博士論文

THE UNIVERSITY OF TOKYO

Evaluating the Effect of an Ethnic Bias on Speech Perception by Non-native Listeners

(非母語話者の発話知覚における民族バイアスの影響評価)

A Thesis

submitted to the faculty of graduate studies

in partial fulfillment of the requirements for the

degree of Doctor of Philosophy

GRADUATE SCHOOL OF ARTS AND SCIENCES, DEPARTMENT OF

LANGUAGE AND INFORMATION SCIENCES

Tokyo, Japan

March, 2020

Marzena Elzbieta Karpinska

カルピニスカ・マジェナ・エルジビエタ

Supervisory Committee

Evaluating the Effect of an Ethnic Bias on Speech Perception by Non-native Listeners

(非母語話者の発話知覚における民族バイアスの影響評価)

by

Marzena Elzbieta Karpinska

カルピニスカ・マジェナ・エルジビエタ

Supervisory Committee

Izabelle Grenon (Department of Language and Information Science)

Supervisor

Tom Gally (Department of Language and Information Science)

Departmental Member

Yuki Hirose (Department of Language and Information Science)

Departmental Member

Yo Usami (Department of Language and Information Science)

Departmental Member

Chris Sheppard (Waseda University, Faculty of Science and Engineering)

Outside member

Abstract

Speech perception is a multimodal process involving the integration of linguistic information with socioindexical information (e.g., speaker's ethnicity). Native English listeners, for instance, may rate utterances by fellow native English speakers as more accented if they believe the speaker to be Asian than if they believe the speaker to be Caucasian. Similarly, native English listeners may also show lower intelligibility of native English utterances when led to believe that the speaker is Asian than when led to believe that the speaker is Caucasian.

The current research investigates whether native speakers of Japanese will similarly demonstrate lower intelligibility and rate native English speech as more accented when made to believe that the speaker is Asian than when made to believe that the speaker is Caucasian. Additionally, it also considers the notion of comprehensibility, that is the listener's perception of how difficult an utterance is to understand.

Eighty native speakers of Japanese listened to the same English utterances presented either as (1) audio-only stimuli, (2) audio accompanied with a picture of an Asian speaker, (3) audio accompanied with a picture of a Caucasian speaker, (4) video of an Asian speaker, or (5) video of a Caucasian speaker. Additionally, an Implicit Association Test was administered to estimate the strength of the listeners' "American = Caucasian" association, that is, the strength of their possible ethnic bias. The results indicate that all Japanese participants associated the idea of being American more strongly with being Caucasian than with being Asian. However, they did not rate the accentedness or comprehensibility of English utterances differently when presented with a picture or video of an Asian speaker than when presented with a picture or video of a Caucasian speaker. Similarly, the speaker's perceived ethnicity appeared to have no effect on intelligibility scores by the non-native listeners in the current study. These results are discussed in the view of

the Reverse Linguistic Stereotyping and the Experience-based models.

Acknowledgements

This thesis would not be possible without the guidance and constant support of my wonderful supervisor Izabelle Grenon. Izabelle did much more than one would have expected from a supervisor. She not only helped me with the research design, data collection, and constantly gave me feedback on my writing but was also always there when I needed her. I would also like to extend my gratitude to all the members of my committee, Tom Gally, Yuki Hirose, Yo Usami, and Chris Sheppard, whose support and valuable comments helped to shape and improve this thesis.

Many thanks go to my boyfriend who always supported and believed in me even though I did not. Thank you Serge, you did much more than I could ever wish for! I also got a lot of help and guidance from my friend, Anna Rogers (Thank you for showing me LaTeX!), without whom I am sure I would not have succeeded. I would also like to say thank you to my fellow student Leslie, who helped me collect the data and bravely printed out the first version of this thesis.

This work would not be possible without the support from Innovative Language learning and all wonderful people working there. Thank you, Peter, Peyton, Mayumi, and all the fantastic ILL staff for the help you offered me! Finally, I would also like to express my gratitude to my friends, who were always there for me when I needed them. Thank you, Sasha, Piotrs (all of you!), Tin, Erin, Marvin, Boon, Yoav, Grant, Myung, Cho, Jacquie, Ola, Asia, Miuka, Kasia, Oliw, Betsy, Paula and Karolina!

Mojej kochanej mamusi, za wszystko co mi dałaś.

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

ANOVA	Analysis of Variance
CEFR	Common European Framework of Reference for Languages
DME	Direct Magnitude Estimation
IAT	Implicit Association Test
L1	First Language
L2	Second Language
NS	Native Speaker
NNS	Non-Native Speaker
RP	Received Pronunciation
SAE	Standard American English

Epigraph

Any knowledge that doesn't lead to new questions quickly dies out: it fails to maintain the temperature required for sustaining life.

- Wisława Szymborska

Chapter 1

Introduction

Language is one of the most prominent ways of communication unique to the human species. We produce contrastive sounds for other humans to perceive them and to map these sounds onto their linguistic counterparts in order to understand the intended message. This process of mapping acoustic sounds onto their linguistic representations is referred to as *speech perception*.

Speech perception was for a long time regarded as a unimodal process, that is a *solely* auditory event (Denes & Pinson, 1973). However, this view has evolved greatly over the past decades (for a review see Massaro, 2002). The variability in speech, which used to be treated as "noise" and hence ignored, was shown to be accounted for by incorporating socioindexical information into the speech perception model (Drager, 2011; Foulkes, 2010).

A growing body of research indicates that this socioindexical information is combined with linguistic information in a socially-weighted processing of spoken utterances (Kleinschmidt et al., 2018; Sumner et al., 2014). For instance, it has been demonstrated that speech perception can be affected by socioindexical factors such as speaker's age (Koops et al., 2008), gender (Strand, 1999), sexual

orientation (Bouavichith, 2019; Levon, 2007) or ethnicity (McGowan, 2015). This effect of socioindexical information is present regardless of whether such factors are real (Babel & Mellesmoen, 2019; Babel & Russell, 2015; Drager, 2011), perceived due to experimental manipulations (Hay, Warren, & Drager, 2006; Rubin, 1992) or simply inferred (Hay & Drager, 2010).

In the age of ongoing globalization and constantly changing demographics, a significant attention has been given to the speaker's ethnicity, especially in the context of speech perception of English by native English listeners. For instance, it has been demonstrated that perceived ethnicity may enhance the intelligibility of Chinese accented English utterances when they are presented with an East Asian face (McGowan, 2015). However, it may also come at a certain cost (Babel & Russell, 2015). A few studies found that perceived ethnicity may lead to reduced intelligibility of a native English utterance or to mistakenly "hearing" foreign accent in a native English speech if the listeners believe that the speaker is Asian (Babel & Mellesmoen, 2019; Babel & Russell, 2015; Kang & Rubin, 2009; Rubin, 1992; Rubin et al., 1999; Yi et al., 2013).

Studies in the 90's and early 2000's were conducted in the United States where with the increase in immigration (The Brookings Institution, 2019) the problem of potential prejudice and stereotypes became more prominent (Lippi-Green, 2012). This led researchers to seek a possible explanation for the effect of perceived ethnicity on the speech perception in a theory that assumed a *negative* bias on behalf of the listener (Kang & Rubin, 2009; Rubin, 1992; Rubin et al., 1999).

This approach was later challenged by numerous research providing evidence that it is the listener's *experience*, rather than their *negative* bias, that leads to the effect of speaker's perceived ethnicity on the perception of native and non-native English speech (Babel & Mellesmoen, 2019; Babel & Russell, 2015; Gnevsheva,

2018; McGowan, 2015). Researchers then argued that listener's own experience creates certain *expectations* and when these expectations are not met an utterance can be perceived as more accented (Babel & Russell, 2015; Gnevsheva, 2018) or can be less intelligible (Babel & Mellesmoen, 2019; Babel & Russell, 2015).

Surprisingly, despite over two decades of research demonstrating the important role of ethnicity as a socioindexical cue, the majority of studies addressed speech perception by native English listeners (Babel & Mellesmoen, 2019; Babel & Russell, 2015; McGowan, 2015; Rubin, 1992) while the effect of speaker's perceived ethnicity on non-native listeners remains unknown limiting the generalizability of reported findings. If the process of speech perception is indeed a socially-weighted model, that is linguistic information is affected by social information, then it is important to investigate whether this model functions in the same way for both native and non-native listeners. If non-native listeners incorporate socioindexical information in the process of speech perception in the same way as native listeners seem to do, then that would potentially confirm the assumptions of the socially-weighted model and expand its' application to non-native listeners. If, however, non-native listeners would be less sensitive to the socioindexical information then that would suggest a need for developing a refined model, which would equally account for the speech perception by native and non-native listeners.

Furthermore, the previous research on the role of ethnic bias on the speech perception focused on the performance of native English listeners from North America, who were raised in a culture where they were potentially exposed to multiple ethnicities on a daily basis (The Brookings Institution, 2019). However, given the important role of experience, it remains unclear if similar patterns would be observed with listeners coming from a different type of cultural background, such as native Japanese speakers. It is true that Japan has welcomed many immigrants

over the past few decades, but they still make up about only 1.4% of the whole population and the majority of them are of Asian ancestry (about 70%), rather than Caucasian ancestry (about 30%)(Statistics Bureau of Japan, 2015). Thus, it is not clear whether native Japanese speakers, born and raised in Japan, are similarly influenced by the East Asian versus Caucasian ethnicity difference during speech processing.

Although Japan still has a relatively small foreign population, the increasing globalization along with the ease of the immigration law in Japan (Toshihiro, 2019) makes the country more accessible to a large number of foreign workers. Hence, one can expect substantial growth in the number of immigrants seeking jobs in the country. Those of them who have little or no knowledge of the Japanese language will most likely be targeting English speaking jobs. If indeed speaker's ethnicity *alone* can affect the perceived level of accent or even intelligibility then this can potentially lead to misjudgment or even discrimination. As it will be discussed later on in this chapter, speaking English with a foreign accent can, for instance, lead to harassment or discriminatory behavior (Harrison, 2014; Hosoda et al., 2012; Munro, 2003). While most of the research on the implications of speaking accented English was conducted in the native context these implications may extend to non-native English listeners. Hence, it is important to investigate whether speaker's perceived ethnicity affects non-native English listeners as well.

The current work aims at confirming whether the ethnic bias related to perceived accentedness or to the actual intelligibility is also present for non-native English listeners, in this case, native speakers of Japanese. In particular, it evaluates the effect of speaker's ethnicity on the perception of ***accentedness*** — the level of perceived accent — ***comprehensibility*** — the listener's perception of how difficult an utterance is to understand (both measured with rating tasks) — and on the —

intelligibility — the actual degree to which the message was understood (transcription task), of native English utterances as perceived by native Japanese listeners. The main research questions of this work are:

1. Are non-native listeners of English affected by the speaker's perceived ethnicity in the same way as native English listeners seem to be?
2. Does priming ethnicity with different types of visual cues (pictures vs. videos) yield different results?

However, *why* are accentedness, comprehensibility, and intelligibility so important? It may seem uncontroversial to argue for the importance of intelligibility and comprehensibility. The solemn function of speech is, in fact, communication. Reduced intelligibility or even comprehensibility may lead to miscommunication or frustration on the part of both speaker and listener. The same is not as clear when it comes to accentedness. Why does accent matter?

The role of accentedness in communication is controversial. Although many accent reduction programs were introduced over the years it seems like having a foreign accent by itself does not necessarily lead to miscommunication (Munro & Derwing, 1995a) and having less accented speech does not necessarily improve intelligibility (Derwing et al., 1998). However, there are other implications of being perceived as having a non-native accent.

One of the reasons why accentedness received so much attention over the past years is its "ability to convey indexical information" (Munro, 2018) a component of speech, which is widely recognized in the sociophonetics literature (Foulkes, 2010). In particular, accentedness can be responsible for positive or negative evaluation of the speaker. While one may argue whether a positive assessment is problematic or not, negative attitudes toward certain accents can potentially lead to harassment or

discriminatory behaviour (Lippi-Green, 2012; Munro, 2003).

For instance, listeners were shown to evaluate non-native speakers as less convincing, less intelligent, and even less attractive than the native American English speakers even though they were evaluating them based only on the audio recordings (Raisler, 1976). Similarly, students of different ethnic background in the United States were shown to judge recordings of the same bilingual Mexican American speaker to be *less* suitable for a high-status job (such as an engineer), less likely to get a promotion, and less competent when he spoke with a Mexican accent than when he spoke in a Standard American English (Hosoda et al., 2012). Furthermore, native English listeners were demonstrated to rate non-native English speakers as less truthful than native speakers of American English even when they believe that the speaker is conveying a message from a fellow native English speaker (Lev-Ari & Keysar, 2010). In other words, being assessed as having a foreign accent can also result in lower credibility.

