ethnic and language backgrounds.

The hypothesis that speech perception in the Sound-Video condition would be easier than the Video-Sound condition was supported. However, the overall error rates for the Sound-Video condition were still high, albeit better than chance. These results suggest that if participants hear a sound they readily form a visual representation of the sound, although the process is not always accurate, especially for non-native speech (Russian) that is similar to English. This may be due to lower entropy afforded by a non-native stimulus image e.g. ambiguous quality or greater cognitive demands e.g. attentional resources to maintain input information. The fact that performance in the sound-video condition was better than in the video-sound condition suggests that hearing a sound allows participants to evoke a corresponding image/prototype better than translating a visual cue into sound. This seems intuitively plausible but it is not assumed in any model of visible speech. Alternatively, hearing a sound first may allow all the necessary information for identification and this information if not ambiguous and complete is sufficient for identification making visual cues redundant. However, this seems unlikely given that matching interacts with modality of presentation i.e. participants made significantly fewer errors with matched pairs when a sound was presented first compared to when a visual cue was presented first. For both modalities of presentation, identification with mismatched pairs was at chance. This suggests that when participants hear a sound, a corresponding image was available but it will be naturally biased toward categories from the native language, resulting in poor performance for the mismatched pair.

The current results add to the body of knowledge about cross-modal perceptual matching and also extend this literature to the area of non-native speech processing. Kaganovich et al (2016) reported results from an experiment with native speakers matching an auditory word input to video-only input (whilst recording ERP). The researchers assumed that visualizations of articulatory organs will be evoked from VO input. These visualizations are possibly stored in immediate visual memory allowing mapping to auditory input. In the case of a mismatch, such as between a non-native auditory signal and extant representations of articulatory organs,
there is a cognitive cost, resulting from the translation of non-native sounds from articulatory gestures to possible visual images. Alternatively, allocation of cognitive resources between audio and visual input may vary depending on whether the inputs are native or not. Our data indeed showed that response latencies to non-native sounds when followed by a mismatched visual image of a native talker from VO were slower compared to when a native sound was followed by a visual image of a native talker from VO. The asymmetry in identification for native and non-native input across matching tasks is nevertheless consistent with theoretical assumptions about native language prototypes (sound based and image based) biasing stimuli input towards identification of audio and video input as native rather than non-native. Such a bias would be expected in any study of AV speech processing but may be more acute when a talker is perceived as from the same cultural background as the participant, thus making heard and seen stimuli more ambiguous and possibly changing response criteria to a more rigorous standard, thus lowering the sensitivity of the identification tasks overall and also generating a tendency towards making a native language choice in the absence of necessary information.

The current study also contributes to the issue of anticipatory planning with visual attributes of speech. For example, in several studies by Sanchez-García and colleagues (2012, 2013), a visual prime (written sentences) was followed by an AV fragment rather than AO stimulus. Participants were asked to make a match/mismatch judgement and the results showed that when participants were given a visual prime followed by a congruent native AV stimulus, responses were faster but not for non-native stimuli. These results led to the conjecture that visual attributes of speech are redundant for anticipatory planning when they do not exist in the native language. The current study extends their conjecture to the vowel/level and cognate word level, with a particular pair of languages-English and Russian.

In sum, the present paradigms failed to find unequivocal evidence that participants make use of visual aspects of speech when making judgements about native and non-native languages. Although there was some suggestion that under certain conditions participants notice the most dissimilar visual features of a non-native sound compared to native language prototypes, the results show that VO identification (seeing the face of the speaker silently pronouncing stimuli) perception of speech, whether native or non-native, is difficult and the effects of assimilation from visual information are not at all robust.

