Visual Discrimination of French and English in Inter-Speech and Speech-Ready Position

by

Joseph Paul D'Aquisto

B.A. Honors, Linguistics & Russian, University of Arizona, 2011

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

The Faculty of Graduate and Postdoctoral Studies

(Linguistics)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

August 2014

© Joseph Paul D’Aquisto, 2014
Abstract

This study investigates the ability of observers to discriminate between French and English using visual-only stimuli. This study differs from prior studies because it specifically uses inter-speech (ISP) and speech-ready tokens rather than full sentences. The main purpose of this research was to answer if observers could successfully discriminate French from English by watching video clips of speakers engaged in ISP and speech-ready positions with the audio removed. Two experiments were conducted; the first experiment focuses on native English vs. non-native English speakers and the second experiment focuses on native English vs. native French speakers which expands further on the data in the first experiment. The results support the view that observers can visually distinguish their native language even in the absence of segmental information.
Preface

All of the work presented in this thesis was conducted at facilities of the University of British Columbia Department of Linguistics. This research was conducted with approval by the University of British Columbia’s Research Ethics Board as part of the research project entitled “Processing Complex Speech Motor Tasks” under the certificate numbers: H04-80337 and B04-0337 supervised by principal investigator Bryan Gick. I was the lead researcher on this part of the project and responsible for all major areas of concept formation, data collection and analysis, as well as the majority of manuscript composition. Bryan Gick was my supervisor and additional committee members were Eric Vatikiotis-Bateson and Rose-Marie Déchaine who were involved throughout the project in discussion and manuscript edits.
Table of Contents

Abstract ................................................................................................................................................ ii
Preface .................................................................................................................................................. iii
Table of Contents ................................................................................................................................. iv
List of Tables ......................................................................................................................................... v
List of Figures ......................................................................................................................................... vi
Acknowledgements ........................................................................................................................... vii

1 Introduction ....................................................................................................................................... 1
  1.1. Background on ISP .................................................................................................................. 3
    1.1.1 Ultrasound Studies of ISP: Gick et al. 2004 ................................................................. 3
    1.1.2 Ultrasound and Optotrak Studies of ISP: Wilson 2006 ........................................ 4
    1.1.3 Ultrasound Studies of Bilingual-Mode ISP: .............................................................. 4
    1.1.4 Electropalatographic Studies of ISP: Schaeffler et al. 2008: .................................. 5
    1.1.5 MRI Studies of ISP: ....................................................................................................... 9
  1.2. Visual Speech Information From Facial Movements ............................................................ 13
    1.2.1 The Contribution of Visual Speech Information ..................................................... 13
    1.2.2 Language Identification From Visual-Only Cues ...................................................... 14
    1.2.3 The Bilingual Advantage in Language Identification of Visual Stimuli .............. 16
    1.2.4 Increased Cognitive Load of Visual-Only Speech................................................... 17
  1.3. Linguistic Information From Facial Movements & Facial Recognition ......................... 18
  1.4 Goal & Questions of This Study ............................................................................................. 22

2 Native/Non-Native Perceivers Pilot ............................................................................................ 24
  2.1. Methods .................................................................................................................................. 24
  2.2. Results .................................................................................................................................... 27
  2.3. Discussion ............................................................................................................................... 32

3 Native English vs. Native French Perceivers ........................................................................ 34
  3.1. Methods .................................................................................................................................. 35
  3.2. Results and Discussion .......................................................................................................... 38

4 Conclusion ..................................................................................................................................... 47

References .......................................................................................................................................... 49
List of Tables

Table 2.1  Tokens Identified by at Least 5 Out of 6 Participants..........................31
Table 3.1  T-Test Results for English & French Groups........................................40
Table 3.2  Tokens Correctly/Incorrectly Identified ..................................................45
List of Figures

Figure 1.1   ISP% for Speakers ..........................................................7
Figure 2.1   Correct Tokens Overall for English & French Groups ...........28
Figure 2.2   Overall % of Correct Tokens by Speaker & L1 Language........29
Figure 3.1   Accuracy % by Native Language ........................................40
Figure 3.2   Confusion Matrix for Subjects Responses ..........................42
Figure 3.3   Mean Accuracy Rate by Stimulus Language ........................44
Acknowledgements

I am grateful to the faculty and staff in the department of Linguistics at the University of British Columbia who have provided me with the skills to work in this field. My supervisor Bryan Gick and committee members Eric Vatikiotis-Bateson and Rose-Marie Deschaine have been great mentors during my time at this institution.

Thanks to my fellow students, researchers and research assistants in the Interdisciplinary Speech Research Lab.

I would also like to give a special thanks to Kyle Danielson from UBC Psychology and Phoebe Wong from UBC Audiology for providing assistance with PsyScope, R, and statistics.
Chapter 1

Introduction

Prior research suggests that adults can discriminate between different languages based solely on visual signals. Ronquest et al. (2010) explored visual cues that observers use to complete language-identification tasks in Spanish and English and found that observers could detect rhythmic differences in syllable-timed vs. stress-timed stimuli. Studies by Weikum et al. (2007) and Weikum et al. (2013) found that there is a correlation between the age of acquisition and the ability to differentiate one's own language from other languages using visual-only cues. Weikum et al. (2007) examined infants ability to discriminate French from English using silent video clips. 4 month old native English speaking infants were studied to see if they could differentiate their native language (English) from an unfamiliar language (French). The 4 month old infants were also compared against 6 to 8 month old monolingual English and bilingual English-French to determine how perception accuracy is affected by age. The stimuli used were sentences recited by three bilingual French-English speakers in each language. Looking time was the method used in order to determine if infants were able to differentiate the languages. If an infant saw trials of different sentences from each language of the same speaker and their looking time increased, this indicated they had noticed the language change. The results from Weikum et al. (2007) showed that 4 and 6 month old monolingual infants looked significantly longer at language switch trials than the 8 month old monolingual infants. This finding supports that infants can visually identify their
native language from an unfamiliar language at 4 and 6 months, but not at 8 months. It was also found that of the 8 month old infants, only the bilingual ones looked significantly longer at the language switch. This means that after 8 months of age, infants ability to discriminate a familiar language vs. a non-familiar declines. The fact that of the 8 month old infants, only the bilinguals who were familiar in both languages had the ability to discriminate French from English supports the previous statement. Weikum et al. (2013) expands on their Weikum et al. (2007) and explored how age of acquisition affects adults ability to discriminate English from French using visual-only stimuli. They used video clips of three balanced French/English bilinguals reciting French and English sentences from the French and English versions of the book “The Little Prince” that ranged 8 to 13 seconds in length. Each of the clips had the sound removed after recording. Weikum et al’s (2013) data showed that adults who had learned English as a first or second language between the ages of 0 to 6 were able to judge whether a speaker appearing in a silent video clip was French or English. Adults who had learned English after the age of 6 failed to discriminate between the two languages. None of the participants had any knowledge of French.

The initial goal of this thesis was to replicate the study by Weikum et al. (2013) using inter-speech and speech ready stimuli as opposed to full sentences. *Inter-speech posture* (ISP) is typically described as the brief pauses that occur between speech when a speaker is in the act of an utterance; *speech-ready position* is an articulatory posture that a speaker assumes when he/she prepares to speak (Ramanarayanan et al. (2013), Gick et al. (2004)).
1.1. Background on ISP

This section will highlight some of the previous research and methods exploring ISP which employ the use of several technologies including: Ultrasound, X-Ray Optotrak, EPG and MRI. Topics explored are the causes that contribute to ISPs and when they happen. How these properties differ across speakers in general are also explained, in addition to significant differences in bilinguals vs. monolinguals. The background is meant to give a history and describe the methods, theories and evidence surrounding ISP.