Finally, prejudice one holds against a certain group of speakers may also evoke *accent stereotyping* (a term introduced in Munro (2003)), that is, some set of emotions or beliefs, which will be activated upon hearing an accent characteristic for a given group of speakers. As Munro (2003) suggests, this may even lead to discriminatory behavior against a particular speaker only based on their accent.

To summarize, all three of the dimensions of pronunciation, that is accentedness, comprehensibility, and intelligibility can have an impact on both the speaker and the listener. While lower intelligibility or comprehensibility may lead to frustration and misunderstanding, non-native accent may impact several areas of life. For instance, speakers marked with non-native accent can be assessed as less convincing or less intelligent. Therefore, the present research investigates whether the speaker's perceived ethnicity may impact non-native listeners' perception in terms of

accentedness, comprehensibility, and intelligibility of native English speech.

Outline

This work investigates the perception of native English speech, however, English language developed greatly into many different dialects. Hence, **Chapter 2** discusses the concept of World Englishes in the context of speech perception. It first describes a model according to which Englishes are classified into three different circles. In the second part, it discusses the way how different Englishes are being perceived by both native and non-native listeners.

Chapter 3 provides the general overview of the current research and defines the accentedness, comprehensibility, and intelligibility in the way these terms are used in the current thesis. Section 3.1 presents an overview of the current research explaining briefly the two experiments described in the present work. Section 3.2 provides definitions of accentedness, comprehensibility, and intelligibility, and explains the methods used in the current study in order to measure these three dimensions. Furthermore, it provides evidence that non-native English listeners can rate the perceived accentedness of English utterance in a way similar to the native English listeners.

Chapter 4 provides evidence for the role of socioindexicality in the process of speech perception. In particular, it reviews research addressing the effect of socioindexical cues, such as the age, gender, or ethnicity of the speaker, on speech perception. Section 4.1 elaborates on the connection between the social and linguistic information reviewing previous work related to the effect of socioindexical cues on the speech perception process. Section 4.2 addresses the main factor of interest of this work, that is the impact of speaker's ethnicity on speech perception. It discusses a number of research investigating the effect of speaker's perceived ethnicity on speech perception by native listeners. Section 4.3 introduces briefly two

competing theoretical frameworks, which aim at explaining the effect of speaker's ethnicity on the perception of a spoken utterance. Finally, Section 4.4 presents a summary of the previous research and discusses the current design and possible predictions in view of the previous findings.

Chapter 5 introduces two experiments conducted to evaluate the research questions of the current work. Section 5.1 describes the Implicit Association Test aimed at investigating the strength of "American = Caucasian" association of the Japanese participants. The test was administrated to the same participants as in Section 5.2. Section 5.2 presents the main experiment of this work evaluating the impact of speaker's perceived ethnicity on speech perception of native English utterances by non-native (Japanese) listeners. More specifically, it describes three tasks in which Japanese listeners, who also took part in the first experiment, were asked to listen to native English utterances and (1) provide an accentedness rating, (2) provide a comprehensibility rating, and (3) transcribe each utterance (which was used to compile a score for intelligibility). While all groups of Japanese listeners were presented with the same audio recordings performed by native English speakers, some groups were made to believe that the speakers were East Asian-looking while others were made to believe that the speakers were Caucasian-looking. Finally, this section also analyzes the possible correlation between the strength of their "American = Caucasian" bias and the mean scores obtained for accentedness, comprehensibility, and intelligibility.

Chapter 6 is a general discussion providing a summary of both experiments and possible implications of the results. In particular, it discusses a possible explanation for the results of the perception experiment interpreted in the view of the strength of implicit "American = Caucasian" bias of the Japanese participants. It also discusses these results in the view of two competing models: the reverse linguistic stereotyping

and the experience-based models while taking into the consideration the situation of English language in Japan.

Finally, **chapter 7** summarizes the results of the current work referring to the two competing theories that can explain these results.

Chapter 2

World Englishes and Speech Perception

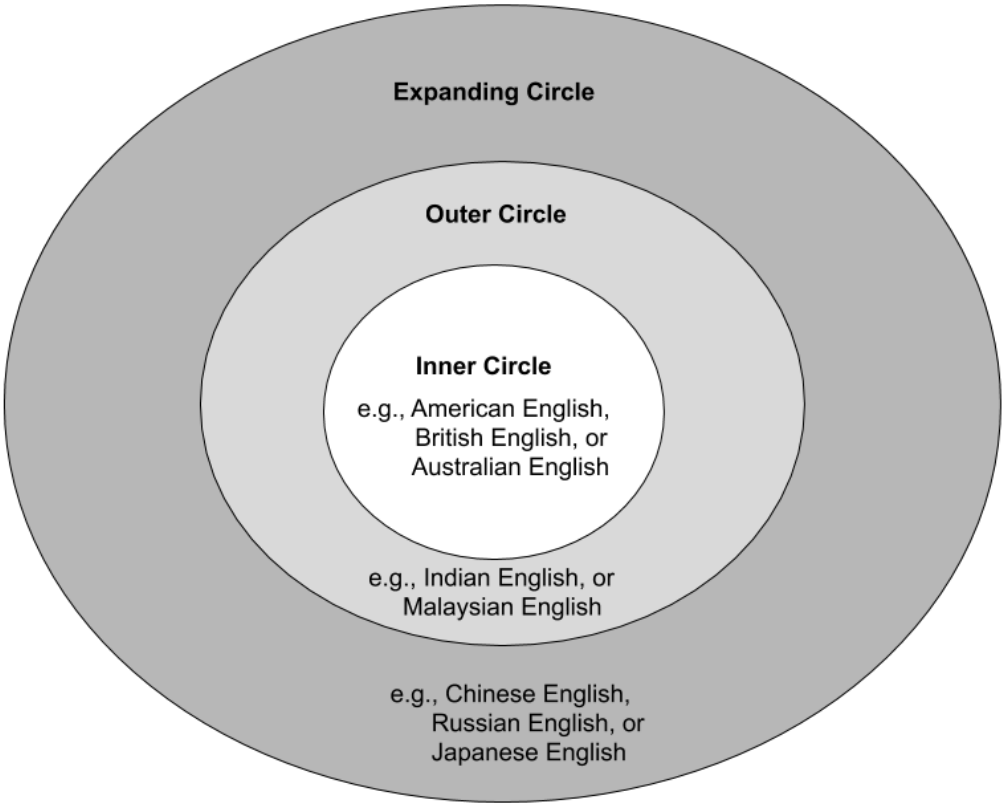
The current work aims at evaluating the effect of an ethnic bias on speech perception by non-native listeners of *English*. However, English evolved greatly in the past few centuries due to the social and political conditions as well as due to the contact with different cultures, other languages, and ideologies (Davis, 2010). In fact, it became hard to talk about English without recognizing the existence of its different varieties, often referred to as World Englishes (WE). This chapter provides a short overview of World Englishes in the context of speech perception.

2.1 Classification of World Englishes

Kachru (1985, 1992) proposed a model where English is classified into three circles (see Figure 2.1). The first circle, called the Inner Circle, includes countries in which most of the population speaks English as their mother tongue. Englishes in this circle developed due to a large-scale migration from England to Australia and North America and they include British English, American English, Canadian English,

Australian English, and New Zealand English.

Figure 2.1: Three Circles of English.



The second circle, referred to as the Outer Circle, includes countries where English is regarded as the *second language* (e.g., India, Malaysia, or the Philippines). These Englishes deviated mostly due to exploitation colonies in Asia and Africa where English was influenced by local languages. While English may not be the native language in these countries, it is often used as a *lingua franca* for communication between, for instance, different ethnic groups.

Finally, the third circle, called the Expanding Circle, includes countries where English is learned as a *foreign language* (e.g., Japan, China, Russia). In these

countries, English does not play any historical role but rather it is learned and used as a way of international communication.

One may assume that because the Expanding Circle includes only non-native speakers of English then the Englishes it includes are necessarily different from Englishes spoken by native speakers. Similarly, the Outer Circle includes very specific Englishes, which are often not taught in schools outside of the countries they developed in. However, even in case of the countries included in the Inner Circle, it is difficult to talk about one English language as these are, in fact, also World *Englishes*. This means that even Englishes in the Inner Circle, which sometimes are referred to as simply "English", do differ from each other in terms of, for example, vocabulary, grammar, or phonetics. For instance, in both Australian English and New Zealand English the vowel /ɪ/ as in *bit* is pronounced differently than in American English (Bell, 1997; Kiesling, 2008). In the Australian English it is pronounced closer to the /i/ vowel as in *beat* while in the New Zealand English it is closer to the /ʌ/ as in *but* (Bell, 1997).

Similarly, there are some major differences between the pronunciation of American and British English (for review see Hosseinzadeh et al., 2015). For instance, Khan and Alzobidy (2018) asked British and American English native speakers to read paragraphs with target words. They found that for words such as *laugh, draft, branch, command, chant, ask, clasp, grass, last, path, gasp* American participants preferred the /æ/ vowel while British participants tended to pronounce the /ɑ/ vowel.

Finally, there are also regional differences within each English from the Inner Circle. For instance, not all native speakers of American English speak in *the same way*. For example, a native speaker from Florida will speak slightly different dialect of American English than a speaker from California (for details see Clopper & Pisoni, 2006; Labov, 1998). All these dialects of American English derived from

social interaction between native speakers in different regions of the United States (Kretzschmar, 2010). Standard American English (SAE), on the other hand, has been established to standardize the language in professional communication and educational system. Kretzschmar (2010) describes SAE as an "institutional construct" that has no native speakers but is "a fact of life" for Americans in formal setting. Kretzschmar adds also that it is a "generalization on a national level of scale abstracted from the speech of educated Americans."

To summarise, World Englishes can be classified into the Inner Circle (e.g., the US or Canada), the Outer Circle (e.g., India or the Philippines), and the Expanding Circle (e.g., Japan or Russia). There is a great diversity among World Englishes, even among Englishes in the Inner Circle where the majority of the population are native speakers. Moreover, for each English in the Inner Circle, there are further regional differences. While those differences may be more apparent in everyday social interaction, the SAE, devoid of regional and socioeconomic characteristics, would normally be used in a formal setting. As it was presented here, English is not a unified language. It is rather a group of languages, which can be jointly referred to as World Englishes. Since the Englishes in this group differ from each other, there will be also some differences in the way in which they are being perceived by native speakers of different variations of English. Hence, the next section discusses the World Englishes in the context of speech perception.

2.2 Perception of World Englishes

Previous section described the three circles of World Englishes. It also presented briefly some examples of how varieties of English in the Inner Circle may potentially differ from each other in aspects such as, for instance, vowel production. This diversity

in production could potentially lead to differences in perception. This section looks at World Englishes, especially these in the Inner Circle, in the context of speech perception as the main aim of the current study is to assess the effect of ethnic bias on speech perception of American English. Thus, it is important to consider how American English may be perceived in contrast to other Englishes from the same circle.

A large body of research evaluated attitudes towards World Englishes of both native and non-native speakers. These kinds of attitudes or beliefs seem to be shaped by listener's experience. Kinzler and DeJesus (2013) presented American children aged 5-10 from Illinois (North) and Tennessee (South) with recordings of Northern and Southern accent. They found out that while younger children did not seem to have any preference towards either type of speaker, the older children (aged 9-10) seem to evaluate Northern accent as *smart* and *in charge*, and Southern accent as *nice*, which is in line with the stereotypes observed in adults.

In another study, Bayard et al. (2001) presented students from Australia, New Zealand, and the United States with recordings of Australian English, New Zealand English, American English, and British English (Received Pronunciation[1]). The students were asked to evaluate speakers on a number of demographic and personality traits. Bayard and colleagues found that, on average, students rated American English most favorably on traits such as status and power, among many others.

The same study was also conducted via the Internet in several Asian countries, such as Malaysia, Indonesia, Korea, China, or Japan, as well as in Europe (Bayard & Green, 2005). Results in Asia confirmed high ratings for American English while British English received low rankings on many traits including status and power. This

[1]Received Pronunciation (RP), also known as BBC English, is considered to be the Standard English in the UK.

may reflect the strong position, which American English seemed to hold, at least at the time of this study, in these countries. Interestingly, in European countries (Sweden, Germany, and Finland) American English was rated high for solidarity but British English achieved high scores in status, prestige, and power.

To summarize, differences in World Englishes can be perceived and shape the attitudes towards speakers of the given variety. Furthermore, regional varieties are present even within one English dialect such as American English. For instance, Californian accent differs from the accent of Florida and accent of New England differs from the accent of New York (Labov, 1998).

These differences make it hard to speak of English as one unified language even though there are many similarities between various Englishes. On the contrary, English is very diverse on many levels, even within the same circle, country or region. Therefore, it is important to address this problem in the current work.