6.3 **AV speech**

AV speech facilitates speech perception but this effect is context specific (Hazan et al., 2006). Some studies report no facilitating effect whereas others report a context specific effect of AV
Visual perceptual assimilation of non-native speech

speech (Bernstein et al., 2013) (Hazan et al., 2002) and some report a facilitatory effect on non-native speech perception when learning a second language (Hazan et al., 2005). The working assumption in this study was that AV presentation would facilitate non-native speech perception over and above AO. The specific prediction tested however was that VO input would show an independent effect on performance and that perceptual assimilation would be demonstrated for VO input.

Mixed results were found, with AV facilitation observed in some experiments but not others. Specifically, in Experiment 1, participants made more errors in the visual-only condition compared to the AV and auditory conditions but there was no advantage of the AV condition over the AO condition. Additive factors logic leads to the inference that AV presentation is no better than audio alone, at least when sub-lexical units vowel and syllables are to be identified. However, results from Experiment 2 (using more rigorously controlled stimuli) found AV facilitation compared with AO and VO presentation. This finding is qualified by the failure to find the AV facilitation effect when compared to the VO in Experiment One, as performance was no better than chance albeit with an imperfect set of stimuli.

There was, however, a significant interaction between presentation modality and item type on AV facilitation. It was easier for participants to identify dissimilar non-native items when they were presented in the AV compared to the AO condition suggesting that for these items visual input provides significant information for identification performance. Given that performance in the discrimination task was conservative (no better than chance) and therefore not sensitive, it is notable that items in the easier condition (dissimilar non-native items) were identified but only in the AV condition. This suggests that there is AV facilitation effect when identification is difficult and/or a more conservative response criterion is adopted (take a guess rather than a hit).

A similar pattern was observed in Experiment 3, which required a Same-Different judgement for vowels and syllables. There was no difference between AO and AV conditions suggesting no AV facilitation. However, item type interacted with modality of presentation. Specifically, it was easier for participants to judge items as the same when they were presented in the AV condition compared to the AO condition. Furthermore, in the AV condition, there was a bias toward accepting items as belonging to the same language suggesting a shifting criterion that was constrained by modality of input. It can be concluded therefore that AV presentation has some additional influence on performance over and above AO (and VO) presentation.
The results from Experiment 4 provide some insight into the mental representation of AV speech. Participants gave higher ratings (judging items as better English items) for ambiguous items (Similar items) in the AV compared to the AO condition. In contrast, they were more likely to judged Dissimilar Russian items as foreign in the AV mode compared to the AO mode. It is apparent from these results that AV presentation introduced a false sense of nativity more often than in the other conditions when veridical perception of items was not optimal (as the in AO and VO conditions). This is a novel report of false perception in the AV discrimination of non-native speech, specifically for more similar languages, namely Russian-English, albeit limited to suboptimal conditions. An alternative view is that speech identification was influenced by response biases rather than false perception of stimulus features and a signal detection approach to analysis would help to clarify these biases.

The results show that for word and sentence perception, there is unequivocal AV facilitation compared to the VO mode but not compared to the AO mode. In the Experiment 7 pilot study, participants performed near ceiling in the AV condition, suggesting that AV information is more than sufficient for identification of native and non-native speech. This is an important finding. Similarly, in Experiment 8 (identification of real and nonsense sentences in AV and VO conditions) participants made significantly fewer errors in the AV condition compared to the VO condition. In fact, participants performed no better than chance with nonsense sentences in the VO condition. In contrast, they were relatively accurate for the nonsense sentences, in both English and Russian, in the AV condition, suggesting that the AV input disambiguated information sufficiently to allow a comparison with extant speech categories. Such findings are consistent with the view that AV perception of native speech is easier when compared to visual-only/lip-reading of sentences (Grant & Seitz, 1998).

In sum, the results failed to show a consistent AV facilitation effect during speech perception of native and non-native speech. However, the data clearly shows that AV speech processing is superior to VO speech processing, albeit not because of auditory input alone since AV was better than audio-only presentation in some experiments. Although there was little evidence of AV facilitation when compared to auditory-only speech, there were hints that an effect is present but depends on context, specifically type of stimuli and task. The results suggest that AV presentation allows for veridical identification of dissimilar non-native items, arguably the easiest and most ecologically valid condition of all conditions tested in the study because there is sufficient information present from multiple modalities.