1.1.1 Ultrasound Studies of ISP: Gick et al. 2004

Gick et al. (2004) explored the existence of a ‘default setting’ or ‘posture’ for articulators and facial muscles when speaking a particular language. They manually measured x-ray films for the following data: pharynx width, velic aperture, tongue body distance from the hard palate (tongue dorsum constriction degree), tongue tip distance from the alveolar ridge (or tongue tip constriction degree), lower-to-upper jaw distance, upper and lower lip protrusion.

Another question raised was whether this default posture is specified as a language’s inventory which is learned from other speakers or functionally derived properties of speech motion. In other words, if there is a default posture it may be part of a language’s inventory or specified target that is held throughout the utterance while being available to learners uninterrupted through the acoustic
signal or it may be used as a rest position that is a feature or property of that language (language-specific). If it is not part of the language inventory and is functional then this could have to do with motor control, for instance a speaker using a large amount of postvelar sounds could display more retraction in their articulators.

1.1.2 Ultrasound and Optotrak Studies of ISP: Wilson 2006

Wilson (2006) explores ISP in Canadian French and Canadian English speakers. Two experiments were carried out in Wilson (2006) using Optotrak and ultrasound imaging in order to address the question of whether ISP is language specific in both monolingual and bilingual speakers and if it is influenced by phonetic context and/or speech mode (monolingual or bilingual). The results from Wilson (2006) show significant differences among the two groups of speakers in regards to position of the articulators. In addition, Wilson's (2006) data also lend support to the notion that there is no differentiation between ISP in bilingual-mode, but favors the idea that a bilingual’s ISP is the same ISP of a speaker's currently most used language.

1.1.3 Ultrasound Studies of Bilingual-Mode ISP

Wilson & Gick (2013) define bilingual-mode as what occurs when a bilingual speaks with another bilingual when both languages are being used, whereas monolingual-mode is described as what occurs when a bilingual is speaking to a monolingual with only one language being used. Additionally, it is also possible to have two
bilinguals speaking to one another using only one language. Wilson & Gick (2013) further tested the question proposed in Wilson (2006) of whether ISP is language specific in both monolingual and bilingual speakers by having eight French-English bilinguals read English and French sentences. The participants’ ISPs were measured with optical tracking of the 3D positions of the lips and jaw while ultrasound imaging was used to track tongue movements. The results from Wilson & Gick (2013) reinforce the hypothesis in Wilson (2006), namely that bilinguals use the ISP of their most-dominant language when in bilingual-mode.

1.1.4 Electropalatographic Studies of ISP: Schaeffler et al. 2008
Schaeffler et al. (2008) used electropalatographic (EPG) data to provide information on tongue-palate contact patterns during speech and non-speech activities such as swallowing and bracing. ISP data were taken from three different tasks performed by English speakers:

1) a read-speech task with single words presented on screen;
2) a picture-naming task with pictures presented on screen;
3) a semi-spontaneous map task that required speakers to describe a simple route to a listener.

ISP data were recorded by identifying significant change in overall contact pattern from the time the prompt appeared to the acoustic onset of speech. ISP was noted when the transition from pre-prompt (non-speech) position and the first speech gesture was neither an interpolation nor a random movement. For the read-speech and picture-naming tasks, data was gathered after an audible beep that
lasted 2, 5 or 8 seconds before the orthographic or picture prompt appeared, continually until after the end of the acoustic output. For the map task, data was gathered continuously.

Schaeffler et al. (2008) mention that EPG works well for cases such as when speakers hold part of their tongue against the roof of their mouth during non-speech, but other tools are needed such as ultrasound for a more detailed understanding. Schaeffler et al. (2008) asked whether ISP’s occurring in spontaneous pauses happen more or less often than those occurring in the prompted pauses and how long a pause has to be to give rise to a measurable ISP. A map task was used in order to determine this. To classify ISPs Schaeffler et al. (2008) first identified whether any notable change occurred in the overall contact pattern between the onset of the prompt and the onset of the acoustic response. An ISP zone was categorized when the transition between the pre-prompt (non-speech) position and the first speech gesture was neither a mere interpolation nor random movement. An ISP was identified in this phase if the kinematic record indicated a motion towards some configuration 1, followed by smooth movement away from it towards the first segment 2, or a clear pause during a continuous motion. Schaeffler et al. (2008) explored how different conditions affect the formation and dynamic structure of the ISP. Of particular interest was whether ISPs occurring in spontaneous pauses occur more or less often than those occurring in prompted pauses, and how long a pause has to be in order to a have measurable ISP. The figure below is an illustration from Schaeffler et al. (2008) showing the percentage of ISPs for speakers in all three tasks.
Figure 1.1 - % ISP for Speakers for the 3 Tasks. Illustration taken from Schaeffler et al. (2008). This shows the proportion of pauses with ISP zones to total number of inter-speech pauses. 86 ISP zones were identified in 140 pause tokens (61.4% of tokens). ISP zones began approximately half a second after presentation of the prompt, 454 ms before the acoustic onset (s.d. 275 ms). There were significant differences between speakers (one-way ANOVA, F(2,85) = 6.892, p=.002; cf. Table 1) but not between speech tasks (i.e. picture naming vs. word list) or following segmental context (i.e. alveolar vs. non-alveolar vs. vocalic onset). Schaeffler et al. (2008) mention that between-speaker differences in speech rate may have had an affect here, but believe suspect habitual differences between speakers were more likely to explain these results.

Schaeffler et al. (2008) investigated the variability among speakers and found that one speaker least familiar to wearing an artificial palate kept his tongue pressed against his palate when waiting for each prompt and when pausing naturally in the map task while another speaker showed a rest position with little to no tongue-palate contact which made the process of identifying ISPs for that speaker harder to detect. Schaeffler et al. (2008) were thus unable to estimate the difference between spontaneous vs. prompted speech on formation of ISPs and no clear conclusion could be drawn in terms of which task elicits reliable ISPs.
Wilson (2006) investigates whether ISP is language-specific in both monolingual and bilingual speakers of Canadian English and Québécois French using Optotrak and ultrasound imaging. Two experiments tested how ISP is related to phonetic context and speech mode (bilingual or monolingual). Wilson (2006) show significant differences in ISP across English and French monolingual groups with English monolinguals having a higher tongue tip, more protruded upper and lower lips, and more narrow horizontal lip aperture. An ultrasound monitor was used to view tongue movements in real time along with an Optotrak (Northern Digital Inc.) 3020 optical tracking system that measured the 3D positions of the lips, jaw and head relative to the ultrasound probe. Optotrak numeric data and ultrasound videos were the data sources measured which was processed via MATLAB, however data had to be pre-processed first. The first step in this required the DV Ultrasound tape to be converted to Adobe Premiere movie files which were resampled in order to ensure later measurements were all on the same scale. Ultrasound movie files were then cropped so that the first frame for each file was the frame immediately following after a clapper was heard in order to establish which Optotrak frames corresponded to the ultrasound frames of interest. Possible periods of rest used for analysis were found by replaying the ultrasound movie files and searching after every sentence for a period of at least 10 frames of no tongue motion. The reasoning behind choosing a 10-frame period vs. a longer or shorter period, was that a 10-frame period was the longest possible rest period where the tongue was considered to be at rest in an average of about 50% of the ISPs across all 24 subjects. If such a period of 10 frames of no tongue motion existed, then the
centre frame of that period was chosen as a "possible rest frame" for analysis, Wilson (2006). English speakers’ jaw ISPs were found to be partially influenced by phonetic context, but the lip and tongue ISPs were not. Wilson (2006) found there was not a significant difference between French and English ISP measurements for the jaw and velum. Upper and lower lip protrusion were greater for English ISP than French ISP in bilinguals perceived as native speakers of both languages, but not bilinguals who weren’t perceived as native in both. The tongue tip, tongue body and tongue root were all farther away from the opposing vocal tract surface in the French group than in the English group.