The aim of this study is to evaluate the effect of an ethnic bias on speech perception by non-native listeners of *English*. Since most of the previous work on the effect of ethnicity on speech perception was done in the US employing American English (Kang & Rubin, 2009; McGowan, 2011, 2015; Rubin, 1992) the *English* in this work will refer to American English. The samples for the current study were recorded by native speakers of American English who, despite coming from different regions of the US, did not have any strong identifiable regional dialect. In order to be comparable with the previous studies, the speech samples recorded for the current study were as close to the SAE as possible.

Chapter 3

Research Overview

The current research aims at evaluating the perception of native English, specifically American English speech by non-native listeners (native speakers of Japanese). The first part of this chapter provides an overview of the current research while the second part defines the accentedness, comprehensibility, and intelligibility in the way the terms are used in the present study, as these three dimensions tend to be defined differently in the literature (for review see Levis, 2006). Finally, it also discusses how these three dimensions will be measured in the current study.

3.1 Method Overview

The main aim of this research is to evaluate if the accentedness, comprehensibility, and intelligibility of non-native English listeners will be affected by the speaker's perceived ethnicity. In particular, it investigates whether native Japanese listeners will perceive native English utterances differently when they believe that the speaker is Asian than when they believe that the speaker is Caucasian.

The current research employed a modification of a matched-guise design, a method well established in sociolinguistic research for measuring language attitudes (Kircher,

2015; Lambert et al., 1960). In a traditional matched-guise task, participants would listen to a series of passages being unaware that some of these are spoken by the same speaker. For example, if the researchers were looking at the difference in attitudes towards two languages, such as Spanish and English, they would record a bilingual speaker, speaking both Spanish and English (two *guises*) and compare participant's ratings for both speech samples of the same speaker. The present study is a variation of the matched guise technique in the way that it presents the same auditory stimuli with different visual cues to different groups of listeners. The main idea behind this approach is to separate the auditory information from socioindexical information (i.e., the speaker's ethnicity) so that the voice variable is kept constant while the ethnicity of the speaker is manipulated.

In the current research, 80 native Japanese listeners were asked to listen to native English utterances produced by 10 native speakers from the United States. Each listener was assigned to one of five experimental groups. All listeners, irrespective of the assigned group, completed a *baseline* condition and an *experimental* condition. In the *baseline* condition, participants listened to the same audio-only stimuli and were asked to rate and transcribe it. The main purpose of the *baseline* condition was to ensure that any differences between the groups in the *experimental* condition will be due to the visual cue that is being manipulated, not the differences between the individual listeners.

In the experimental condition (matched-guise design) the (1) control group listened to the audio-only stimuli while the other groups listened to the same stimuli but with either (2) a picture of an East Asian face, (3) a picture of a Caucasian face, (4) a video of an East Asian speaker, or (5) a video of a Caucasian speaker. All participants were asked to rate the utterances for accentedness (how native-like the speech was) and comprehensibility (how easy the speech was to understand). They were also asked to

provide transcription of each utterance to measure the intelligibility. Additionally, all participants completed an implicit association test (Devos & Banaji, 2005; Greenwald, McGhee, & Schwartz, 1998) in order to estimate the *strength* of their "American = White" association. A strong association like that, if present, could lead to rating native utterances as accented when presented with an Asian face. Furthermore, it could also cause a drop in intelligibility ratings, when an utterance was presented with an Asian face as compare to a Caucasian face and audio-only stimuli.

3.2 Terminology

Research on speech production, especially in the second language (L2) context, often evaluates pronunciation treating it as an entity consisting of several partially independent dimensions (Munro, 2018). Munro and Derwing (1995a), Kennedy and Trofimovich (2008), Derwing and Munro (2009), and Hansen Edwards et al. (2018), among many others, discuss the pronunciation in terms of three particular dimensions: *accentedness*, *comprehensibility*, and *intelligibility*. These three attributes describe different aspects of speech in which an utterance can be evaluated. This section provides the definitions of these three dimensions and briefly discusses the way they were assessed in the current study.

3.2.1 Accentedness

Accentedness has been usually defined as the degree in which an utterance differs from (or is similar to) native speech (Derwing & Munro, 2009; Isaacs & Thomson, 2013; Jułkowska & Cebrian, 2015; Kennedy & Trofimovich, 2008; Munro & Derwing, 1995b, 1999). It has also been a common agreement that accentedness can be treated as a continuum where on one side of the scale are utterances produced by a native

speaker and on the other side are utterances with features typical for a non-native speaker.

In the present research, I will adopt the definition proposed by Kennedy and Trofimovich (2008) and define accentedness as "how closely the pronunciation of an utterance approaches that of a native speaker." This will allow to easily construct a 9-point Likert scale, which is the most common and effective method for accentedness evaluation (Derwing & Munro, 2009). In the current study, 1 represents a strong non-native accent (i.e., an utterance that differs significantly from a model native speaker speech), whereas 9 represents native accent (i.e., an utterance indistinguishable from a native speaker).

Native versus Non-native Raters

Researchers generally agree that accentedness is a perceptual phenomenon on the part of the listener (Thomson, 2017), that is, it is something a listener can perceive, and hence, something that the listener can evaluate. However, one may question whether non-native listeners would perform in the same way as native listeners. This subsection provides a short overview of the past research showing that native and non-native speakers are able to judge accentedness in a comparable way.

Native listeners seem to be exceptionally good at distinguishing native from non-native accent (Derwing & Munro, 2009). For instance, untrained native English listeners were able to correctly categorize native and French-accented English utterances just by listening to samples as short as 30 ms (Flege, 1984). Munro et al. (2003) challenged this idea even further by presenting native English listeners with 6 to 10 second long stimuli played backwards. In a backwards speech, features such as segmental, lexical or grammatical information are no longer available. Yet, listeners were still able to distinguish between native and non-native accent above chance

level. Furthermore, this ability to assess accentedness does not seem to be related to the listener's experience, as both listeners who are experienced with a particular non-native accent, as well as novice native listeners, appear to be rating utterances by non-native speakers in a comparable way (Isaacs & Thomson, 2013). This finding suggests that the listener's experience does not influence the ratings.

While it is clear that native listeners have little to no problem with providing consistent accentedness ratings, it may seem unclear whether non-native listeners can also rate accentedness in a consistent way. However, recent studies suggest that non-native listeners exhibit similar rating patterns to native listeners (Crowther, Trofimovich, & Isaacs, 2016; Wester & Mayo, 2014). Furthermore, non-native listeners from different L1 groups generally agree on the accentedness ratings (Crowther et al., 2016). Indeed, even non-native listeners who have *no* familiarity with the language in question demonstrated rating patterns comparable to those of the native listeners and to those of other non-native listeners who are familiar with the rated language (Major, 2007). All these findings suggest that knowledge about the language being evaluated may not be even necessary and that listener's L1 may not affect the ratings.

3.2.2 Comprehensibility

Comprehensibility usually refers to the listeners' perception of how difficult it is to understand a given utterance (Derwing et al., 1998; Jułkowska & Cebrian, 2015; Kennedy & Trofimovich, 2008; Munro & Derwing, 1995b; Saito et al., 2016). The word "perception" used in this definition is very important. It indicates that comprehensibility can be very subjective and may differ depending on the individual. It is the listeners' *belief* of how difficult the utterance is to understand and it may be different from the actual understanding. For instance, an utterance

may be perceived as difficult to understand, yet it can still be perfectly understood. On the other hand, an utterance may be perceived as relatively easy to understand, yet it may not be fully understood by the listener.

Similarly, as for accentedness, comprehensibility ratings are also traditionally collected on a Likert scale (Thomson, 2017). Derwing and Munro (2009) suggest that the best performing scale for comprehensibility ratings is, as for accentedness, a 9-point Likert scale. On a scale like that, which was also employed in the current study, 1 represents an utterance that was very easy to understand while 9 represents an utterance that was very difficult to understand, although the reverse scaling is also common (see, e.g., Thomson, 2017).

3.2.3 Intelligibility

Intelligibility denotes the extent to which a speaker's message was understood by the listener (Derwing & Munro, 2009; Isaacs & Trofimovich, 2011; Jułkowska & Cebrian, 2015; Kennedy & Trofimovich, 2008; Munro & Derwing, 1995a, 1995b; Nelson, 1982). As it was mentioned at the beginning of this chapter, the terms intelligibility and comprehensibility are sometimes used interchangeably (Levis, 2006). However, many researchers have argued that these two dimensions should be disentangled (Kennedy & Trofimovich, 2008; Munro & Derwing, 1995a). While comprehensibility is a subjective measure, something that the listener perceives, intelligibility is more objective and refers to the actual level of understanding.

But how does one measure the *actual* understanding? There is no easy answer to this question. Unlike accentedness or comprehensibility, there is no single measure to determine how much a listener understood. Hence, intelligibility has been measured by assigning a True or False value to a statement (Munro & Derwing, 1995b), by multiple-alternative forced choice identification tasks (Bundgaard-Nielsen et al., 2011;

Hayes-Harb et al., 2008; Thomson, 2017), or by a cloze test (Rubin, 1992). However, the most popular method seems to be a transcription task where either a correctly transcribed keyword (McGowan, 2015) or number of correctly transcribed content words (Derwing & Munro, 1997; Kennedy & Trofimovich, 2008; Sheppard et al., 2017) are being scored. As Derwing and Munro (2009) point out, none of these methods tell the whole story and choosing one over the other will, most likely, be closely related to the purpose of the experiment.

The present research employed the transcription task. While it may be unclear whether words transcribed correctly were actually understood, a transcription of short sentences is less affected by listener's ability to memorize longer passages than a cloze test as presented in Rubin (1992). Moreover, a transcription task does not rely on the listener's vocabulary size to the same extent as a True or False task. This is because one is capable of transcribing the words that are relatively novel to them. Finally, a transcription task does not allow for guessing in the same way as a True or False task.

To summarize, there are three dimensions across which speech can be, and often is, evaluated. *Accentedness* is about how similar an accent is to that of a native speaker, *comprehensibility* is about how difficult an utterance *feels* to be understandable, and *intelligibility* is the actual understanding. Both accentedness and comprehensibility are subjective measures often evaluated using a Likert scale while intelligibility can be measured with a transcription task. Although accentedness may seem difficult to evaluate by non-native listeners, both native and non-native listeners appear to be rating the accent of native and non-native utterances in a relatively comparable way (Crowther et al., 2016; Major, 2007; Wester & Mayo, 2014). The current research employed the matched-guise design in order to investigate how these three dimensions, that is accentedness, comprehensibility, and intelligibility, may be affected by the speaker's perceived ethnicity as operationalized by both pictures and videos of East

Asian and Caucasian speakers.

Chapter 3 provided an outline of the current research and defined the accentedness, intelligibility, and comprehensibility as used in this work. Chapter 4 provides more information about the socioindexical cues discussing how socioindexical information in general and ethnicity in particular may affect speech perception. It also presents two competitive models incorporating social information into the speech perception process.

Chapter 4

Socioindexical Cues and Speech Perception

Language is set in a social context. Social information is encoded in our speech and it may interact with linguistic information (Foulkes, 2010). Since the current research investigate the possible link between socioindexical cues (specifically ethnicity) with speech processing by non-native speakers of English, section 4.1 demonstrates that a connection has been found between various socioindexical cues and speech processing by native listeners. Section 4.2 specifically reviews the literature pertaining to the impact of ethnicity on speech perception. Section 4.3 introduces speech perception models that endeavor to account for these phenomena. Finally, section 4.4 summarizes the previous findings, discussing the current research design in relation to the previous research. It also provides predictions for the outcome of the current perception experiment.

4.1 The Interaction between Social and Linguistic Information

Spoken language conveys more than purely linguistic information. While each utterance has its *linguistic* meaning, it also carries a second layer of *socioindexical* information related to the speaker‘s gender, age, sexual orientation, regional background, or ethnicity (Foulkes & Hay, 2015). This socioindexical information - at least to some degree - is retained in memory alongside the linguistic knowledge (Foulkes, 2010) and it can be accessed during the process of speech perception (Kleinschmidt et al., 2018; Sumner et al., 2014).

There is strong evidence for a connection between the social and linguistic aspects of speech where once could impact the other (for review see Drager, 2010; Thomas, 2002). The earliest studies addressing the presence of socioindexical information concentrated mostly on whether or not social information can be successfully extracted from speech. It has been demonstrated that listeners can consistently infer, among many other social aspects, speaker's ethnicity (Purnell et al., 1999; Trent, 1995; Tucker & Lambert, 1969), socioeconomic status (Shuy, 1969; van Bezooijen, 1988), and sexual orientation (Munson & Babel, 2007; Munson et al., 2006) from auditory stimuli alone.