Another hypothesis tested here was related to differences in AV speech asynchrony between Russian and English languages. Karpov and colleagues (Karpov et al., 2013), using a talking
head and artificial speech found that English and Russian have different time discrepancies between auditory and visual speech units. Specifically, English is characterised by a greater asynchrony than Russian. Their findings suggest that asynchrony between articulation and sound may be an AV cue for English-speaking listeners who are processing Russian speech. The hypothesis tested here was that native English speakers are sensitive to such information during visual processing of Russian speech, such as sentences. However, there was no support for this hypothesis as participants were at chance for visual sentence processing. The most obvious difference between the present findings and Karpov et al. is use of natural speech compared to synthesised speech and use of a talking head. We note that studies show differences between auditory-lead asynchrony and visual-lead asynchrony, with the latter being more difficult to notice (Conrey & Pisoni, 2006).

Another reason for the lack of asynchrony effect at the syllable/word level here is the way that experimental stimuli were presented. Schwartz and Savariaux (Schwartz & Savariaux, 2014) argued that the lag of an auditory signal behind a visual signal is only specific under certain experimental conditions - at the start of speech sequences or so called ‘preparatory sequences’ of speech. They defined these as the gestures that are visible but produce no sound. This was the case when single vowel/syllables and words here. In contrast, chained sequences of speech, or so-called co-modulatory gestures of speech, are the visual gestures which produced both visual and auditory signals more or less in synchrony. These types of gesture were more likely to be present with presentation of sentences than words and it is possible that participants detected asynchrony in AV speech when presented with the native and non-native sentences, leading to ‘near ceiling’ performance in that condition.

6.4 Implications of the study

The present study was not concerned with perceptual training of non-native speech and there was no consideration of implications for second language learning. However, the findings do have implications for this field. Specifically, the results highlight two novel points regarding non-native speech perception. First, under conditions of limited repetition (up to 4 repetitions per item) and no feedback, participants derive little value from VO speech and appear to have limited ability to notice visual features that are not consistent with their native language. This may have implications for second language teaching (albeit based on null results) in so far as audio immersion may be sufficient for learning the speech features of a non-native language. Secondly, the findings highlight a limitation of AV speech when discriminations are easy and these should be considered when perceptual training is conducted. Several researchers argue
that AV facilitation is not automatic (Bernstein & Auer, 1996; Hazan et al., 2005). The current study extends this view to non-native speech perception when limited exposure is given and no feedback or training is provided. Erdener (D. Erdener, 2016) argued that L2 instruction should include AV material, including online instructional video. Erdener commented that some commercially available online computer-aided language learning (CALL) systems claim to improve L2 pronunciation, the drawback is that the feedback is provided via the speech processing software. However, Erdener points out that the speech clarity and accent judgments made by CALL systems still do not match those made by human listeners. Erdener argued that teaching L2 with orthography combined with AV speech materials can be beneficial, especially if L1 and L2 orthographies match. The current study adds to Erdemer’s call to investigate further the impact of AV materials on L2 acquisition for different linguistic groups and under different conditions.

Another aspect of the study, which has a unique component compared to AV L2 studies, was the ranking of AV items. The results suggested that modality of presentation impacts on how native and non-native speech is perceived. Participants judge similar non-native items as better exemplars of English if they are presented in the AV format compared to the AO. In contrast, they are more likely to judge dissimilar foreign items as less English when they are presented in the AV mode. Practical implications of this result may be that processing English speech from a fluent bilingual speaker, and possibly even accented English speech will be easier in the AO mode. On the other hand, the fact that certain dissimilar items were judged as more foreign in the AV mode may imply that tuning or possibly focusing attention on specific aspects of non-native speech is easier in the AV format.