Wilson (2006) note that variation in anatomical size and proportion could be an explanation. Some of the other factors for this difference between the two language groups could also be thing as: higher tongue tip in English due to sides of tongue being tethered to the roof of mouth and molars; French jaw being open more and more widely vs. English because of high frequency of [a] in French compared to English; English lips being neutral whereas French lips are more rounded and active in spreading and rounding. The factors accounting for the differences of articulators in French and English go beyond the scope of this paper, but they are worth noting.

1.1.5 MRI Studies of ISP

Ramanarayanan et al. (2013) explore human speech production using real-time magnetic resonance imaging (MRI) of the vocal tract. Ramanarayanan et al’s. (2013) procedure extracted frames correlating to ISP pauses, speech-ready and absolute
rest position from MRI sequences of speech read by 5 English speakers. In addition the procedure extracted image features that were used to measure vocal tract posture at these time intervals. Their analysis determined that there are significant differences between vocal tract posture during ISP and absolute rest position before speech. The results from Ramanarayanan et al. (2013) lend further support to the idea that vocal tract positions differ during positions at rest, speech-ready and ISP. Ramanarayanan et al. (2013) state that the default setting can be defined as the set of postural configurations that vocal tract articulators tend to be deployed from and return to in the process of producing fluent and natural speech; in addition they can be language-specific or speaker-specific. A postural configuration can vary, for example, keeping lips in a rounded position or keeping the tongue retracted into the pharynx throughout an entire speech utterance.

Ramanarayanan et al. (2013) ask what articulatory or acoustic variables are used to obtain these postures? Prior to their paper, the manner of control used by the speech “planner” during the execution of these postures had not been addressed yet in a comprehensive manner using speech articulation data. Also, Ramanarayanan et al. (2013) focus on understanding speech within spoken American English while considering the effects of speaking style (read vs. spontaneous) and position within an utterance and analyzing its postural motor control characteristics. Postures occurring in silent pauses before speech (speech-ready and absolute rest) and during speech are examined in order to eliminate most of the factors that may happen due to articulatory postural variations that are required to produce speech sounds. For example, one speaker may display a wider
opening of the mouth to produce /r/ sound while another speaker may have a much narrower opening when producing that same sound, or employ the use of a different muscle or movement.

As previously mentioned, Gick et al. (2004) claimed the existence of a language-specific default position and also that speech rest positions are specified in a manner similar to actual speech targets. Ramanarayanan et al. (2013) mention that further analysis of positioning within an utterance and speaking style could have important implications for understanding the speech motor planning process. Gick et al. (2013) analyze vocal tract posture in order to answer the following questions: (1) Do articulatory postures occurring in grammatical ISP pauses differ from those in an absolute rest position and also from speech-ready posture (or pre-speech posture)?, (2) What can be concluded regarding the degree of active control exerted by the cognitive speech planner (as measured by the variance of appropriate variables that capture vocal tract posture) in each case?, and (3) Do articulators vary between read and spontaneous speech? The question regarding read vs. spontaneous speech builds upon a previous study by Ramanarayanan et al. (2009) that explored the hypotheses that pauses at major syntactic boundaries (i.e., grammatical pauses), but not ungrammatical (e.g., word search) pauses, are planned by a high-level cognitive mechanism while also controlling the rate of articulation around these areas. In that study MRI was used to measure articulation at and around grammatical and ungrammatical pauses in spontaneous speech. Ramanarayanan et al. (2009) found that grammatical pauses were found to have an appreciable drop in speed at the pause itself vs. ungrammatical pauses, which
supported their hypothesis that grammatical pauses are indeed choreographed by a central cognitive planner. Since it was shown that different speaking styles can affect the articulators, Ramanarayanan et al. (2013) explored read vs. spontaneous speech further. Ramanarayanan et al. (2013) point out that no imaging technique can give a complete view of all vocal tract articulators, which can make the analysis of vocal tract posture difficult. There have been developments in real-time MRI that can examine the midsagittal vocal tract during speech production which provides a way to measure the articulators. Ramanarayanan et al. (2013) used American English speakers reading a simple dialog in a conversation with the experimenter such as “what music do you listen to…,” “tell me more about your favorite cuisine…,” etc.) in order to elicit spontaneous spoken responses while inside an MRI scanner. One result found in Ramanarayanan et al. (2013) was that vocal tract postures occurring in absolute rest positions are more extreme and significantly different from those occurring in ISPs. Specifically, Ramanarayanan et al. (2013) found values of several variables (not velic aperture) during both read and spontaneous ISPs to be significantly higher than those during non-speech rest intervals which suggests a more closed vocal tract position with a smaller jaw angle and a narrow pharynx at absolute rest compared to articulatory settings occurring just before speech (speech-ready) and during speech (ISPs). Ramanarayanan et al. (2013) argue that this may indicate that during the instances of non-speech rest position the tongue may be resting more nestled in the pharynx of the individual and that the mouth is more closed.

Additionally, Ramanarayanan et al. (2013) found that rest positions
displayed relatively high differences compared to speech-ready and ISP positions. The trend was typically seen for the read ISPs. Ramanarayanan et al.'s. (2013) methodology is fairly robust to rotation and translation and does not require much manual intervention while also giving a meaningful comparison across speakers. The results from Ramanarayanan et al. (2013) using real-time MRI measurements of vocal tract posture show that (1) there is a significant difference in default rest postures compared to speech-ready and inter-speech pause postures, (2) there is a significant trend in most cases for variance between ISP pauses, which appear to be more controlled in their execution vs. rest and speech-ready postures, and (3) read and spontaneous speaking styles also exhibit differences in articulatory postures.

1.2. Visual Speech Information From Facial Movements

There have been many works including Ronquest et al. (2010) and Soto-Faraco et al. (2007) investigating perception of visual-only speech also referred to as "lipreading" or "speech reading".

1.2.1 The Contribution of Visual Speech Information

Earlier work on visual-only speech has also shown that speech perception is multimodal and that visual signals can both enhance and alter it. Several prior studies including Soto-Faraco et al. (2007), Munhall & Vatkiotis-Bateson (1998) and Sumby & Pollack (1954), suggest that visual speech information can aid in understanding spoken messages in 1) noisy conditions, 2) second languages and 3) instances where it is conceptually difficult to understand. When listening to speech
in noisy conditions the face can provide extra information that increases perceptual accuracy, Sumby & Pollack (1954).

Munhall & Vatikiotis-Bateson (1998) note that each individual speaker differs in the amount and clarity of phonetic information they provide while different speaking styles can also affect how the visual signal influences judgment. A typical problem throughout much of the past audio-visual studies is a lack of information regarding the visual stimuli other than the gender of the speaker. Soto-Faraco et al. (2007) asked the question of how much information can be revealed from visual-only speech signals and expanded upon prior research showing visual-only signals can relay information to perceivers. Listeners modify the use of visual information depending on the recording conditions (Vatikiotis-Bateson et al. (2007). Ronquest et al. (2010) replicate and expanded further upon Soto-Faraco et al. (2007) using two languages that differ in rhythmic classification and timing in order to examine the contribution of rhythmic information in visual-only language processing. Ronquest et al. (2010) found that both the monolingual and the bilingual observers completed the task successfully which further supported earlier results of Soto-Faraco et al. (2007).

1.2.2 Language Identification From Visual-Only Cues

Ronquest et al. (2010) mention how a significant amount of research has demonstrated that rhythmic information can be identified in auditory speech and that listeners can distinguish between languages in the absence of lexical or segmental information by relying solely on linguistic rhythm and durational cues. It
has also been shown that in terms of rhythm, stress-timed languages such as English have a higher variation in vowel duration than syllable-timed languages such as Spanish which has considerably less vowel duration. One experiment in Ronquest et al. (2010) tested whether observers could differentiate rhythmic differences of syllable-timed vs. stress-timed stimuli by examining visual-only cues in language identification tasks. The experiment used monolingual and bilingual Spanish–English participants using a two-alternative forced choice (2AFC) task. The stimuli used in this experiment consisted of visual-only video clips of English and Spanish sentences spoken by male and female bilingual speakers. They also sought to examine additional cues available to observers for language identification. Another experiment in Ronquest et al. (2010) focused on the use of rhythmic cues in language identification which explored rhythmic differences and contributions to the perception of visual-only signals. Their results show that language identification can occur from visual signals alone and that observers are able to identify some lexical items from a visual-only display, but that the amount of available lexical information in this modality is very limited. Some specific words such as common lexical items and phrases were identified more accurately vs. less common ones, but overall the percentage of correct word identification was low. Their results also showed that observers were able to identify languages based on their rhythmic differences in visual-only stimuli.