The strength of this link can vary, and some socioindexical features are certainly more salient than others (Foulkes, 2010; Sumner et al., 2014). However, the connection itself seems to be bidirectional (Kleinschmidt et al., 2018). That is, while social information can be *extracted* from what we hear, it can also *affect* what we hear and how well we understand it. For instance, Niedzielski (1999) demonstrated that listeners may shift their perceptual boundaries based on the information they were given about the speaker's origin. In her study, listeners from Detroit listened to

sentences spoken by a fellow native speaker from Detroit. They were asked to identify vowels in indicated words by choosing one token from a synthesized continuum. About half of the listeners were led to believe that the speaker was from Canada, while the other half were led to believe that the speaker was from Detroit. Both Detroit and Canadian speakers produce the diphthong /aʊ/ as in *about* as a raised nucleus. However, listeners from Detroit are unaware that they also follow this pattern and strongly associate this raising with Canadian English. Niedzielski found that participants who believed the speaker to be Canadian were more likely to indicate that they heard a raised nucleus than participants who believed the speaker to be from Detroit.

A similar effect of vowel shift in perception due to the speaker's regional background was found by Hay, Nolan, and Drager (2006). In their study, they merely implied the nationality of the speaker by writing *Australian* or *New Zealander* on the answer sheet. All listeners were listening to the same speaker from New Zealand. However, the group who had *Australian* on their answer sheet was more likely to indicate that they heard vowels similar to Australian English than the group who had *New Zealander* on their answer sheet. The same effect was found in a follow up study where the nationality was primed only by the presence of plush kangaroos and koalas associated with Australia or plush kiwis associated with New Zealand (Hay & Drager, 2010).

Another study found an effect of perceived gender. For instance, Strand and Johnson (1996) discovered that a variation of the McGurk Effect (McGurk & MacDonald, 1976) is also present for the gender of the speaker. They played a gender-ambiguous continuum of fricatives /s/ and /ʃ/ embedded in carrier words *sod* and *shod* with videos of male and female speakers (within-subject design). The results indicated that listeners were shifting their perceptual boundary to lower

frequencies for the male speakers and higher frequencies for the female speakers. In other words, listeners perceived the same fricative *differently* depending on the perceived gender of the speaker. A similar effect for vowels was found later in a follow up study. Listeners were demonstrated to alter their boundary for the vowels /ʊ/ and /ʌ/ as in *hood* and *hud* pair when presented with, or asked to imagine, a male or female face saying the words (Johnson et al., 1999).

Several studies recognized the impact of perceived age as a socioindexical factor in speech perception. For instance, Koops et al. (2008) conducted an identification experiment in which they played auditory stimuli to listeners from Houston along with a face of a young, middle-aged, or elderly woman. There is a tendency for older Houstonians to merge pre-nasal /ɛ/ and /ɪ/ vowels, often referred to as the PIN/PEN merger. Koops and colleagues found that listeners took significantly longer to identify words with these vowels when the lexical item was presented with an elderly face than when it was presented with a middle-aged face.

Similarly, Drager (2011) researched the shift in vowel perception in New Zealand English depending on two factors: (1) the perceived age of the speaker, and (2) the actual age of the listener. They found that older listeners tend to perceive more lexical items as members of the TRAP set — set of words with /æ/ vowel like in the word *trap* — rather than DRESS set — set of words with /ɛ/ vowel like in the word *dress* (Wells, 1982) — when they were presented with a photography of a young person. This finding is in line with the ongoing chain shift in New Zealand English, where TRAP vowels raise to the space of DRESS vowels, a novel process that would naturally be more apparent for the younger generation. Hence, presenting the stimuli with a young face would create a "congruent" condition, which could also potentially lead to better intelligibility.

This idea was explored by Walker and Hay (2011), who suggested that a match

between the age of the speaker and his lexical choice, that is a "congruent" condition, may facilitate understanding. They presented listeners with words used more by older speakers, words used more by younger speakers, and words that are age-neutral, all of them spoken by both younger and older voices. The results of their experiment indicate that words were recognized more easily when the age of the speaker matched the "age" of the lexical item. This finding was later replicated by Kim (2016) where the stimuli were, just as in Walker and Hay (2011), blocked by the speaker (i.e., participants heard all words uttered by the younger speaker and then all words uttered by the older speaker) and also by Kim and Drager (2018) where the stimuli was *not* blocked by the speaker (i.e., presented in a random order) to avoid any expectations before the presentation of the stimuli.

To summarize, numerous studies provide strong evidence for an interaction between social and linguistic information during speech processing. This social and linguistic information interact with each other in a bidirectional way. That is, socioindexical factors can not only be inferred from a spoken utterance, but they can also alter listeners' perception of that spoken utterance. Moreover, if both social and linguistic information is presented in a congruent condition, as per listener's expectation, social information can facilitate processing of the utterance. Conversely, the incongruent condition may hinder speech perception.

These socioindexical cues, among many others, include the speaker‘s perceived age, gender, regional origin, sexual orientation, and ethnicity. For instance, listeners can shift their perceptual boundaries based on the information they were given or even suggested about the speaker's gender (Johnson et al., 1999; Koops et al., 2008; Strand & Johnson, 1996) or the speaker's origin (Hay & Drager, 2010; Hay, Nolan, & Drager, 2006; Niedzielski, 1999). Furthermore, the speaker‘s perceived age can facilitate speech processing if it matches the "age" of the lexical item (Kim, 2016;

Kim & Drager, 2018; Walker & Hay, 2011). Similarly, the listener's perception of vowels can be altered by the speaker's perceived age (Drager, 2011). The next section reviews work conducted on the impact of the speaker's perceived ethnicity on speech perception separately, as this is the factor of interest in the current thesis.

4.2 Ethnicity as a Socioindexical Cue

The previous section demonstrated how perception can be altered by the speaker's perceived age, gender, or regional background. This section focuses on the speaker's ethnicity, providing evidence that perception of native English listeners (or native listeners in general) may be altered by manipulating the speaker's ethnicity while keeping all other variables constant. The first part of this section presents research on the effect of ethnic bias on the perception of (1) native speech, followed by the effect of ethnic bias on the perception of (2) non-native speech, and finally the effect of the ethnic bias on the perception of (3) native speech contrasted with non-native speech.

4.2.1 The Effect of Ethnic Bias on the Perception of Native Speech

A vast body of research employed native speaker's voice, which was usually presented with a picture of either Asian or Caucasian guise (the matched-guise technique) in order to investigate the effect of ethnic bias on speech perception of native listeners. For instance, Rubin (1992) played two mini-lectures (in humanities and sciences) delivered in SAE by a native English speaker from central Ohio to four groups of native English listeners (undergraduates from a large southeastern university). Each group listened to one of the two lectures while being presented with either a picture of an Asian (Chinese) or a picture of Caucasian guise. All listeners were asked to rate the

accentedness of the speaker on a 7-point Likert scale and to complete a cloze test where every seventh word was deleted (52 blanks). Each listener saw only one face, listened to only one lecture, and provided exactly one accentedness rating. Additionally, one intelligibility score was calculated for each listener by scoring the exact matches from the cloze test. Rubin found that listeners who were presented with an Asian face, regardless of the topic of the lecture, demonstrated poorer intelligibility than listeners who were presented with a Caucasian face. Moreover, the accentedness rating yielded comparable results, that is, listeners who saw the Asian face rated the speaker as more accented than listeners who saw the Caucasian face. Rubin interpreted these findings as evidence for listeners' having a negative bias. He argued that listeners who saw the Asian face could not "hear objectively" and imagined a foreign accent when it was not present.

Almost two decades later, McGowan (2011) found a similar effect of the speaker‘s perceived ethnicity on the perception of accent by native English listeners. He presented single words, rather than mini-lectures, delivered in SAE by a native English speaker from San Diego, California, to two groups of native English listeners in an identification task using an eye-tracking apparatus. The listeners were undergraduate students recruited at the University of Michigan. Prior to the presentation of the spoken words, one group saw a picture of an Asian face while the other saw a picture of a Caucasian face. In the identification task, the listeners saw two pictures and then heard one word. They were asked to look at the picture, which represents the word while their eye-movement was tracked. At the end of the experiment, the participants were asked if their speaker had an accent. Given the binary question, all listeners in the Asian face condition unanimously reported hearing a foreign accent while all listeners in the Caucasian face condition did not. While the results of this study are in line with Rubin (1992), McGowan argued

against the negative bias and hypothesized that listeners might have interpreted the question, which was asked *after* the task, as something like "Did you see an Asian face?"

In yet another between-subject study, Rubin et al. (2015) investigated how perceived ethnicity may affect the evaluation of health care assistants, as opposed to university instructors, in terms of their language proficiency, personal characteristics, and professional competence. Listener's intelligibility was also measured by their proper understanding of the message. As in the previous studies, native English listeners, recruited from seniors centers from a large southern city in the US, were asked to rate the same native English voice (SAE) paired with either a Caucasian or a Mexican guise. This time, apart from the pictures, listeners were also provided information about the guise, such as their names and hometowns. Rubin and colleagues found that listeners in the Caucasian group rated the speaker's accent to be closer to Standard American English than listeners in the Mexican group. They were also more likely to evaluate the same speaker more positively in the Caucasian condition than in the Mexican condition. Rubin and colleagues, just as Rubin (1992), interpreted the results in terms of negative stereotyping.

Hanulíková (2018) provided evidence for the effect of an ethnic bias on the perception of accentedness for a language other than English. Just as Rubin (1992), she played two mini-lectures recorded by the same native Dutch speaker to a group of native Dutch listeners (mostly students). The choice of Netherlands was intentional, as Hanulíková explained, since the populations of the Netherlands is more diverse and more multilingual than the population of the United States in early 90's where the original Rubin's study was conducted. The listeners were presented with with Moroccan and Dutch guises (within-subject design) in clear and adverse conditions (between-subject design). Contrary to Rubin (1992), Hanulíková

did not find any effect of speaker's ethnicity on intelligibility (cloze test) or comprehensibility (7-point Likert scale). However, she found that in the adverse condition listeners perceived the Moroccan guise as *more* accented than the Dutch guise. There was also a significant negative correlation between intelligibility and accentedness ratings. Participants who rated the speaker to be more accented tended to have lower intelligibility scores. Hanulíková interpreted these results as being partially in line with Rubin's findings having argued that this effect was weaker for listeners who had more experience with non-native speakers and therefore were less negatively biased.

Babel and Russell (2015) also found an effect of the speaker‘s perceived ethnicity on both intelligibility and accentedness ratings. Contrary to the other studies, Babel and Russell used recordings of native English speakers of Asian and Caucasian origins using the pictures of the *actual* speakers and not guises. The researcher employed recordings of 120 sentences embedded in noise spoken by 12 native speakers of Canadian English born and raised in Richmond, British Columbia. Half of the speakers were of East Asian complexion and half were of Caucasian complexion. The listeners were native speakers of Canadian English recruited from the University of British Columbia (UBC) community in Vancouver. This choice was intended, as Canada itself, and Vancouver in particular, have a very diverse population with many immigrants from both mainland China and Hong Kong. To back up these claims, Babel and Russel cite a first-year UBC students survey from 2012 saying that 39% of domestic and international UBC students are Chinese, whereas Cantonese and Mandarin account for 30% of first-year students' first language. Given this diversity and the number of immigrants in Canada, the researchers hypothesize that listeners' experience may influence their expectations regarding the speaker.

Babel and Russell asked their participants to listen to and transcribe each of the 120 sentences and to rate a subset of these sentences for the accentedness on a 9-point Likert scale. This procedure is similar to the one employed in the current study. Half of the stimuli for each task was presented with a picture of the speaker (either Asian or Caucasian) and half of the stimuli was presented as audio-only (within-subject design). Babel and Russell found that Chinese Canadians were, on average, *less* intelligible than Caucasian Canadians. While the results in audio-only condition were more convergent, in picture condition Chinese Canadians were noticeably *less* intelligible than Caucasian Canadians. In addition, Chinese Canadians were rated as *more* accented in the picture condition compared to the audio-only condition while Caucasian Canadians were rated as *less* accented in the picture condition compared to the audio-only condition. Yet again, just as in Rubin (1992), simply seeing an Asian face seemed to affect both the accentedness and intelligibility of the speaker.

In addition to the speech perception study, Babel and Russell (2015) administrated an Implicit Association Test (IAT) designed to identify listeners' bias towards Asian speakers. In their IAT, the same participants as in the speech perception task had to classify "Asian" and "Caucasian" surnames along with positive (e.g., vacation) and negative (e.g., death) words. The results of the IAT (referred to as D scores) indicated that the majority of participants associated Asian surnames with negative lexical items and Caucasian surnames with positive lexical items. That is, the participants exhibited a negative implicit bias towards the Asian ethnicity. However, the correlation between D scores and the difference in accentedness ratings of the audio-only stimuli and stimuli presented with Asian faces was *not* significant. Similarly, the correlation between D scores and the difference in intelligibility score of the audio-only stimuli and stimuli presented with Asian faces was also *not* significant.