Another distinctive aspect of the current study was investigation of AV and VO processing of cognates. Given the cognates were pronounced differently, it was of interest that participants failed to notice visual correlates of differences in the VO mode, as is evident in participants accepting visual images as native. This is important, as cognates are often the first words to be acquired during second-language acquisition (at least for adults). The current findings suggest that cognates do not have privileged access into mental representation of visual non-native speech, at least within the limited presentation rate used in the current study.

The difference between video-sound and sound-video cross-modal matching suggests that the process of ‘translating’ the visual correlates of non-native speech into auditory speech is not effective, at least under current experimental conditions. In contrast, the process of translating sound to visual correlates is more effective. Moreover, there is a benefit in providing a written cue, even for non-native speech, as it facilitates mapping to prototypes/exemplars of native
tive assimilation of non-native speech

speech. Contrast between native and non-native stimuli facilitates discrimination, as matched pairs generally result in better performance. These are important findings in relation to early stages of acquiring new language for naïve participants, which implies that certain methods of instruction or training (e.g. incorporating written input, providing matched versus mismatched language contrasts) may be beneficial. In addition, the results showed that slowing the rate of presentation does not facilitate non-native speech processing when performing a cross-modal matching task. Instead, participants showed they were more adept at processing quick stimuli, even non-native speech, compared to slow speech. This suggests that perceptual adaption to rapid speech is robust, even in the very early stages of encountering a new language.

Finally, the data contribute to knowledge about cross-linguistic comparative studies. Cross-linguistic differences are commonly reported and the current study was the first to investigate systematically VO and AV processing of the Russian language by naïve Australian English-speakers. The results of the current study can be contrasted with other cross-linguistic studies showing that it is possible to discriminate language based on the visual attributes of speech. It also provides impetus for further comparative studies with language pairs from unrelated and related language families more detailed analysis of the reasons for the observed differences.

6.5 Future direction

The study used an ecologically valid paradigm with natural speech stimuli. However, this did not allow the same precision of investigation as given with synthetic stimuli (e.g. (P. Kuhl, 1995). A similar design with visual stimuli could identify best and worst exemplars of a sound for future investigation of visemes and visual assimilation/visual magnet effects. This could allow a more detailed investigation of non-native speech processing, similar to experiments in the domain of auditory speech processing. A difficulty with adopting this approach, however, is how to establish a system of 'visemes' for native and non-native languages of most interest. Further issue is related to the experimental paradigms used in the current study. Whilst most experiments in the current study replicated previously used experimental paradigm (eg. language identification, same-different, match-mismatched paradigms), none of them included asking participants to recall what is being said. The latter exerts additional demands on immediate and short-term memory but provides further insights into processing of native and non-native speech.

Another issue is that the current study used ranking of stimuli for auditory-only and audio-visual stimuli. Ranking of visual-only speech may be informative when doing cross-linguistic studies, especially considering bilingual speakers may have a ‘visual accent’. Another issue is
Visual perceptual assimilation of non-native speech

individual variability in lip-reading ability and ideally individual measures will be correlated with performance in visual-only conditions. Several studies use this design (e.g. Kaganovich et al., 2016) and measure a baseline lip-reading ability that is correlated with other measures (e.g. accuracy, response time). This study presented one bilingual speaker and ideally several speakers would be used to control speaker effects (e.g. intelligibility) that are often reported (e.g. (Yehia et al., 1998). Better control for the degree of bilingualism of a speaker e.g. use of independent professional bilingual linguists as judges for rating the speaker’s input would also control for possible confounding variables.