Additionally, observers were able to identify stimuli that were temporally reversed which Ronquest et al. (2010) argue, eliminated lexical information but retained rhythmic differences, however this is debatable and certain scholars claim
that simple reversal of syllables can shift timing attributes. All of the participants in Ronquest et al. (2010) performed significantly above chance in language identification tasks both in forward and reversed conditions despite language background or prior linguistic experience. Since observers were able to identify words in the backward condition, this supports the idea that rhythmic differences are a cue that aids in language identification and that vowel duration and rhythmic differences among languages can affect how languages are perceived and identified in visual-only speech. Results from Ronquest et al. (2010) also support the idea that the visual signal by itself is sufficient for an observer to correctly identify the language being spoken, expanding upon previous research confirming that prior linguistic experience, lexical information, rhythmic structure, and utterance length can play a role in visual-only language identification.

1.2.3 The Bilingual Advantage in Language Identification of Visual Stimuli

Soto-Faraco et al. (2007) did a similar study to Ronquest et al. (2010) using Spanish and Catalan which are more similar to each other than Spanish and English. Soto-Faraco et al. (2007) suggested future studies should examine observers’ ability to discriminate or identify languages that are less closely related than Spanish and Catalan. Soto-Faraco et al. (2007) examined whether monolingual and bilingual observers could discriminate Spanish from Catalan using visual-only speech stimuli. Soto-Faraco et al. (2007) used two groups of bilinguals (Spanish dominant, Catalan dominant) and three groups of monolinguals (Spanish, Italian, English) that participated in language identification tasks. They found that the bilingual
observers discriminated the languages better than the monolingual Spanish
observers who still performed above chance and that the English and Italian
monolingual observers who had no experience with either language were not
successful at the task. This implies that knowledge of at least one of the languages
is necessary in order to accurately discriminate visual-only stimuli.

Soto-Faraco et al. (2007) concluded that prior experience with the specific
languages or at least one of them is one primary factor aiding in successful
discrimination. Soto-Faraco et al. (2007) also mentioned that several different
features of the stimuli affected discrimination, including length of the utterance
and the number of distinctive segments or words present in the stimuli.

1.2.4 Increased Cognitive Load of Visual-Only Speech

de los Reyes Rodríguez Ortiz (2008) mentions that processing visual-only speech
signals has a higher demand mentally as opposed to perceiving oral speech signals.
Speech reading requires a certain level of skill in deduction because one must have
the ability to finish what one cannot hear from oral information. Another
consideration discussed in de los Reyes Rodríguez Ortiz (2008) is the correlation
between level of intelligence and speech reading accuracy that has raised certain
questions by scholars. de los Reyes Rodríguez Ortiz (2008) poses the example
argued that when a person has an IQ below 80 they will have certain difficulties in
processing visual-only speech, however, this issue is still under heavy debate. de los
Reyes Rodríguez Ortiz (2008) explains how memory is also considered to be related
to accurately interpreting visual-only speech since high accuracy levels of speech reading generally occur when the perceiver has a high level of working memory, and show that among prelingually deaf people, the best speech readers were those that possessed higher levels of intelligence and more intelligible speech. While there may be a correlation between a participant’s IQ and their performance in laboratory conditions, interpretation of the judgments are hard to measure.

1.3. Linguistic Information From Facial Movements & Facial Recognition

Campbell & Massaro (1997) investigate how a speaker’s face shows linguistic information in face-to-face interactions. In their study it is mentioned how prior studies have shown that participants with normal hearing are able to speech read without regular training and that both normal hearing and hearing-impaired individuals can be trained to recognize visible consonant and vowel phonemes or ‘visemes’. Campbell & Massaro (1997) ask which features actually convey the information in the face required for speech reading. Studies such as Munhall & Vatikiotis-Bateson (1998), Munhall et al. (2004) and Vatikiotis-Bateson et al. (2007) have shown the features in the lower half of the face including the jaws, lips and cheeks are what conveys information to process speech reading. Campbell & Massaro (1997) mention prior studies such as Summerfield (1979) where subjects had to read speech in conditions: 1) whole face displayed, 2) videos where only the lips were shown and 3) a moving ring representation of the lips. Campbell & Massaro (1997) explain the results found in Summerfield (1979) that identification accuracy increased as much as 42.6% in videos where the whole face was displayed.
vs. videos where only the lips or the moving ring were shown. Campbell & Massaro (1997) describe how faces are thought to be perceived as both individual features and structural relations among features. Second order structural relations are described as those that remain constant in all stimuli where first order ones that do not. For example, they regard the nose, mouth and eyes as second-order, but the jaw as first-order. Campbell & Massaro (1997) describe how different structural order relations have been thought to be visually identified independently of one another and that prior scholars have hypothesized that facial recognition is processed similarly to visual speech recognition. However, it is difficult not to point out the problems mentioned in Campbell & Massaro (1997) particularly in regards to structural relations because the mouth, nose and eyes can move and there are muscle components for these body parts as there are for body parts classified as first order structural relations (Vatikiotis-Bateson et al. 2007).

Conrey & Gold (2006) discuss how normally hearing perceivers generally are able to understand visual-only speech or ‘lipreading/speech reading’ but speakers vary in how easy they are to understand. Despite the speaker variability of information given during visual-only speech there are strategies used by participants that also have an affect on what information they can receive. The accuracy of visual-only speech perception, also known as “speech intelligibility”, has been shown to vary across different speakers and some speakers are consistently easier than others to speech read.

Conrey & Gold (2006) note that more studies in variability of auditory speech than visual-only speech and how it has been shown that auditory speaker variability
can lower speech intelligibility. One such example given by Conrey & Gold (2006) is when a list of words is read by different speakers which contributes lower memory recognition in an instance when a previously heard word is presented in a novel voice as opposed to that of the original speaker. Even though both speakers and observers can vary on how intelligible and accurate their speech and perception is, it has been found that visual intelligibility of speakers tends to remain consistent across different observers. The question that arises is whether it is the properties of the speaker or the overall perceptual strategies across the observers. An example of this is given in Conrey & Gold (2006), speaker A who only moves his/her lips is known to be less intelligible than speaker B who moves his/her lips similar to speaker A but also uses additional jaw movements. It is not known if observers are getting more information from speaker B’s jaw movements or if the observers are using some other cues from speaker B that give a higher accuracy level. It is possible that the observer only looks at jaw movements and as a result sees speaker B as more clear since speaker A has no jaw movements, but if observers looked at the lip movement maybe they would be able to read both speakers at an equal accuracy level. One possible explanation could be that it is due to an individual speaker’s properties or speech habits that have an effect on an observer’s perceptual strategy. One example of this is given in Lansing and McConkie (2003) who reported observers focusing their gaze more on a speaker’s mouth when visual-only speech was presented vs. visual and auditory speech presented together. If this is true then this would support the argument that it is the speakers’ properties that ultimately affect perception and not the perceptual strategies of the observers.
Munhall et al. (2004) demonstrated how head movement plays a role in speech intelligibility employing the use of a custom animation system to create four different audiovisual versions of 20 Japanese sentences: 1) recorded natural head motion & recorded facial motion; 2) zero head motion & recorded facial motion; 3) double head motion (amplitude of head movement doubled in all six degrees of freedom) & recorded facial motion; and 4) auditory-only video with the screen blacked out. Best performance was achieved when participants identified sentences with natural head & facial motion continuing from best to worst in the following order: zero head motion & recorded facial motion, double head motion & recorded facial motion, auditory-only. This makes sense that the auditory-only condition displayed the worse performance based on many previous investigations showing that by removing one of the signals decreases intelligibility, but the head movement factor is quite interesting here and should be noted for future investigations relating to this current study.