Contrary to Rubin (1992), Babel and Russell (2015) explained their results with the listener's experience. They argued that in a multicultural and multilingual environment, such as Vancouver, listeners may *expect* to hear an accented speech from a speaker that looks East Asian just because they interact with non-native English speakers of East Asian origins in everyday life. They explained the reduced intelligibility in the audio+picture condition for Chinese Canadians in terms of a mismatch effect. When participants are presented with an Asian face they expect, due to their experience, to hear a foreign accent. However, hearing a perfect Canadian English accent creates a mismatch effect, which leads to lower intelligibility. Babel and Russell's conclusion is supported by the fact that the same speakers in the audio-only condition were more intelligible.

To sum up, the speaker's perceived ethnicity may affect the perception of native speech, whether it is American English, Canadian English (Babel & Russell, 2015; McGowan, 2011; Rubin, 1992; Rubin et al., 2015), or even native Dutch speech (Hanulíková, 2018). Native English listeners were rating native English utterances as more accented when presented with an East Asian face than when presented with a Caucasian face (McGowan, 2011; Rubin, 1992), or as audio-only stimuli (Babel & Russell, 2015). Moreover, the same native English speech was *less* intelligible when presented with an Asian (Chinese) face than when presented with a Caucasian face (Rubin, 1992) or as audio-only stimuli (Babel & Russell, 2015). Similarly, native English speech was rated as more native-like when the listeners believed that the speaker is Caucasian than when listeners believed that the speaker is Mexican (Rubin et al., 2015). Moreover, native Dutch speech in noise was rated as *more* accented when presented with a Moroccan guise than when presented with a Caucasian guise (Hanulíková, 2018). However, no effect of the speaker‘s ethnicity was observed for native Dutch speech in the clear speech condition. Similarly, there

was no significant effect of the speaker's ethnicity on the perception of native Dutch speech for the comprehensibility ratings and intelligibility scores in neither clear nor adverse listening conditions.

4.2.2 The Effect of an Ethnic Bias on the Perception of Non-Native Speech

The previous section reviewed studies that investigated the effect of ethnic bias on the perception of native speech with a focus on native English speech. This section surveys studies that examined whether speaker's ethnicity may alter the perception of *non-native* speech, in particular non-native English speech.

Rubin et al. (1999), just as Rubin (1992), presented the same mini-lecture delivered in English by a native speaker of Dutch to two groups of native English listeners (between-subject design). The Dutch speaker had only a mild foreign accent and was asked to imitate North American English. The listeners for this experiment were recruited from a southeastern university in the United States and had encountered, on average, 1.6 instructors who were not native English speakers. Just as in Rubin (1992), one group listened to the lecture while being presented with an Asian guise while the other group listened to the same lecture while being presented with a Caucasian guise (between-subject design). The Asian guise was presented with a Chinese name and Taiwan hometown while the Caucasian guise was presented with a Caucasian name and American hometown. Like in other studies, participants in the Asian guise group showed *lower* intelligibility than participants in the Caucasian group (cloze test). They also evaluated lecture quality, friendliness of the instructor, and his teaching competence significantly *lower* than the group, which saw the Caucasian guise. Rubin and colleagues interpret these findings in terms of negative stereotypes on behalf of the listener. They conclude that this negative bias may affect students' perception of

foreign instructors.

McGowan (2015) questions the negative bias providing evidence that seeing an East Asian face can *enhance* the intelligibility of Chinese accented English. The researcher played a series of sentences recorded by non-native English female speaker (native speaker of Mandarin Chinese) to three groups of native English listeners (between-subjects design). The listeners in McGowan's study were native speaker of American English, recruited at the University of Michigan and University of California. Similarly as in Rubin (1992), one group saw a picture of an East Asian face and one group saw a picture of a Caucasian face. McGowan also included a control group, which saw only an ambiguous silhouette. Listeners were asked to transcribe sentences of Chinese accented English presented in noise. McGowan found that listeners who saw the Asian face were significantly *better* at the transcription task than listeners who saw the Caucasian face. In other words, seeing an Asian face while hearing a Chinese accent (congruent condition) improved the intelligibility of Chinese-accented English. He argued that listeners were better at the transcription task when presented with an Asian face than when presented with a Caucasian face because, based on their experience; they were *expecting* to hear Chinese accent and this is what they actually heard. On the other hand, listeners in the Caucasian face condition were *expecting* to hear a native accent so when they heard Chinese accent instead, it made it more difficult for them to understand the utterance.

Contrary to Rubin et al. (1999) and McGowan (2015), one early study also employing non-native English speech did *not* find any effect of ethnicity on neither accentedness ratings (7-point Likert scale) nor intelligibility (cloze test). Rubin and Smith (1990) presented native English listeners, recruited at the University of Georgia, with mini-lectures delivered by two non-native English speakers (native

speakers of Mandarin Chinese). Each listener listened to only one mini-lecture delivered by one of two female speakers in one of two accentedness levels (moderate or high) on one of two topics (science or humanities). During the task, the listeners were presented with either an Asian (Chinese) face or a Caucasian face. After listening to the lecture, each listener provided exactly one accentedness rating using a 7-point Likert scale. Similarly, one intelligibility score was computed for each listener by scoring only the exact matches on a cloze test (52 blanks). In addition to this, listeners were asked to complete a questionnaire including questions such as whether the speaker was Asian or Caucasian as well as some more fine-grained questions about listener's stereotypes and beliefs. While this setting is similar to Rubin (1992), in this study, Rubin and Smith did *not* report any effect of ethnicity on accentedness ratings or the intelligibility scores. They only found that more accented speech was perceived as more Asian compared to the less accented speech and that the degree to which the students *believed* the speech to be accented was a good predictor of the way they were rating instructor's teaching abilities. The more accented the speech was perceived to be, the lower ratings did the instructor receive for her teaching abilities. This may come to no surprise if we take into account the fact that 40% of participants choose to drop a class after finding out that the teacher is a non-native speaker of English on at least one occasion. Rubin and Smith conclude that students should be encouraged to take classes taught by non-native instructors in order to get familiar with the non-native speech. In their view, giving the non-native instructor the benefit of the doubt could potentially lead to greater satisfaction with the academic work as well as to better listening abilities.

Zheng and Samuel (2017) questioned whether the effect of speaker's ethnicity observed in Rubin (1992) is present at the perception level. They hypothesized that listeners may not *perceive* non-native accent but rather *decide* that they hear a

non-native accent when presented with an Asian guise. Zheng and Samuel incorporated non-native English stimuli prepared by blending native American English recordings with Chinese accented English recordings. Unlike the other studies, Zheng and Samuel (2017) presented native English listeners, undergraduates from Stony Brook University, with a synthesized continuum ranging from slightly accented to moderately accented speech (single words). Listeners were then asked to listen to each word and to rate the accentedness on a 4-point Likert scale. One group of listeners was presented with an Asian face while the other group was presented with a Caucasian face. After a break, the listeners were asked to perform the same task but this time the pictures were reversed, that is the Asian group saw Caucasian guise while the Caucasian group saw Asian guise. The first half of this experiment was, in fact, a between-subjects design similar to Rubin (1992) and just as in Rubin's study Zheng and Samuel reported the same effect of speaker's perceived ethnicity. However, the results in the second part were reversed, that is the participants perceived the Caucasian guise as *more* accented then the Asian guise. When analyzing the data from both parts collectively, the results have changed and the effect of ethnicity was not significant anymore.

Zheng and Samuel (2017) then repeated this experiment with the same audio stimuli presenting native English listeners with dubbed videos rather than just pictures of the speaker. According to the researchers, videos reduce the demand characteristics when compared to pictures, meaning that the listeners are less likely to guess the purpose of the experiment and act accordingly. This time they did *not* find any effect of face by analyzing only the first half of the data. They did, however, find a weak effect of the speaker's ethnicity when analyzing the data collectively, that is Asian guise was rated as more accented than the Caucasian guise. While Zheng and Samuel do not negate the role of ethnicity as a

socioindexical cue, they argue that the ethnicity of the speaker affects listeners' *interpretation* rather than *perception*.

To summarize, non-native English speech was less intelligible when presented with an East Asian face than when presented with a Caucasian face (Rubin et al., 1999). Moreover, the same instructor was rated as less friendly and less competent when native English listeners were led to believe that he is Asian than when they were led to believe that he is Caucasian (Rubin et al., 1999). On the other hand, Chinese accented speech was *more* intelligible when presented with an East Asian face than when presented with a Caucasian face (McGowan, 2015). Furthermore, in one early study, no difference was found in terms of accentedness and intelligibility of Chinese accented speech between an Asian (Chinese) guise and a Caucasian guise (Rubin & Smith, 1990). However, the accent ratings correlated positively with ratings of teaching qualities, the more accented the speech was perceived, the lower was the teacher rated for the teaching qualities (Rubin & Smith, 1990). Finally, Zheng and Samuel (2017) provided evidence that presenting listeners with pictures may bring demand characteristics and yield results different than when presenting participants with videos.

Rubin et al. (1999) argued that the effect of ethnic bias on speech perception is due to a negative bias on behalf of the listeners. On the other hand, McGowan (2015) provided evidence in favor of an experience-based approach. He argued that it is not the negative bias but rather the listener's experience that affects the listener‘s perception. Finally, Zheng and Samuel (2017) questioned whether the perception is being affected and argued that this effect may take place not on the perception level but rather on the interpretation level.

It seems that when non-native English speech is incorporated into the design then the effect of the speaker‘s perceived ethnicity is less apparent than when native English

speech is employed. While one study discovered an effect of an ethnic bias on the intelligibility of Dutch accented English (Rubin et al., 1999), another did not find any effect of ethnic bias on the intelligibility of Chinese accented English (Rubin & Smith, 1990). Yet another study found that the effect may vary depending on how the guise is created (picture vs. video). Finally, one study found a positive effect of speaker's perceived ethnicity on the perception of Chinese accented speech (McGowan, 2015). This inevitably brings some doubts about whether it is indeed the *negative* bias and not merely listener's expectations that lead to a different evaluation of native and non-native English speech when presented with an East Asian guise than when presented with a Caucasian guise.

4.2.3 The Effect of an Ethnic Bias on the Perception of Native Speech Contrasted with a Non-Native Speech

The previous section reviewed studies, which investigated the effect of ethnic bias on the perception of non-native speech. This section surveys studies evaluating the effect of ethnic bias on the perception of native English speech contrasted with non-native English speech.

Rubin et al. (1997) presented native English listeners from a southeastern university in the US with a message about AIDS delivered by a speaker with either a moderate South Asian accent, heavy South Asian accent, or Standard American accent. The moderate and high South Asian speech samples were recorded by a bilingual male speaker from New Delhi while the native English sample was recorded by a native male speaker of North American English who was asked to adjust his pitch and speed to match those of the bilingual speaker. Each listener listened to only one of three samples, which was presented with a picture of either South Asian

male or Caucasian male. Participants were then asked to summarize the message as a measure of intelligibility and to complete a simple questionnaire. While Rubin and colleagues did not find any effect of ethnicity on the recall score, they found that Caucasian guise was rated higher for interpersonal attractiveness than the Asian guise when the speaker had North American accent. They conclude that no effect of speaker‘s ethnicity on the intelligibility should be approached with “guarded optimism” as the overall recall rate was under 20%, which may suggest that the nature of this message carries some emotional barriers to effective listening.

de Weers (2019) provided additional support for the possible lack of the effect of ethnic bias on the intelligibility of native and non-native English utterances. De Weers presented native English listeners with native and non-native English utterances embedded in noise. The native listeners in this study came from Canada (20), the UK (7), the US (2), South Africa (2), and Ireland (1). Unlike in Rubin et al. (1997), the listeners listened to all utterances and saw both East Asian and Caucasian faces (within-subject design). The stimuli for this study consisted of statements recorded by two female native speakers of Canadian English and two female native speakers of Japanese. Each utterance was paired once with an East Asian face, once with Caucasian face and once with no face at all. Listeners were asked to rate the accentedness of each statement on a 7-point Likert scale and to choose whether the statement was true or false (intelligibility). Additionally, the response time was recorded as a measure of comprehensibility, with longer time implicating that the message more difficult to understand. Similarly to Rubin et al. (1997), de Weers did not find any effect of ethnicity on intelligibility. Moreover, there was also no effect of ethnicity on the accentedness ratings or comprehensibility as measured by the response time. She argued that the null effect of speaker‘s perceived ethnicity could be due to the research design. While most studies used

between-subject design with no control group (McGowan, 2015; Rubin, 1992; Rubin et al., 1999; Rubin & Smith, 1990), de Weers employed a within-subject design.