Another issue is that Australian English-speaking participants only were tested, while cross-language studies are generally more informative. Specifically, a double cross-language study of English and Russian-speaking participants would be informative in clarifying results here so as to exclude possible artefacts (e.g. the effect of speaker) or to more generalised effects (e.g. effects of language background). A pilot-study conducted with a small group of bilingual English-Russian-speakers (unreported), found similar trends to the native English-speakers. Therefore, future studies could use cross-language manipulations with speakers of languages from different language families matched across different variables. In addition, studies adopt a longitudinal approach in to test whether language proficiency impacts on speech processing. In the case of the current study, naïve participants only were tested and the possibility remains that they would perform differently with the same stimuli if the L2 proficiency were greater.

The current study used ecologically valid clear speech whilst many studies have demonstrated an AV effect in noisy or suboptimal conditions. Using different contexts, such as pure speech, or filtered speech (Burnham et al., 2015) would test the effect under different conditions.

Furthermore, exploring individual differences between participants was not investigated in the current study. The topic of individual differences in perception of visual speech (speech reading) is not well investigated, despite many studies in both the fields of speech reading and AV speech reporting on individual differences between participants (MacLeod & Summerfield, 1990).

In recent years, the topic of individual differences has been gaining new momentum, with a number of studies looking specifically into the reasons for such variability (Agnès Alsius, Wayne, Paré, & Munhall, 2016). The majority of AV speech studies look into speech input under degraded condition, embedded in noise (speech-in-noise tasks; SPIN). Within the field of speech reading, spatial working memory and verbal processing speed have emerged as some of the cognitive factors which could account for individual differences (Feld &
Sommers, 2009; Lidestam, Lyxell, & Andersson, 1999). Another factor implicated in the ability to speech read is the capacity for perceptual synthesis (Watson, Qiu, Chamberlain, & Li, 1996).

New techniques, such as a spatial frequency filtering technique, provided new insights into potential factors contributing to individual variability in speech reading. The technique allows deletion of specific spatial frequencies of the image as a type of manipulation of the resolution of silent speech (Munhall, Kroos, Jozan, & Vatikiotis-Bateson, 2004; Wilson, Alsius, Pare, & Munhall, 2016). For instance, Wilson and colleagues (Wilson et al., 2016) presented 2 groups of participants (one with the highest speechreading ability, the other with the lowest speech reading ability) with silent clips of a speaker pronouncing syllables processed with spatial frequency either filtering or unfiltered. The result showed that the group that were very good at speech-reading, were more adversely impacted upon by the removal of high spatial frequency information compared to poor speech readers. Wilson and colleagues concluded that individual differences are at least partially due to the ability to extract fine-detailed facial information from the visual image. They hypothesised that this is due to superior ability in mapping optical signal to visual speech representation (Wilson et al., 2016). How important are fine facial detail for individual variability in AV speech perception? Some studies which did not examine individual differences showed that facial detail input is not as essential when auditory input is also present (MacDonald, Andersen, & Bachmann, 2000; Munhall et al., 2004).

Alsius and colleagues (2016) specifically looked at the topic of individual differences in AV speech in noise recognition. They utilised low-pass spatial frequency filtering together with eye tracking and compared participants across the AV benefit factor. Their participants were divided into two groups-the first was the group that benefited from visual information the most, and the other group was the one with the last benefit from the visual information. Alsius and colleagues also used two levels of linguistic complexity, isolated words and sentences. They first screened all participants for speech-reading ability of isolated words and sentences when stimuli was presented unfiltered under visual-only mode. They collected response distributions for the audio-visual unfiltered, auditory-only condition, the visual-only condition, and AV benefits scores. Their results showed that the participants who benefited most from the addition of visual information were more negatively impacted upon by the removal of high spatial frequency information compared to the group which showed low visual gain. These effects were maintained across two linguistic levels, words and sentences. The results also showed that individual differences as a function of gaze fixation were observed. The group that showed the highest AV benefit fixated longer on the mouth region.
compared to the other group. The researchers interpreted their results as evidence that efficient use of fine facial detail information and longer gaze fixation on the mouth regions for words underlie individual differences in AV speech perception in noise. Thus, for some individuals, as well as degraded auditory input, visual input can also lead to reduction in AV speech perception. Extending the finding of Alsius and colleagues to the field of second language acquisition would facilitate greater understanding of individual differences that underlie speech-reading and AV processing of non-native and native speech.