Conrey & Gold (2006) discuss how measurements of lip opening and vowel duration have been found to be generally good judges of perceptual cues for identifying vowels among some speakers, but this also varies across speakers. Furthermore, perceptual distance from consonant phonemes has been measured by using multidimensional scaling (MDS) analysis and has shown that speakers who were more intelligible had greater correlations between distance from phonemes on the MDS analysis. However, one problem in Conrey & Gold (2006) is that participants looked at visual stimuli that included the markers used on the talkers so this may have had an effect on observers’ perception. Separating the two factors
of physical variability in the visual information and the perceptual strategies among observers described in Conrey & Gold (2006) can be done with a technique called ideal observer analysis, Geisler (2004). It can be used to give the amount of physical information available in a perception task by looking at observers who produce the best possible performance on a certain perceptual task. The goal of Conrey & Gold’s (2006) study was to test whether cross-speaker variability in visual-only speech perception happens because of 1) differences in information available across talkers or 2) different perceptual strategies among observers. Conrey & Gold (2006) established that talker variability in visual-only speech perception happens because of both variability of physical information and the perceptual strategies of observers.

1.4 Goal & Questions of This Study

I hope in this thesis to answer the question of whether utterances occurring in what is defined as ISP and speech-ready position can be identified visually using similar methods as those described in Weikum et al. (2013) and Soto-Faraco et al. (2007), namely whether participants can discriminate between different languages from silent video clips. Are we able to discriminate between two different languages simply by looking at the visible articulatory positioning of one’s face when the speaker is in ISP and/or speech-ready position? Are perceivers able to better identify ISP and speech-ready tokens in the language(s) they speak vs. the language(s) they do not? For instance, would a native English speaker watching a
silent video of a person speaking be able to better judge whether the speaker in the video is speaking English vs. a different language unfamiliar to the observer?

Based on Weikum et al’s (2013) results, if it is assumed ISP and speech-ready tokens are perceived in a similar way as full sentences it is expected that observers are able to differentiate two different languages if they have familiarity with at least one of them? If ISP and speech-ready position are language specific, this postural information should be more available to perceivers who are familiar with the target language(s) in question. A pilot study was conducted in order to test the hypothesis that perception of ISP is more robust when perceivers are familiar with the target language in Chapter 2 followed by a more elaborate study of the same question in Chapter 3.
Chapter 2

Native/Non-Native Perceivers Pilot

An initial pilot study was conducted wherein participants with or without previous knowledge of English and French observed short, silent video clips of a bilingual French-English speaker’s face during pre-speech and inter-speech postures.

2.1. Methods

6 participants from two different language groups participated in this study (3 native English speakers with non-native exposure to French, and 3 non-native English speakers with no previous knowledge of French). All 3 of the native English speakers were female between the ages of 26-38 and from the US or Canada, while the non-native English speakers were 2 males and 1 female between the ages of 24-43. The female and one of the males were native Japanese speakers from Japan and the other male, a native Slovenian speaker from Slovenia. All non-native English speakers started learning English as an L2 language between 9-12 years of age and did not have any knowledge of French. None of the observers in either group had any reported speech or hearing impairments. Below is a more detailed background on each of the individual speakers:

Native English Speaker 1:
Female from Portland, Oregon USA.

Native English Speaker 2:
Female from Dallas, Texas USA.
Native English Speaker 3:
Female from Vancouver, British Columbia Canada. Had some exposure to French while living in Montreal, Quebec while studying at a university. Her courses were taught entirely in English.

Non-native English Speaker 1:
Female from Japan that learned English from a Japanese tutor in grade 6 approximately once a week. Tutoring during this time focused only on writing and reading. She systematically started to learn English from a Japanese teacher from grade 7 in public school in Japan. She did not speak with native speakers of English until the age of 18.

Non-native English Speaker 2:
Male from Japan started learning English at around 9-10 years of age.

Non-native English Speaker 3:
Male from Slovenia started learning English at the age of 12.

The stimuli presented were produced from recordings of one balanced bilingual Canadian French/English speaker recorded on video while having a casual conversation with another balanced bilingual Canadian French/English speaker who was off camera. While both speakers were Canadian and spoke English and French, they were from different provinces (Alberta and Quebec) and spoke different varieties of “Laurentian French” (also known as “Québécois French”). One speaker was from Western Canada (Alberta) who spoke “Western Canadian French”; the other from Montreal speaking “Montreal Canadian French”, and also had a Swiss French substrate. Due to having only one camera it was decided to have the one speaker off camera because recording both speakers sitting facing each other with one camera would introduce difficulty in getting a straight angle on both speaker’s faces. The speakers engaged in two 10-minute conversations, one in English and the other in French. A total of 80 short clips, 40 from each language
conversation were extracted using Final Cut Pro on a Mac computer with a total of two conditions, 1) when the speaker was in the act of an ISP (x20) and 2) when the speaker was in speech-ready position (x20). The methods and selection criteria for determining ISP and speech-ready tokens were as follows:

**ISP:**

If the speaker had already started a speaking a phrase, an extraction was made anywhere within that phrase where there was no auditory speech. For example, if the speaker said a sentence such as, “It was a hot day yesterday” an extraction could be made during the timeframe between the words ‘it’ and ‘was’ or ‘day’ and ‘yesterday’ etc.

**Speech-Ready:**

If the speaker was not in the act of a phrase. This could consist of moments when the speaker was just listening to the other speaker ranging all the way up to when the speaker starts to move articulators but before audible speech is made. Some examples included when the speaker was sitting and nodding his head while listening to the other speaker, or when the speaker was about to utter the sentence “Where did you say you were from?” in which the extraction could be made starting from anytime while the speaker is merely in the act of listening up to the point where the speaker opens his mouth to articulate /w/, but before the onset of the vowel. There were no control factors for specifying the length of the stimuli or body and head movements included within those stimuli extractions.

The audio was removed from all of the clips during extraction. Clips ranged from approximately 100 milliseconds to 3 seconds in length. Participants were
tested in a sound-controlled room looking at stimuli via MS PowerPoint. They watched all 80 tokens for both languages in each of the conditions. The tokens were arranged in a randomized order using an online list randomization tool and observers had to judge which language the token was. Observers were not told that they were viewing clips of speakers in ISP or speech ready position. The administrator of the experiment sat next to the observer inside the sound booth the whole time and navigated to the next token when the subject was ready. This was done in order to make sure observers didn’t accidentally skip tokens. Observers stated their answer as French or English and the experimenter wrote down the answer. Observers were allowed to request replay of each token as many times as they wanted to before they made a judgment. At the end of the experiment each participant was given an explanation of the point of the research. All of the participants were asked questions about their performance such as “Did you find the discrimination task easy or hard?” etc. Several of the participants explained some of they tactics they employed in discriminating between the two languages. The most common tactic described was looking at the opening of the mouth and identifying a token as French if the mouth opening was small or rounded while identifying it as English if the mouth was open wide.

2.2. Results

The results were mixed across all observers for all tokens. In this pilot study, there were too few participants to run statistics, however results strongly suggest that tokens were identified based on chance. Figure 2.1 below illustrates the correctly identified tokens overall and for each language among the different observers.
Looking at the figure more closely we see that the first column represents the total number of tokens judged correctly with a total possible number of 80. English speaker 1 identified 45 and English speakers 2 and 3 each identified 37. Looking at the non-native English speakers, we see similar results. Non-native English speaker 1 identified 46 of the overall 80 tokens correctly while non-native English speaker 2 identified 41 and non-native English speaker 3 got 39 correct. Figure 2.2 shows the same numbers for participants broken down by percentage. The native English speakers had a combined accuracy rating ranging from 46.25% to 56.25% for all tokens whereas NN English speakers had a combined accuracy rating from 48.75% to 57.5% for all tokens.