Gnevsheva (2018), on the other hand, provided more evidence that listeners *expect* Caucasian speaker to be a native speaker of English while at the same time, just as McGowan (2015), questioning the negative bias. In her study, Gnevsheva presented native English listeners with recordings of 18 non-native speakers of English (9 Korean and 9 German) and 6 native English speakers (2 from New Zealand, 2 from the UK, and 2 from the US). Listeners, who were themselves native speakers of English from New Zealand, were asked to rate accentedness of short utterances in one of three conditions: (1) audio-only condition, (2) video-only condition, or (3) audiovisual condition (between-subjects design). While Korean native speakers were being rated consistently as accented across *all* three conditions, German native speakers were rated as *less* accented in the video-only condition compared to the audio-only condition and as *more* accented in the audiovisual condition compared to audio-only condition. Moreover, in the soundless video-only condition listeners rated the accentedness of native English speakers and native German speakers in a comparable way, which may suggest that they were not able to infer the foreign accent based on the video alone. Gnevsheva explains that the listeners may have *expected* native English speech from a Caucasian speaker; hence they rated native and non-native Caucasian-looking English speakers in the same way in the video-only condition. Gnevsheva argued that these expectations could also cause the German native speakers to be rated as *more* accented in the audiovisual condition than in the audio-only condition. Gnevsheva interpreted her results in terms of a mismatch effect, that is, the listeners in audiovisual condition *expected* to hear native English speech from a Caucasian speaker, yet when they heard a non-native accent that accent "stood out" even more as it was unexpected.

Gnevsheva also argued against negative bias endorsed Rubin (1992), as Korean native speakers were rated in a comparable way across all three conditions. She explains her result entirely in terms of the listener‘s experience, which would lead to certain expectations.

Contrary to Gnevsheva, Yi et al. (2013) observed the effect of the speaker‘s ethnicity on the perception of English spoken by native Korean speakers. They presented monolingual native speakers of American English, recruited at the University of Texas, with English sentences recorded by either two native speakers of American English or two native speakers of Korean (between-subject design). In each pair of speakers, one speaker was a male and one speaker was a female. Each group of participants listened to audio-only and audiovisual recordings of both speakers (within-subject design). Listeners were then asked to transcribe each sentence and the intelligibility score was computed by scoring the correct transcription of the content words. Yi and colleagues found that while both native and non-native English speakers were *more* intelligible in the audiovisual condition than in the audio-only condition, this effect was significantly greater for native speakers.

In an additional experiment, Yi et al. (2013) presented the same stimuli to another group of six native speakers of American English, also recruited at the University of Texas. The listeners were asked to listen to *all* of the stimuli, that is to both native and non-native samples presented in both audio-only and audiovisual condition. The sentences were presented in a randomized order and the listeners were asked to rate the accentedness of each utterance on a 9-point Likert scale. Yi and colleagues found that native English speakers were perceived as *less* accented in the audiovisual condition when compared to the audio-only condition while Korean speakers were perceived as *more* accented in the audiovisual condition when

compared to the audio-only condition. It is worth noting, however, that this difference in accentedness ratings, albeit significant, was only about 1.8% (0.162 points on a Likert scale) for Korean speakers and 0.9% (0.081 points on a Likert scale) for native English speakers, which may raise some questions as to the validity of the analysis especially since the authors do not report the effect size. Given the small difference and small sample size (only 6 listeners), it is possible that with more listeners, this effect would not have been significant.

Yi and colleagues have also asked all of their listeners to take an Implicit Association Test (IAT) in order to assess whether they were implicitly associating being American with being Caucasian. Participants were presented with a series of pictures of East Asian and Caucasian faces and of American and foreign places in a series of discrimination tasks. Their implicit bias was then measured by computing the response time in the congruent condition (American places and Caucasian faces sharing the same response key) and incongruent condition (American places and Asian faces sharing the same response key). Yi et al. (2013) found that all of their participants had an implicit bias towards the American+Caucasian pairing, that is, they associated being American with being Caucasian. The researchers further investigated the relationship between the IAT effect and the intelligibility scores. The results indicated that participants with stronger "American = Caucasian" bias had better intelligibility scores in the audiovisual condition for native English speakers than for non-native English speakers. At the same time, there was no such effect for the audio-only condition. Yi and colleagues conclude that listener-related factors, such as "non-linguistic visual bias", play a significant role in the process of speech perception.

Babel and Mellesmoen (2019) contributed to this discussion by incorporating native and non-native English voices into a within-subject design while presenting

videos of the actual speakers. The researchers recorded videos of two females, native speakers of Canadian English and two females, native speakers of Spanish and Mandarin Chinese, uttering high and low predictability sentences embedded in noise. Just as Yi et al. (2013), Babel and Mellesmoen did not employ the traditional matched-guise design but rather used the actual speakers. One of the speakers in both native pair and non-native pair was Asian while the other was Caucasian. Participants for this study were native or near native (age of acquisition ¡ 5) English listeners recruited at a university in Canada. All participants listened to all speakers and were asked to transcribe what they heard. High predictability sentences were presented first followed by low predictability sentences. Surprisingly, the intelligibility of Caucasian native speakers and Asian non-native speakers both *increased* in the second part of the experiment while intelligibility of Asian native speakers and Caucasian non-native speakers decreased in the second part of the experiment. Babel and Mellesmoen explain their results with listeners' expectations. They argue that listeners adapted better to speakers whose ethnicity matched their actual accentedness as per stereotypes (but not *negative* bias) that Asian speakers will speak with a foreign accent and Caucasian speakers will speak with a native English accent.

To summarize, the effect of ethnic bias on speech perception seems to be more complex when both native and non-native voices are included. While some studies reported no effect of the speaker‘s ethnicity on the intelligibility scores as measured by a cloze test (Rubin et al., 1997) or by evaluation of true/false statements (de Weers, 2019), others found an effect on intelligibility as measured by a transcription task (Babel & Mellesmoen, 2019; Yi et al., 2013). Similarly, the effect varied for the accentedness ratings. While Gnevsheva (2018) did not find any effect of the speaker‘s ethnicity on Korean accented English, Yi et al. (2013) reported an effect of

ethnicity on Korean accented English. Interestingly, the expectations that Caucasian speaker will be a native English speaker can also affect Caucasian-looking non-native English speakers (Gnevsheva, 2018). Finally, Gnevsheva (2018) and Babel and Mellesmoen (2019) provided additional evidence attributing the effect of ethnicity on speech perception by native English listeners to the listeners *experience* rather than their *negative* bias argued by Rubin (1992) and later by Kang and Rubin (2009).

This section presented studies evaluating the effect of the speaker‘s perceived ethnicity on the speech perception of native and non-native English speech. While some researchers attributed this effect to the negative bias on behalf of the listener (Kang & Rubin, 2009; Rubin, 1992; Rubin et al., 1999; Yi et al., 2013) others provided additional evidence that it is not the speakers‘ bias but rather their experience that affected their speech perception (Babel & Mellesmoen, 2019; Babel & Russell, 2015; Gnevsheva, 2018; McGowan, 2015). The next section discusses in more detail these two approaches introducing the Reverse Linguistic Stereotyping (Kang & Rubin, 2009; Lippi-Green, 2012) model and the experience-based models (Johnson, 2006; Kleinschmidt et al., 2018; Sumner et al., 2014).

4.3 Socially-weighted Models of Speech Perception

The previous sections discussed how socioindexical cues interact with linguistic information in the process of speech perception. Therefore, socioindexical features should be accounted for in a speech perception model. This section discusses two fundamentally different approaches to this problem. The first part of this section introduces the Reverse Linguist Stereotyping (RLS) model (Kang & Rubin, 2009), while the later part discusses experience-based models (Johnson, 2006; Kleinschmidt

et al., 2018; Sumner et al., 2014)

4.3.1 Reverse Linguistic Stereotyping

Rubin (1992) attempts to explain the effect of perceived ethnicity on speech perception in terms of a *negative bias* on behalf of the listener. He argued that listeners who saw an Asian face were "*incapable* of hearing objectively." In particular, Rubin advocates that listeners hold a negative social bias against, for instance, Asian-looking English speakers, and that this bias leads to a negative social evaluation, which results in reduced intelligibility or perception of a foreign accent even when it is not present.

This idea was endorsed by Lippi-Green (2012), who argued that listeners' negative stereotypes towards Asian-looking English speakers lead to a *communicative breakdown*. In other words, listeners *choose* not to pay attention to the English utterance when the speaker is Asian-looking.

Kang and Rubin (2009) put this idea into a new theoretical model, which they called the Reverse Linguistic Stereotyping (RLS). In their view, RLS is supposed to be "a converse of the linguistic stereotyping hypothesis." Kang and Rubin explain that while in the linguistic stereotyping language is the trigger for certain behavior, in the RLS it is the object of stereotyping. They argue that listeners would, therefore, extend their negative beliefs about certain groups of speakers (like Asian-looking English speakers) to individual members of these groups.

While RLS model is still present in the literature, it has been challenged by several studies. McGowan (2011), for instance, successfully demonstrated that listeners can actually *benefit* from seeing an Asian face if it is combined with non-native English speech, that is, when the accent matches certain expectations of the listener. This led to a popularization of the experience-based models. While there may be a *negative* bias towards Asian-looking English speakers, it seems that it is not present at the

level of unconscious perception as it was argued by Rubin and colleagues (McGowan, 2015). The next part of this section introduces an alternative to the RLS — the experience-based models.

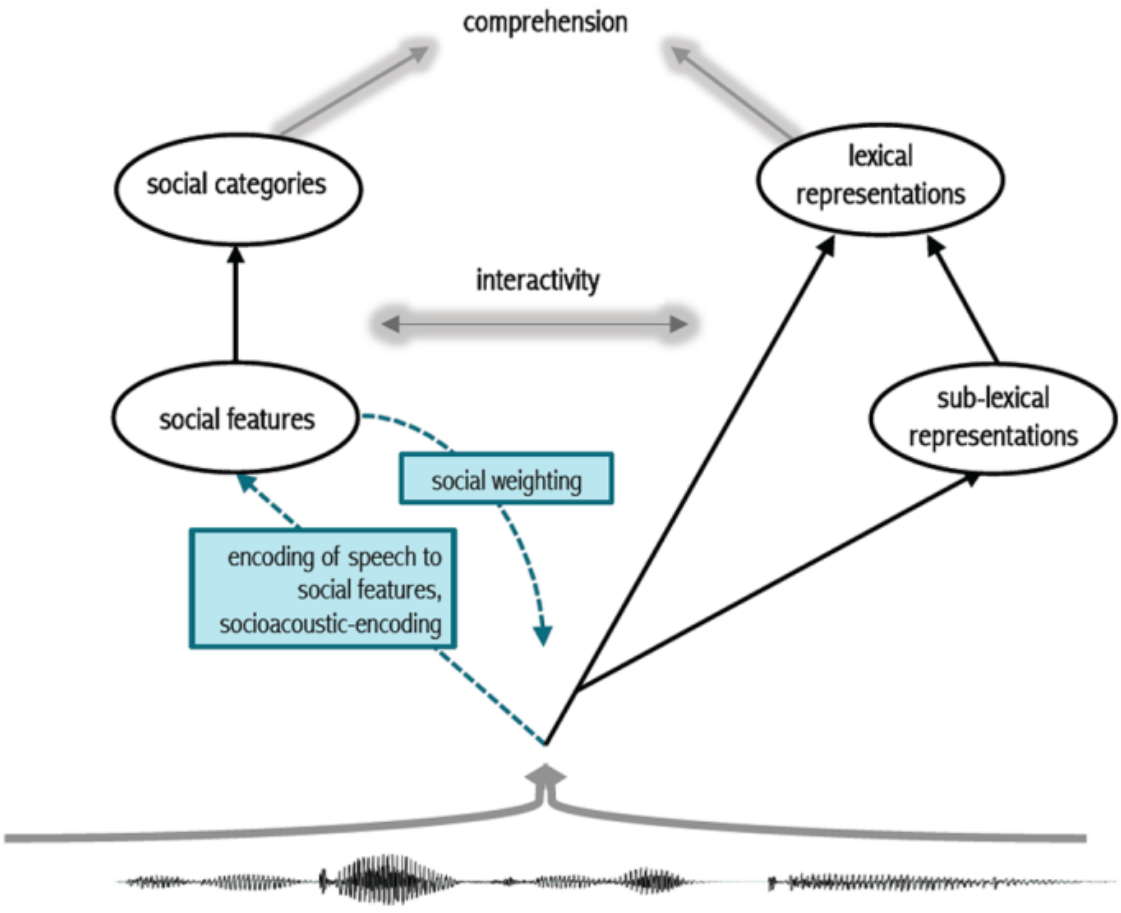
4.3.2 Experience-based Models

Experience-based models are powerful in their ability to account for both negative and positive effects of the speaker‘s perceived ethnicity. This means that they can explain the results that RLS is incapable of explaining. Moreover, they can easily account for not only ethnicity but also other socioindexical information.