Another important factor for future investigation to emerge from the current study is the type of stimuli presentation. Whilst the study manipulated the speed of stimuli presentation, no manipulation was present to exaggerate non-foreign speech. Instead natural speech was used without specific instructions for the speaker to exaggerate certain aspects of non-native speech. There is some indication of the value of hyper-articulation in perception of foreign speech and this could be potentially explored in the future. The study of Uther and colleagues (2007) compared infant-directed speech to foreigner directed speech. They compared British women’s interaction with their infants, foreign adults, and control British females. They found that vowel hyper-articulation was present in both infant-directed and foreigner-directed speech. However, foreigner-directed speech had significantly less positive and more negative affect than infant-directed speech, as well as lower voice pitch. The findings were interpreted as supportive of Kuhl’s view (2010) of vowel hyper-articulation as emphasising extremes of the language-specific phonetic prototypes. Uther and colleagues (2007) suggest that hyper-articulation in foreign-directed speech serves a similar purpose by emphasising differences between non-native vowels. Similar results were obtained by Papousek and colleagues (1991) who found exaggerated pitch contours and pitch range in a tonal foreign-directed speech. In contrast, Biersack and colleagues (2005) showed no increases in pitch range in foreign-directed speech.

Another important element in the current study is the use of a single speaker. There are also neuro-behavioural studies that show the benefit of using a consistent speaker. Mani and Schneider’s study (2013) asked the question of whether consistent pairing of different speakers’ faces with different sounds impacts on discrimination of the sounds. Using brain ERP results, the study compared data from the participants previously exposed to consistent speakers to groups of participants who had inconsistent pairing. The results showed that phoneme detection was improved when consistent speaker-sound pairing was used. The researchers concluded that there is a strong influence of visual speaker identity in speech processing. Future studies should ideally include both single and multiple speakers to control for speaker effects.
In addition, language familiarity as an important factor warrants further future investigation. One explanation for differences between the present findings and reports of VO effects (e.g., Ronquest et al., 2010; Soto-Faraco et al., 2007) is language familiarity. Testing languages with lesser and greater familiarity for Australia English speakers would clarify whether language familiarity does contribute to the current findings. Furthermore, none of the experiments used here disambiguated the foreign language. Including such a condition as a control could clarify how much knowledge of the language may be contributing to the results. A similar criticism could be made about temporal manipulations of speech, such as informing participants about the reversal of stimuli, which has an impact on performance strategies (Ronquest, 2010). In relation to language familiarity, there was no control over word frequency/familiarity, as well as word salience. Controlling for these factors could be informative (e.g., Soto-Faraco, 2007). Furthermore, the study used cognates but did not contrast these with non-cognate words. The latter control would clarify whether there is a ‘reversed’ cognate effect, such as that observed in some acoustic studies (Kouačević, 2012; Temnikova & Nagel, 2015). Cognate effects depend on language proficiency (Midgley et al., 2011) and including bilinguals in addition to monolinguals could provide some answers regarding this issue. The combination of behavioural and physiological studies, such as that described in Midgley and colleagues (2011) and in Bice and Kroll (Bice & Kroll, 2015) but including AV and visual-only presentations may also shed light on processes that are not observable through behavioural measures alone, but can be seen in neural responses.

Whilst the current study attempted to tap into the differences asynchrony of speech inherent in Russian and English, it did not specifically vary the asynchrony of stimuli. The latter could be useful in tapping into the underlying processes involved in AV integration (Navarra et al., 2010).