Figure 2.1-Correct Tokens Overall for English & French Observers
Based on the data from the pilot study it does not appear that the participants’ ability to differentiate between ISP and speech ready positions across English and French stimuli has much to do with their age of acquisition, but this factor was not tested systematically in this study. The data here show that ISP and speech ready tokens do not yield similar results as the full sentences in Weikum et al. (2013). If ISP and speech-ready tokens in this study had been perceived the same way as the full sentence and word tokens in Weikum et al. (2013), it would have been expected to see the non-native English speakers do worse on the language identification tasks here. However, all participants across both groups appeared to have close accuracy ratings and both groups also seemed to have identified tokens at chance. In fact, the non-native speakers as a whole did slightly better at identifying tokens as a whole vs. native English speakers, although there was
variation in individual speakers for both groups in terms of who did better at identifying English tokens vs. French tokens.

It was decided to look at other factors more closely in order to determine what other visual information an observer uses to make a judgment. In order to make a qualitative evaluation, tokens that were identified correctly amongst all or most observers were analyzed further in an attempt to better understand what cues are used in order to make a judgment. Of the 80 tokens only one was correctly identified for all six observers, but there were 8 additional tokens correctly identified by five of the six observers and these nine tokens were given closer inspection. The token identified by all six observers showed the speaker’s mouth open very wide which could be a likely cause why it was correctly identified as English since most of the participants described how they thought tokens displaying a wider mouth opening were English. Also, observers tend to perceive French speech as having more lip rounding and less exaggerated or wide jaw openings and movement. Other observations including head nods, eyebrow movement and tight-lip closures may also be giving information as to how to judge the target language.

A further explanation of the differences between ISP and speech-ready stimuli as they are defined in this thesis needs to be addressed. These two types of stimuli are quite different from each other. Movements occurring in ISP are dynamic transitions between words, whereas speech-ready position may involve little to no movement depending on the individual speaker. The ISP tokens contained more coarticulatory segmental speech content and were on average much shorter in duration than speech-ready tokens. ISP tokens were typically less than half a second
in duration, while speech-ready tokens were as long as three seconds. Because of these differences, it is possible that participants may have had more or less difficulty in identifying one condition compared to the other. Future studies should fully compare the differences in the perception of ISP and speech-ready stimuli across each language.

Descriptive comments on the nine most frequently correctly identified tokens are shown in Table 2.1 below:

<table>
<thead>
<tr>
<th>Identified correctly by</th>
<th>Token type</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>All participants</td>
<td>English ISP 17</td>
<td>Mouth open very wide</td>
</tr>
<tr>
<td>5/6 participants</td>
<td>English ISP 02</td>
<td>No lip rounding, little to no protrusion</td>
</tr>
<tr>
<td>5/6 participants</td>
<td>English ISP 05</td>
<td>No lip rounding, little to no protrusion</td>
</tr>
<tr>
<td>5/6 participants</td>
<td>English ISP 08</td>
<td>No lip rounding, little to no protrusion, head nodding</td>
</tr>
<tr>
<td>5/6 participants</td>
<td>English Speech Ready 06</td>
<td>Significant lower jaw movement, eyebrow raising</td>
</tr>
<tr>
<td>5/6 participants</td>
<td>English Speech Ready 07</td>
<td>Head nodding, lips pressed together tight</td>
</tr>
<tr>
<td>5/6 participants</td>
<td>English Speech Ready 13</td>
<td>Head nodding, lips pressed tightly together, some eyebrow movement</td>
</tr>
<tr>
<td>5/6 participants</td>
<td>French Speech Ready 02</td>
<td>Mouth open w/ slight rounding</td>
</tr>
<tr>
<td>5/6 participants</td>
<td>French Speech Ready 12</td>
<td>No actual lip movement, but position of lips slightly open, head moves sideways</td>
</tr>
</tbody>
</table>

Table 2.1 - Tokens Identified by at Least 5 Out of 6 Participants
2.3. Discussion

The goal of this study was to expand upon data from Weikum et al (2013) with ISP and speech ready stimuli in order to determine if age of acquisition of a language contributes to adults' ability to distinguish different languages in visual-only speech. Based on the results it appears that age of acquisition alone is not likely to determine this ability. Observers need other informational cues from the visual stimuli in order to make a correct judgment, but the question remains what those cues are exactly. The tokens identified correctly by all or most speakers showed various properties in the visual signal that included head nodding & movement, eyebrow raising & movement, lip closures & openings and lip protrusion. These factors may affect how an observer perceives silent ISP and speech-ready stimuli. There may also be additional information that was not seen or mentioned here that play a role in this task. It is not known whether these additional factors will better help native English speakers or non-native English speakers for this study. Based on the information covered so far it appears that all observers are no better or worse at using visual cues to identify tokens no matter what language they speak or when they learned it.

At the end of the experiment observers were asked if they found certain tokens harder or easier to judge and every participant across both groups stated the task was highly difficult and that no tokens were easier than others. Participants did not know during the experiment that they were judging inter-speech and ISP tokens, they only knew they were looking at either French or English, but observers could take note of duration of the token. All of the tokens were very short in duration.
compared to full sentence and word tokens used in Weikum et al. (2013) but some of the tokens in this study were still significantly shorter than others.

Looking at table 2.1 more closely shows that the tokens English ISP 01, English ISP 05 and English ISP 08 had no lip rounding with little to no protrusion. These properties may have contributed to participants perceiving these tokens as English. We know that 5 of the 6 participants correctly guessed these tokens, but what if tokens displaying similar properties had been extracted from the French conversation? The same can be said of English ISP 17 token which displays a wide opening of the mouth, and which was identified correctly by all 6 of the participants. Is this because observers associate wide mouth opening with English? To what extent are observers using these visual properties to make a decision? While the properties of specific tokens are noted, they were not analyzed in detail which is something that should be expanded upon in future related studies. Based on the results it does not appear that participants were able to better identify speech-ready tokens from inter-speech tokens regardless of what their native language was, but we do not have concrete evidence for this and more work will need to be done in this regard.
Chapter 3

Native English vs. Native French Perceivers

In order to better understand the results of the study presented in Chapter 2, another study was conducted with more perceivers, and with both native French-speaking and native English-speaking perceivers. The participants in this experiment consisted of both native English speakers with L2 knowledge of French and native French speakers with L2 knowledge of English. Weikum et al. (2013) did not use any French speakers in their study and found that observers only needed to have a degree of proficiency in one of the two languages to discriminate in a same/different task between English and French sentences. For this second experiment it was not expected that native French speakers would discriminate better overall than native English speakers or vice versa. The main hypothesis for this second experiment is that native speakers are better able to perceive interspeech and pre-speech postures in their first language given only visual information. An additional question for the second experiment was whether or not English speakers and/or French speakers would identify the same tokens correctly as the majority of the participants in the first experiment (shown in Table 2.1). Also would French speakers be able to better discriminate certain tokens better than English speakers and vice versa?
3.1. Methods

The total number of observers included 7 native English speakers and 7 native French speakers. The native English speakers were all females between the ages of 21-28 from western Canada (6 were from BC and 1 was from Alberta). The native French speakers were 4 males and 3 females with 6 of the speakers being from France and 1 female speaker being from Quebec, Canada. The French speakers were 20-34 (5 of the speakers were exactly 20 years of age) years old. Below is a more detailed background for each speaker:

**English Speaker 1:**
21 year old female born in Vancouver. Started using English from birth, and studied French from grades 7-12 (age 13-18). Highest level of education in French is grade 12 level. There was a gap of not speaking or reading French from age 18-20. Currently only uses French occasionally for work as of recent (from Jan 2014) for tutoring purposes.