The base for experience-based models is the exemplar theory (Foulkes, 2010; Foulkes & Hay, 2015). Exemplar theory assumes that episodic traces are being stored in memory in order to be activated when presented with a consistent social category (e.g., speaker's ethnicity). This means that listeners, depending on their previous experience, may *expect* a non-native English accent upon seeing an Asian face. In this case, presenting Asian face with a native English accent would create an incongruent condition in which the utterance maybe be less intelligible and appear to be more accented. Similarly, presenting Asian face with a non-native English accent, especially one that can be associated with an Asian face such as a Chinese face, would create a congruent condition which, as a result, would increase the intelligibility (Babel & Mellesmoen, 2019; McGowan, 2015). This approach would also explain the results reported in Gnevsheva (2018). Korean native speakers were rated similarly accented across all three experimental conditions (audio-only, video-only, and audiovisual) not because the listeners were discriminating against Asian speakers but because they *expected* them to have a foreign accent. On the other hand, German speakers were rated as *more* accented in the audiovisual condition than in the audio-only because the listeners *expected* them to sound like

native English speaker, so that seeing a Caucasian face paired with a non-native speech created a mismatch effect and led them to rate the utterances as more accented. The same German speakers were rated as *less* accented in the silent video-only condition than in the audio-only or audiovisual conditions because the listeners, yet again, were expecting a native English accent from a Caucasian speaker.

Figure 4.1: Socially-weighted speech perception model.



Note: Reprinted from Sumner et al. (2014).

The exemplar theory laid grounds for the development of a socially-weighted speech perception model. Sumner et al. (2014) introduce a dual-route model in

which an utterance is being parsed into multiple social and linguistic information. In this one-to-many approach, the speech signal is being mapped simultaneously to social representations and to lexical representations either directly or through smaller sub-lexical representations (see Figure 4.1). This process allows to construct a *link* between the social representations and linguistic representations, which interact with each other. Sumner and colleagues call this process *social weighting*. In their view, it is the social weighting that allows to choose particular linguistic representations over others. One important feature of this model is that it does not rely strongly on raw frequencies. For instance, Sumner and colleagues hypothesize that word *smiling* would be recognized more easily when pronounced in a happy voice than when pronounce in a sad voice or when pronounced in a neutral voice. Although listeners hear the word *smiling* uttered in a neutral voice far more often than in a happy voice, they can parse the happy emotion (here a socioindexical cue) from the speech and use this information in the process of social weighting to recognize the word more easily.

Kleinschmidt et al. (2018) proposed an additional explanation of "social weighting" introducing the ideal adapter model. The ideal adapter model, in line with the exemplar theory, assumes that listeners can learn and store socioindexical information as accessible categories. Listeners can then only *probabilistically infer* how possible a linguistic unit is based on the cue distribution. Kleinschmidt and colleagues describe this relationship between inferences and cue distributions using Bayes theorem. Thus, listeners depend on their previous *experiences* in order to estimate how likely a cue is, given the social category and to make probabilistic inferences.

As it was demonstrated above, the experienced-based models provide a rich framework that can successfully incorporate ethnicity of the speaker along with

other social information into the speech perception model. In this way, they are much more productive than the RLS, which is based on the assumption that there is a negative bias toward Asian-looking, or more generally non-Caucasian, English speakers. One potential strength of the experience-based models is that they can account for both negative (Babel & Russell, 2015; Rubin, 1992) and positive (McGowan, 2015) effects of socioindexical cue as well as on the effect of ethnicity on the perception of non-native Caucasian English speakers (Gnevsheva, 2018). The next section provides a summary of this chapter along with predictions for the performance of non-native listeners.

4.4 Summary of the Previous Findings and Predictions for Non-native Listeners

Several studies researched the effect of ethnic bias on the speech perception by native listeners, in particular by native listeners of English (e.g., Babel & Mellesmoen, 2019; McGowan, 2015; Rubin, 1992). Ethnicity of the speaker, whether actual or just perceived due to the experimental manipulation, was shown to alter the accentedness ratings (Babel & Russell, 2015; Gnevsheva, 2018; Kang & Rubin, 2009; McGowan, 2011; Rubin et al., 1999, 2015; Yi et al., 2013; Zheng & Samuel, 2017) and the intelligibility scores (Babel & Mellesmoen, 2019; Babel & Russell, 2015; McGowan, 2015; Rubin, 1992; Yi et al., 2013) of native English listeners from the United States (Rubin, 1992; Rubin et al., 2015, 1997; Yi et al., 2013), Canada (Babel & Mellesmoen, 2019; Babel & Russell, 2015), and New Zealand (Gnevsheva, 2018).

Moreover, having a "non-native" face does not necessarily induce a negative effect. While pairing native speech with East Asian face seems to *reduce* intelligibility (Babel

& Mellesmoen, 2019; Rubin, 1992), pairing non-native speech with East Asian face appears to *enhance* it (Babel & Mellesmoen, 2019; McGowan, 2015).

It is also important to mention that few studies did *not* report any effect of speaker's ethnicity on the accentedness ratings and intelligibility scores (de Weers, 2019; Rubin et al., 1997; Rubin & Smith, 1990). All these studies employed non-native English utterances, either exclusively or contrasted with native English utterances. Hence, it is possible that this effect is more apparent when only *native* English speech is presented with an East Asian face and a Caucasian face than when non-native English speech is paired with an East Asian face and a Caucasian face.

The effect of ethnic bias on speech perception was first explained by the RLS — a theory, which assumes negative bias on behalf of the listener. The first mentions of this *negative* bias in the context of speech perception and ethnicity of the speaker appear in the early 90's when the globalization and multiculturalism, while already in motion, were less pronounced. Hence, negative stereotypes appeared to be a reasonable explanation for this phenomenon.

However, nearly 20 years later McGowan (2011) provided evidence for a positive effect of East Asian ethnicity on the perception of non-native English speech questioning the negative bias hypothesis. With the ongoing globalization and constantly changing demographic of the United States and Canada, where most of these researches were conducted, this discovery led to a new theory based on the experience of the listener. The experience-based model, or more specifically socially-weighted speech perception model (Sumner et al., 2014), provided the more versatile framework to account for the effect of ethnicity on the speech perception by native English listeners by widening the application to negative as well as positive effects.

The current research evaluates whether an effect of ethnic bias will also be

present for the non-native English listeners from Japan, where the number of foreign residents remains relatively low (1.4%) with the majority of them being of Asian origins from countries such as China where the official is not English (Statistics Bureau of Japan, 2015). In order to do so, this study employs only native English voices (SAE) since some previous studies reported no effect of the speaker‘s ethnicity when non-native English voices were also employed (e.g., de Weers, 2019; Rubin & Smith, 1990). Furthermore, since the majority of previous studies employed either male (Hanulíková, 2018; Rubin et al., 1999, 2015, 1997) or female (de Weers, 2019; McGowan, 2015; Rubin et al., 1999; Rubin & Smith, 1990) speakers, the current study also investigates the possible effect of speakers' gender, as well as the possible interaction between gender and perceived ethnicity. It is possible, for instance, that this effect of ethnic bias will be stronger for one gender than the other. Finally, the current study explores whether the effect of ethnic bias will be stronger (or weaker) for guises presented with video stimuli than for guises using pictures.

In the current study, sentences recorded by native English speakers were evaluated for the perceived accentedness on a 9-point Likert scale as in Yi et al. (2013) and Babel and Russell (2015). Furthermore, intelligibility was measured with a sentence transcription task as in Yi et al. (2013), Babel and Russell (2015), and Babel and Mellesmoen (2019). While some other studies also included cloze test administered right *after* the listening task (e.g., Hanulíková, 2018; Rubin, 1992) this seems to be too challenging for non-native speakers as it requires not only a good understanding of the content of the sentence but also good short term memory. Hence, it can be questioned, even in the case of the native English listeners, whether the difference in intelligibility was due to the perceived ethnicity or just ability of the listeners to memorize the utterances properly. Additionally, comprehensibility was measured on a 9-point Likert scale to evaluate how difficult the utterances felt like to the non-

native listeners. Speaker's ethnicity could be inferred from pictures of East Asian and Caucasian guises (e.g., Babel & Russell, 2015; McGowan, 2015; Rubin, 1992; Rubin et al., 1997; Rubin & Smith, 1990) or with videos of East Asian and Caucasian guises (e.g., Babel & Mellesmoen, 2019; de Weers, 2019)[1].

In addition to the perceptual tasks, an Implicit Association Test (IAT) was administrated as in Yi et al. (2013) in order to establish if non-native English listeners in the current study were implicitly associating being an American with being Caucasian and, more importantly, evaluate the strength of this association. While American is not an equivalent for a native English speaker, as there are native English speakers in Canada, Australia, New Zealand or the UK, among many others, since the dialect of English used in the current study is the American English, it is possible that associating the concept of American with being Caucasian may translate into associating native speaker of American English with being Caucasian. Furthermore, Kubota and Fujimoto (2013) claimed that Japanese native speakers tend to associate the concept of being native English *teacher* with being Caucasian. This claim was born from the idea embraced by Kubota, Fujimoto, and other researchers that Caucasians or Americans are being "worshiped" in Japan. This phenomenon is often referred to as *hakujin sūhai* ("worshiping Caucasians") and *Amerikajin sūhai* ("worshiping Americans"). If the claim made by Kubota and Fujimoto is indeed true, then participants showing a strong implicit "American = Caucasian" association, may very likely be affected by the speaker's perceived ethnicity when listening to native American English utterance.

Specifically, if the IAT reveals that non-native English listeners in the current study have an *implicit* association "American = Caucasian" then they may rate the same native American English utterances as *more* accented and *less* comprehensible

[1]While de Weers (2019) used the same matched-guise design as in the current study, Babel and Mellesmoen (2019) used videos of the actual speakers (that is, they did not use a guise).

when presented with an Asian face than when presented with a Caucasian face. Their intelligibility scores may also be *lower* for the Asian guise than for the Caucasian guise. Furthermore, this effect may be potentially stronger for the picture stimuli than for the video stimuli (Zheng & Samuel, 2017) as pictures bring *demand characteristics*, that is a situation where participants guess the concept of an experiment and act accordingly — rate the native stimuli in Asian face condition as, for instance, more accented (see McCambridge, de Bruin, & Witton, 2012; Rosenthal & and, 2009).

If, on the other hand, native Japanese listeners do not show any strong "American = Caucasian" association (that is, they regard both East Asian-looking speakers and Caucasian-looking speaker as equally American), then there should be no effect of speaker's face regardless of whether the native American English speech is presented with an East Asian guise or a Caucasian guise. The next chapter first investigates this implicit bias and then describes the main perception experiment.

Chapter 5

Experiments

This chapter describes two experiments used to evaluate the effect of an ethnic bias on the perception of native American English speech by native Japanese listeners. The perception tasks (section 5.2) were performed first. Japanese participants were asked to rate utterances recorded by native English speakers from the United States for their accentedness and comprehensibility. They were also asked to transcribe the utterances in order to measure their intelligibility. The whole experiment lasted about 2 hours. Upon completing the main experiment, each participant was offered a short break. After the break, the participants were asked to perform the Implicit Association Test (IAT), described in section 5.1, which took about 5 minutes. Since the IAT was meant to verify one of the assumptions of the current study, its results will be presented first. However, it was done last by the participants in order to make sure that doing the IAT would not, in any way, influence the participants' performance on the perception tasks.

5.1 Implicit Association Test

Chapter 4 provided evidence that the speech perception process can be affected by the listener's beliefs or stereotypes about the speaker. Experienced-based models predict that someone who exhibits a *strong* "American = Caucasian" association is more likely to assess native English utterance as being accented when presented with an Asia face than someone who has *weak* or none "American = Caucasian" association. Similarly, someone with a *stronger* "American = Caucasian" association is more likely to transcribed native English utterances presented with an Asian face less accurately than someone how has *weak* or none "American = Caucasian" association. Hence, in order to better analyze the effect of speaker's perceived ethnicity on speech perception by non-native listeners one would have to make sure those listeners *implicitly* associate being native English speaker (in this case American) with being Caucasian, that is they showed a *strong* "American = Caucasian" association.