Finally, the processing of non-native speech was not facilitated by slow presentation here and this merits further investigation. Zhao (Zhao, 1997) reported effects of listener control over speech rate during L2 comprehension by allowing participants and argued that speed of presentation needs to match individual internal ‘reference’ of slow speech. Including such an experimental condition may also uncover the effect, which was not evident in the design of this study.
6.6 Conclusion

This study explored perception of native and non-native speech, drawing on auditory speech perception studies showing that native speech is perceived via sound categories with coherent structure and that non-native speech is assimilated according to the same representations. The design of the study included a pair of languages, English and Russian, from the same family (Indo-European), not previously compared in the AV and visual speech domains. The study validated non-native (Russian) and native (English) experimental stimuli, consisting of vowels and syllables based on their similarity. The study replicated findings using auditory presentation, showing that non-native stimuli are assimilated according to similarity to native categories. However, the hypothesis that similar constructs are used for the processing of visual aspects of speech was not supported, thus finding no consistent and systematic evidence for visual perceptual assimilation. Furthermore, the results demonstrated a limited benefit of AV speech compared to auditory speech for non-native speech processing. The study also investigated visual perception of cognate words and sentences and found little evidence for language discrimination based on the visual aspects of speech. In fact, participants were unable to discriminate real from temporally reversed speech in the VO condition. Moreover, they were unable to reliably translate a visual image into the corresponding auditory sound. Cognates did not show a pattern of privileged access to a native mental lexicon in relation to the visual aspects of cognate processing. The study also drew attention to the importance of orthographical input in the very early stages of encountering a new language. It showed that there was a benefit in providing a written native cue/prime when processing non-native visual aspects of speech. It also highlighted the fact that the presence of a contrast between native and non-native stimuli facilitates perception, including in the very early stages of L2 acquisition. In addition, it uncovered an interesting bias towards accepting non-native speech as more native in the audio-visual mode compared to the auditory-only mode. The study also demonstrated that naïve participants performed better with normal and quick speech than slow speech, showing robustness of perceptual adaptation to time-modified speech. Finally, the study highlighted the need for further cross-linguistic comparative studies exploring visual and audio-visual speech for novel language pairs.
References


Faris, M. M., Best, C. T., & Tyler, M. D. (2016). An examination of the different ways that non-native phones may be perceptually assimilated as uncategorized. *The Journal Of The Acoustical Society Of America, 139*(1), EL1-EL1.


Visual perceptual assimilation of non-native speech

Visual perceptual assimilation of non-native speech


Visual perceptual assimilation of non-native speech


Visual perceptual assimilation of non-native speech


Appendix A: Instructions

Language identification experiments instructions
Imagine that you are a secret agent who intercepted coded (audio/video/audio-video) messages sent by a spy. Some of the coded messages consist of English sounds while others include non-English sounds. The first step in decoding them is to sort them into English and non-English items. If the item has only English sounds, press the right YES button. If the item has non-English sounds, press the left NO button.

During the practice you will see a person’s mouth and the items will be all English sounds. For example the first practice item is CHA. This item consists of the sounds CH and A. Both are the English sounds so press the YES button.

This is the end of the practice. In the following experiment you will see another person’s whole face. When you hear items with non-English sounds press the left NO button. When you hear items with only English sounds, press right YES button.

Congratulations, you are half way through. Press the black switch to continue.

Thank you, this is the end of this part.

Matching experiments instructions
In this experiment you will see a bilingual person saying both English and non-English sounds. These sounds will be presented in pairs and will be either the same or different. For the practice you’ll hear a non-English sound SI followed by the same sound. Since they are the same, press the YES button. Then you will hear the same non-English sound SI followed by the English sound SA. Since they are different, press the NO button.

This is the end of the practice. In the following experiment you will hear English and non-English sound pairs. Each pair will be presented several times over the course of the experiment. If the sounds in each pair match, press the YES button. If the sounds do not match, press the NO button.