**English Speaker 2:**
23 year old female born in Vancouver. Started using English from birth and is used predominantly. She studied French from elementary to high school from age 9 to 18 focusing mainly on grammar and vocabulary, but not spoken French. Her French schooling was not an immersive environment. She also studied in French in university from age 20-21 at an intermediate level with more focus on spoken French. Currently she rarely uses French, speaking it only once in a while.

**English Speaker 3:**
25 year old female born in Surrey, BC. Started learning French around grade 4 or 5 and took French courses up until grade 11. Returned to studying French with courses during first year of University in 2006/2007 in Detroit where French courses were equivalent to high school level. Currently only uses English.

**English Speaker 4:**
24 year old female born in New Westminster, BC. She studied French from grades 6-8 when she was between 11-13 years old and only used in school for French class.

**English Speaker 5:**
23 year old female born and raised in Quesnel, B.C. She started using French at the age of 5 in a French Immersion school for 13 years. She had one gap year afterwards
and continued to take French courses at university for 3 more years. She took 100/200 university level courses with no gaps. At university she only used French within her courses. In 2011, she studied French for 4 months in Nantes, France, taking intermediate level courses. She used French on a daily basis for the majority of her stay in all settings. Since then she has not taken any further French courses.

**English Speaker 6:**
28 year old female born in Victoria, BC. English was used for all of her education and she uses English everyday at school, work and home.

**English Speaker 7:**
21 year old female born in Edmonton, Alberta. She attended school and university taught exclusively in English. She’s spoken English since birth and has very little experience with French, having taken one or two beginner classes around the age of 12. Her French use was limited to the classroom and she has not used it since. English is the only language she’s ever used on a regular basis.

**French Speaker 1:**
20 year old female born in Brittany, France. She has used the language daily even in Canada having always been surrounded by French people. She started learning English at school at 8 years of age and had only a few hours of English lessons per week until the end of middle school. She took more advanced classes in English literature and history in high school (approximately 8 hours per week). Her classes at university were exclusively taught in English, mostly by native speakers, and she used this language to write academic papers, oral presentations and to converse with international classmates. She was considered as fluent in English according to her French university.

**French Speaker 2:**
20 year old male born in Niort (west of France, near Atlantic ocean). Started learning basic English at the age of 8 at school (numbers, alphabet, introductions...etc.) with more comprehensive instruction beginning around middle school (age 11). Throughout middle school and high school years, teaching of English was not outstanding. He had sufficient English grammatical structure and vocabulary and became interested in improving on his own time. His pronunciation improved from watching English movies and TV shows. When he entered university in Paris, he began meeting exchange students from Britain while continuing English studies where all his teachers were native English speakers (from Canada, Ireland and USA) which was not the case in grade school. Since coming to UBC, he uses French every day with fellow exchange friends but has no problem attending a class or having conversations with foreign friends in English.
French Speaker 3:
20 year old female born in Orsay, France. Started learning English in primary school around 8-9 years old. She has never stopped using French, but English was part of her education as her parents spoke English sometimes for their jobs though she never had conversations in English in their daily lives. Since she has had the chance to live and study at UBC a whole academic year her English has improved. In high school and at her home university in France, she had English classes for 3 or 4 hours maximum per week.

French Speaker 4:
20 year old male born in Saint-Cloud, France. He started to learn English at school when he was 9 and has been speaking English continuously from that time. His language of instruction from 3 to 18 years old was French. During his second year at university in France he had courses in French and English. During his year at UBC his instruction has been in English only. He uses French everyday with friends and by reading website pages. He also uses English everyday watching series, reading internet sites and speaking with foreign friends.

French Speaker 5:
34 year old female born in Québec Canada. She attended French grade schools with 1 class of English per week taught in elementary and high school. She started using English on a regular basis at the age of 16. Since the age of 20 she has used written and spoken English at work and reads papers in English at university.

French Speaker 6:
20 year old male born in Nimes, France. He started studying English at the age of 12 but rarely practiced, using it only 2 hours per week in class. In 2011 he started studying English more intensely. In August 2013 he came to Vancouver and started using English on a daily basis. Recently he received the maximum level on his English exam through his French university that is equivalent to the European C2 level.

French Speaker 7:
Male from France. In 2013, spent the summer in Vancouver with everyday exposure to English. He used both French and English on a daily basis.

The procedure introduced some changes from the pilot study described in Chapter 2. The stimuli presented were not completely identical to the stimuli presented in the experiment in Chapter 2 due to those changes in the presentation format which is further discussed below.
While the signal content of each stimulus was the same as before, experiment 1 used MS PowerPoint slides to present to observers while the experimenter sat next to the subject and the experimenter was the one navigating through the stimuli while also writing down the observers’ responses. For this second experiment, the video stimuli were presented through the PsyScope (Cohen et al., 1993) application which also recorded all participants responses along with their response times while also eliminating the need for the experimenter to be in the room with the participant. At the beginning of each session, participants went through a training session consisting of 20 trials in a randomized order. There were 3 different tokens in the training session. At the beginning of each trial token the stimuli would appear with the video paused at the very beginning. The participant would then be required to press the spacebar for the video to play. Once the video was finished the screen would go blank, at which point the observer was required to enter ‘1’ for English or ‘9’ for French before they were able to proceed to the next trial token. At the end of the training session there was an instruction screen that indicated that the actual experiment was starting which then used the same 80 tokens that were used in the experiment described in Chapter 2.

All of the tokens were randomized through PsyScope. The participants had full control as to when to start the play of each token. All participants viewed the stimuli in a sound booth or a quiet, empty room.

3.2. Results and Discussion

Statistics for accuracy were calculated using R, which took proportion correct (PC) and total correct (TC) by speaker for each individual token. For example the token
English ISP1A had a PC value of 0.571 and TC value of 4 for the English group which means 4 of the 7 English speakers identified this token correctly with a 57% accuracy rate. Overall there did not appear to be a significant difference between the accuracy rate of the English and French speakers across overall tokens. Overall English speakers accuracy rate across all tokens was near 50% whereas overall French speakers accuracy rate across all tokens was slightly above 50% shown in table 3.1 below.

When language of the stimulus is not a factor, speakers perform equally on the task (mean English group accuracy = 49.11%, mean French group accuracy = 52.50%, t (1118) = -1.14, p = .257). This is to show that the groups did similar in task performance. When speakers’ accuracy on their native language was compared to speakers’ accuracy on their L2 language, irrespective of what their native language was a significant difference was observed, (mean L1 accuracy = 54.29%, mean L2 accuracy = 47.32%, t (1118) = 2.33, p = .020). However, when you look at the results based on the first language of the participants, the significant difference in accuracy only emerges for native English speakers and not the native French speakers. Two-way T-Tests comparing participants divided into two groups (English L1 speakers, French L1 speakers) were run to compare their performance on English vs. French tokens using R, and the values are shown for each of the groups in Table 3.1 below:
The mean English and French resp. is the accuracy rate for the English and French tokens. Despite an insignificant difference among the English and French group across tokens overall, it was shown that both groups had a higher accuracy rating when judging tokens in their respective languages. The English group showed around a 10% higher accuracy rate in the English tokens while the French group appeared to be more balanced but still showed higher accuracy when judging French tokens. These results are illustrated in figure 3.1 below.

<table>
<thead>
<tr>
<th><strong>English group T-test results</strong></th>
<th><strong>French group T-test results</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>t= -2.46</td>
<td>t= 0.84</td>
</tr>
<tr>
<td>df= 558</td>
<td>df= 558</td>
</tr>
<tr>
<td>p-value = 0.014</td>
<td>p-value = 0.399</td>
</tr>
<tr>
<td>mean English resp= 55.00%</td>
<td>mean English resp= 50.00%</td>
</tr>
<tr>
<td>mean French resp= 44.64%</td>
<td>mean French resp= 53.57%</td>
</tr>
</tbody>
</table>

Table 3.1-T-Test Results for English and French Groups

![Figure 3.1 - Accuracy % by Native Language](image-url)
The results from the second experiment in this chapter lend support to the hypothesis posed in Chapter 1 that observers are able to better identify ISP and speech-ready postures visually in the languages they speak natively vs. the ones they do not. This shows that language experience and background contributes to the ability to use these visual signals that have no segmental background. The results also allow us to address another question that was posed in the introduction to this chapter: whether observers who are native French speakers are able to discriminate certain tokens better than observers who are native English speakers and vice versa?

It is important to note again the differences between ISP and speech-ready because of the different information available in the two conditions. It was decided to investigate the ISP and speech-ready stimuli more closely between the two groups. A Confusion Matrix was created in R to illustrate how French and English participants responded to the different stimuli in each of the conditions (ISP &
speech-ready) across each language shown in figure 3.2 below.

Looking at the matrix (lighter color shades represent a higher number of responses while darker color shades show a lower number of responses), it shows that the French subjects responded ‘French’ and ‘English’ equally for all ISP stimuli (left column). This suggests that French participants were responding or guessing roughly evenly for all ISP stimuli regardless of the language. The English subjects responded ‘French’ more when the ISP stimuli were French and responded ‘French’ less when the ISP stimuli were English and vice versa for English. So the English participants seemed to do better at identifying ISP stimuli than the French participants who appeared to be guessing across all ISP stimuli. For speech-ready
stimuli, French subjects appeared to pick ‘English’ less when the speech-ready stimuli were French while picking ‘French’ more when the speech ready stimuli were French so they did well on identifying French stimuli in the speech-ready condition. However, French subjects seemed to pick ‘English’ and ‘French’ equally when the speech-ready tokens were in fact English. The English subjects appeared to be more biased toward picking English in the speech-ready condition overall because it shows they picked ‘English’ more whether the stimulus language was English or French, and the English participants picked ‘French’ regardless whether the speech-ready stimuli was English or French. D-prime scores were calculated in R and the results were: ISP for English L1 subjects (0.01790544), ISP for French L1 subjects (0.12566135), Speech-ready for English L1 subjects (-0.03697919), Speech-ready for French L1 subjects (0.05435102).

In addition, an interaction plot (shown in figure 3.3 below) was created showing the mean accuracy rating for speech-ready and ISP tokens across both language groups. Looking at the interaction plot, it appears that there was not a significant difference for either language group in the ISP condition. The English group identified ISP tokens whether they were English or French with about 50% accuracy. The French group identified ISP tokens slightly better when the stimuli were English (55%) than when they were French (50%), but this did not appear significant. The significance seems only to appear for the speech-ready tokens and shows the English group is around 20% better on English speech ready stimuli than French ones. The French group also had a higher accuracy in identifying French
speech-ready stimuli than English ones though the gap was not as wide for them as it was for the English group.

![Figure 3.3-Mean Accuracy Rate by Stimulus Language](image)

Two-way and three-way ANOVAs were calculated in order to test the importance of stimulus language, subject’s L1 language, and stimuli condition (ISP, speech-ready). The two-way variance of stimulus language and subject’s L1 language was found to be significant, while the three-way variance of stimulus language, subject’s L1 language, and stimuli condition was found to be highly significant.

- Stimulus language: Subject L1 = Df(1), Sum Sq (1.36), Mean Sq (1.358), F Value (5.476), Pr(>F) 0.01945
- Stimulus language: Subject L1: stimuli condition = Df(1), Sum Sq (2.32), Mean Sq (2.3223), F Value (9.365), Pr(>F) 0.00227

Qualitative properties of tokens that were most correctly and least correctly identified in experiment 2 were also noted just as they were in experiment 1. Table 3.2 below shows a list of those tokens either correctly or incorrectly identified by all members in each group.

<table>
<thead>
<tr>
<th>Group</th>
<th>Token</th>
<th>Correct/Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>English speech ready 4A</td>
<td>Correct</td>
</tr>
<tr>
<td>English</td>
<td>French ISP 5A</td>
<td>Correct</td>
</tr>
<tr>
<td>English</td>
<td>English speech ready 8A</td>
<td>Incorrect</td>
</tr>
<tr>
<td>English</td>
<td>French speech ready 13A</td>
<td>Incorrect</td>
</tr>
<tr>
<td>French</td>
<td>English speech ready 9A</td>
<td>Correct</td>
</tr>
<tr>
<td>French</td>
<td>English speech ready 15A</td>
<td>Correct</td>
</tr>
<tr>
<td>French</td>
<td>English ISP 8A</td>
<td>Correct</td>
</tr>
<tr>
<td>French</td>
<td>English ISP 14A</td>
<td>Correct</td>
</tr>
<tr>
<td>French</td>
<td>French speech ready 11A</td>
<td>Correct</td>
</tr>
<tr>
<td>French</td>
<td>English ISP 6A</td>
<td>Incorrect</td>
</tr>
<tr>
<td>French</td>
<td>English ISP 12A</td>
<td>Incorrect</td>
</tr>
</tbody>
</table>

Table 3.2 - Tokens Correctly or Incorrectly Identified by All Members of Each Language Group
None of the tokens were correctly or incorrectly identified across all speakers in both groups. Inspecting the descriptive data, there were 2 tokens that were always identified correctly by all participants in the English group (English speech ready 4A, French ISP 5A) and identified 2 tokens that were always identified incorrectly (English speech ready 8A, French speech ready 13A). Chapter 2 discussed individual stimuli tokens that were correctly/incorrectly identified by all or most participants in experiment 1; experiment 2 also shows certain stimuli tokens that were perceived as correct or incorrect among participants in the two language groups. Looking at table 3.2 we see that the all the participants in the English group correctly identified English speech ready 4A token, while they all incorrectly identified English speech ready 8A token. English speech ready 4A showed the head tilted with the mouth open, so it is possible that the participants in the English group associate these properties with the English language which may have contributed to their perfect accuracy rating on this particular token. On the other hand, English speech ready 8A token which displayed some lip rounding, was incorrectly identified among all English participants perhaps due to lip rounding being thought of as having French-like properties. French speech ready 13A token was also incorrectly identified amongst all English participants, this token showed head nodding and lip tightening and perhaps these properties are generally perceived as features of English to the English participants. It is difficult to say what exactly is contributing to each of the individual participant’s perception across both groups, but it would be valuable to consider the descriptive properties of these individual tokens in future work.
Chapter 4

Conclusion

The data shown from the two experiments suggest that an observer’s native language does affect his or her ability to discriminate visually across languages in ISP and speech ready tokens. Weikum et al.’s. (2013) data already showed that it is possible to distinguish languages visually when using stimuli containing segmental speech information in the form of full sentences and words. Weikum et al. (2013) also explained how the length of stimulus can play a factor in how easy or difficult it is for observers to make a judgment. Their data support the hypothesis that the longer the stimuli the more information observers can use to make an effective judgment. Prior to the present thesis, stimuli such as ISP and speech-ready tokens used in the present experiment had not been tested in a study of visual-only perception. Despite the tokens in the present thesis being much shorter in length than those used by Weikum et al. (2013) and largely lacking segmental speech information, the present results nevertheless indicate that native speakers are able to visually identify speech-ready postures of their native language. It is important to note that these speech-ready postures were longer in duration but contained less segmental coarticulatory information than the inter-speech postures, suggesting that perceivers are using non-segmental visual information about facial posture to distinguish between English and French.

There remain several things to explore here. One possible step is to analyze more closely the tokens that were identified all correctly or incorrectly by the
English or French groups as shown in table 3.2 to identify qualitatively the features perceivers are using (either correctly or incorrectly) to identify their native language. Based on observations in the present study, these features are likely to involve more than just articulator positioning, such that gestures including head movements, nodding, eye movement, etc. will have an effect on how different observers respond. Future studies should focus on extralinguistic information such as this in order to better understand the correlation between perception and gesture. Also, given the variation observed across individual participants, it will be important for future studies to include relatively large numbers of observers.
References