For this experiment we filmed a bilingual person saying both English and non-English sounds. We then separated the sounds from the video image and paired them so that some of them match while others do not.
For practice, you will hear the non-English sound/see a matching pair of the video SI followed a matching image of this sound. Since they match, you have to press the YES button. Then you will hear the same non-English sound SI followed by the image of the English sound SA. Since the image does not match the sound, press the NO button. This is the end of the practice. In the following experiment you will hear English and non-English sounds. Each of them will be repeated several times over the course of the experiment. If the image matches the sound press the YES button. If it does not match the sound press the NO button.
Appendix B: Stimuli with IPA transcription

<table>
<thead>
<tr>
<th></th>
<th>Russian</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowels</td>
<td>ə (identicial), ɨ (similar), i (dissimilar)</td>
<td>ʌ(Identicial), i:(Similar), æ(dissimilar)</td>
</tr>
<tr>
<td>Consonants</td>
<td>m(Identicial), s(similar), r (dissimilar)</td>
<td>m(Identicial), s(similar), r(dissimilar)</td>
</tr>
<tr>
<td>VC</td>
<td>im, is, ir, am, as, ar, im, is, ir</td>
<td>æm, æs, ær, am, as, ar, i:m, i:s, i:r</td>
</tr>
<tr>
<td>CV</td>
<td>mi, si, ri, mà, sa, rà, mi, si, ri</td>
<td>mæ, sæ, ræ, ma, sa, ra, mi:, si:, ri</td>
</tr>
<tr>
<td>CVC</td>
<td>mim, sis, cir, màm, sàs, ràr, mim, sis, cir</td>
<td>mæm, sæs, rær, màm, sàs, ràr mi:m, sì:s, rì:r</td>
</tr>
</tbody>
</table>

English Cognate Words

FLAG, CHAOS, TENNIS, TEXT, FILM, TALC
ZEBRA, CRAB, CRITIC, OPERA, TRANSPORT, FACTOR
BULLETIN, GALLOP, DELEGATE, DIPLOMAT, INSTITUTE, CATALOGUE
ARGUMENT, GRANITE, IMMIGRANT, INTERVAL, CARAVAN, VETERAN
CLASSIFICATION, INTONATION, SITUATION, COALITION, SANCTION,
EVACUATION, DISCRIMINATION, AGGRESSION, TRADITION, PROFESSION,
RESOLUTION, REPRODUCTION

Russian Cognate Words

Флаг, хаос, теннис, текст, фильм, тальк.
зебра, краб, критик, опера, транспорт, фактор.
Бюллетень, галоп, делегат, дипломат, институт, каталог.
аргумент, гранит, иммигрант, интервал, караван, ветеран.
классификация, интонация, ситуация, коалиция, санкция, эвакуация.
дискриминация, агрессия, традиция, профессия, резолюция, репродукция

English Sentences

(from Revised List of Phonetically Balance Sentences (Harvard Sentences), 1969. Lists 11 and 12)

1. Oak is strong and also gives shade.
2. Cats and dogs each hate the other.
3. The pipe began to rust while new.
4. Open the crate but don’t break the glass.
5. Add the sum to the product of these three.
6. Thieves who rob friends deserve jail.
7. The ripe taste of cheese improves with age
8. Act on these orders with great speed.
9. The hog crawled under the high fence.
10. Move the vat over the hot fire.
11. The bark on the pine tree was shiny and dark.
12. Leaves turn brown and yellow in the fall.

Russian Sentences
Дуб крепкий и даёт тень.
Собаки и кошки ненавидят друг друга.
Труба начала ржаветь ещё новой.
Откройте ящик, но не сломайте стекло.
Добавь эту сумму к результату этих трёх.
Воры, которые обваривают друзей заслуживают тюрьму.
Хороший вкус сыра улучшается со временем.
Действуйте по этим указаниям с огромной быстротой.
Ёжик подполз под высокий забор.
Подвинь этот кателок над горячей водой.
Кора от сосны была блестящая и темная.
Осенью листья становятся коричневые и жёлтые.