This kind of association (i.e., "American = Caucasian") is in its principle identical to the one described in Devos and Banaji (2005). Devos and Banaji used an Implicit Association Test (IAT) in order to test the strength of "American = White" association of three groups of undergraduate Yale students. One group consisted of Caucasian American, one group consisted of Asian Americans, and one group consisted of African Americans. A small number of Asian and African participants was not born in the United States, however, this factor did not affect the results. Devos and Banaji found that all three groups showed a strong "American = White" association suggesting that even Asian Americans asthat all three groups showed a strong "American = White" association suggesting thatsociated the concept of being American with being Caucasian significantly stronger than being Asian or African American. In this study, I adopted their methodology and implemented a similar IAT in order to evaluate the strength of

implicit "American = Caucasian" association of Japanese listeners who will later perform the accentedness and comprehensibility rating, as well as transcription task of native American English utterances.

5.1.1 Rationale for the IAT

An Implicit Association Test (IAT) is a common method to measure the strength of *implicit* associations between some concepts (e.g., "republicans" or "democrats") and attributes (e.g., "positive" emotions or "negative" emotions) (Greenwald, McGhee, & Schwartz, 1998). An individual may, for instance, hold a positive attitude towards republicans and negative attitude towards democrats (or vice versa). However, if a traditional questionnaire is employed, the individual may not report this bias either because they are unwilling to do it or because they are simply unaware of it (Greenwald, McGhee, & Schwartz, 1998).

The IAT is commonly used in psychological research in order to assess a wide range of biases. For instance, the `gender-career` IAT measures whether an individual implicitly associate being *male* with *career* and being *female* with *family*, and the `weight` IAT measures whether an individual has preferences toward thin people relative to obese people. In fact, the popularity of IAT in psychology and later in other fields led to founding in 1998 `Project Implicit`, a non-profit organization led by researchers from Harvard University, University of Washington, and the University of Virginia, which is dedicated to online data collection from various IATs (Greenwald, Banaji, & Nosek, 1998). At the time of writing this thesis, `Project Implicit` features 14 different IATs such as Religion IAT, Age IAT, or Weapon IAT each of which is testing some kind of implicit associations or bias [1].

[1] It is important to stress here that this *bias* does not always refer to a *negative bias*. It simply shows a way in which an individual may evaluate certain concepts.

The IAT has also been gaining popularity in linguistic research where it is used not only as the main research tool (e.g., Babel & Russell, 2015) but also as an alternative to a survey designed to measure participant's implicit bias prior to the experiment (e.g., McGowan, 2011).

The IAT measures how strong is a relation between two concepts, such as "Caucasian" or "Asian", and two attributes, such as "American" or "foreign." Both, concepts and attributes, can be presented as visual cues (pictures), as written words, or as audio files (e.g., Pantos & Perkins, 2012) in a discrimination task. An individual can either have preferences towards "American = Caucasian" pairing, "American = Asian" pairing, or no preferences (no implicit bias) towards any of pairings suggesting that both Asian-looking and Caucasian-looking speakers are regarded as equally American. While traditional ethnic IATs would investigate the implicit "American = Caucasian" bias of native American English speakers (Devos & Banaji, 2005; Yi et al., 2013) the current study investigates whether such a bias would be present for non-native English speakers. In order to do so, the current study follows the procedure described in Devos and Banaji (2005) and replicated later in Yi et al. (2013) and in (Yi et al., 2014) with some necessary modifications. The participants in this study were native Japanese speakers for whom the word "foreign" (外国の *gaikoku no* in Japanese) is very often associated with "not Japanese" rather than the intended, "not American". Therefore, the name of this category was changed to "Japanese" in order to contrast with something being "American." As a result, the target concepts in this study - "Caucasian" (白人 *hakujin*) and "Asian" (アジア人 *ajiajin*) - were contrasted with attributes - "American" (アメリカの *Amerika no*) and "Japanese"(日本の *Nihon no*).

The concepts in the current study were presented as black and white pictures of East Asian-looking and Caucasian-looking people, while the attributes were mostly

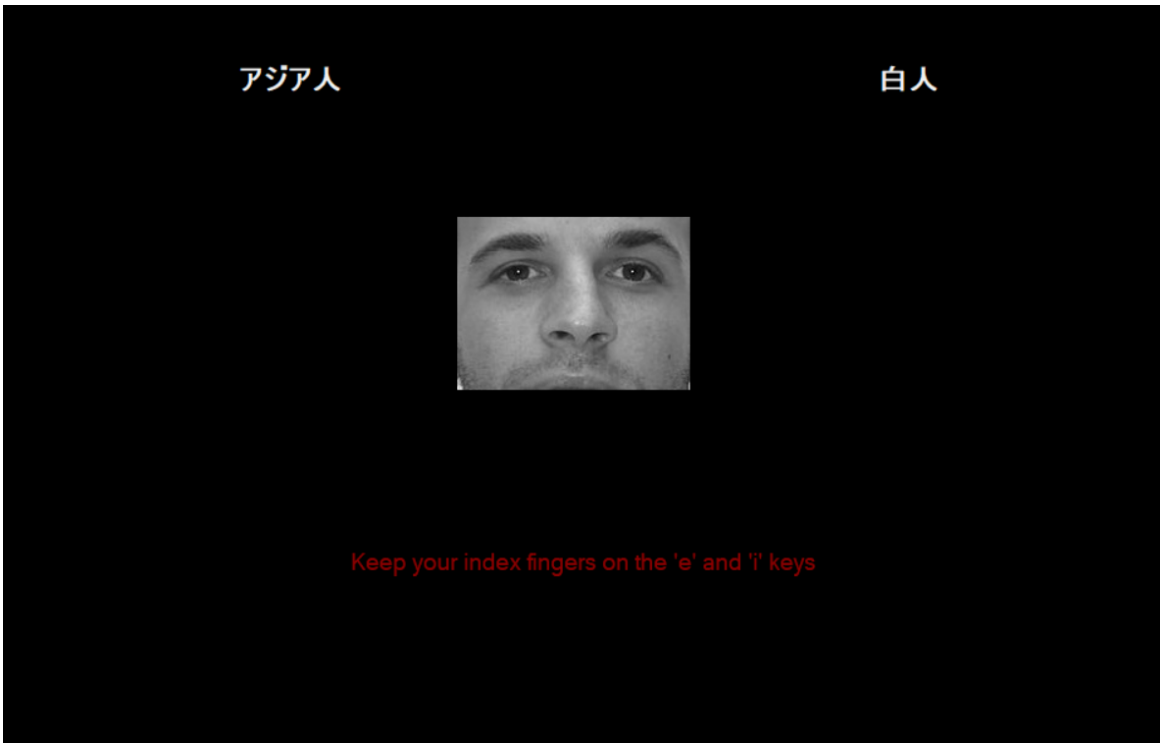
names of places (such as Tokyo) and symbols (such as $) relating to either something being American or Japanese. They were all written in the Latin alphabet as using the Japanese writing system would mean that all "American" symbols and places would be written in *katakana* (Japanese syllabary used for foreign words and names) while all "Japanese" symbols would be written in *hiragana* and Chinese characters.

In a typical IAT the stimuli are presented in a series of discrimination tasks (usually 5), which are used to compute the implicit association effect (often referred to as the D score) based on the response time, not on the number of correct and incorrect answers (Greenwald, McGhee, & Schwartz, 1998). If a participant makes a mistake, feedback is provided, and the person had to press the correct key before moving forward. The IAT used in the current study included a series of 5 tasks:

(1) Initial target-concept discrimination: Participants were asked to categorize a set of pictures belonging to one of the two target concepts, "Caucasian" or "Asian" like in Figure 5.1. They were asked to choose to which group the item (picture) that is currently being displayed on the screen belongs by pressing the key associated with the given concept. For example, in Figure 5.1 if the participant pressed the "`i`" key corresponding to the concept of "Caucasian" he or she would receive feedback that their response was correct by moving to the next trial. All other keys are blocked to avoid registering any accidental key strokes. This task was used to familiarize participants with the stimuli and the procedure.

(2) Associated attribute discrimination: Participants were asked to categorize a set of words belonging to one of the two attributes, "American" or "Japanese", such as names of American and Japanese places or currencies. The words were presented in English to avoid using Japanese orthography where all American attributes would be

Figure 5.1: IAT Step 1: Initial Target-Concept Discrimination.



Note: The label on the left side, associated with the `e` key, says "Asian" (アジア人), the label on the right side, associated with the `i` key, says "Caucasian" (白人). Here the participant is supposed to press the `i` key associated with the label "Caucasian."

written in a syllabary for foreign words – *katakana* while Japanese attributes would be written in another syllabary – *hiragana* and Chinese characters (see Figure 5.2). Participants were asked to choose which attribute, American or Japanese, is currently displayed on the screen by pressing the key associated with that given attribute. For example, in Figure 5.2, if the participant pressed the "`e`" corresponding to the attribute "Japanese", he or she would receive feedback that his or her response was correct by moving to the next trial. Similarly to the previous task, this task was used to familiarize participants with the stimuli and the procedure.

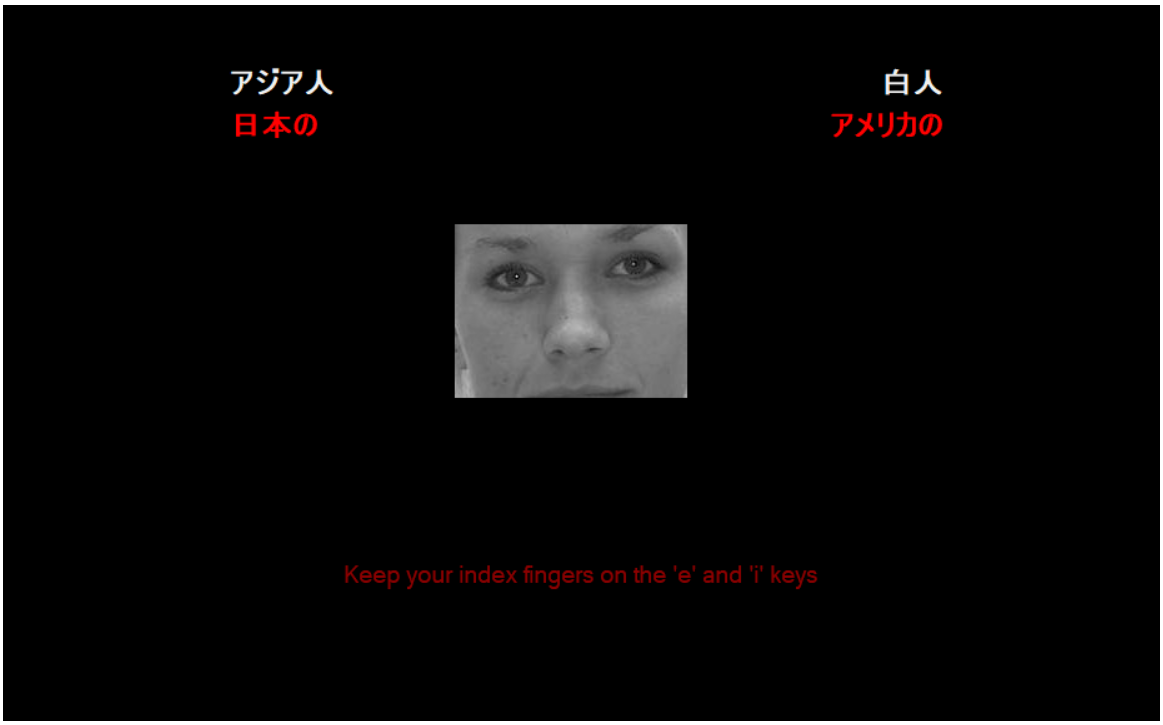
(3) Initial combined task: Participants were asked to categorize a set of pictures belonging to one of two target concepts, "Caucasian" or "Asian" or words belonging to

Figure 5.2: IAT Step 2: Associated Attribute Discrimination.

Note: The label on the left side, associated with the `e` key, says "Japanese" (日本の), the label on the right side, associated with the `i` key, says "American" (アメリカの). Here the participant is supposed to press the `e` key associated with the label "Japanese"

one of the target two attributes, "American" or "Japanese." The items were appearing on the screen one at a time and the participant had to choose which one is currently displayed on the screen by pressing the key associated with that concept or attribute (see Figure 5.3). The same key on the keyboard was linked to a congruent concept-attribute pair (i.e., "Caucasian" and "American"). Similarly, the other congruent pair (i.e., "Asian" and "Japanese") shares one response key. Results from this part *are* used when computing the final score. The idea behind this task is that it is *easier* to answer when closely related items share the same response key. Hence, participants with stronger "American = Caucasian" association are expected to answer *faster* in this task than in the (5) Reversed combined task, where the items are presented in the incongruent pairing.

Figure 5.3: IAT Step 3: Initial Combined Task.



Note: The labels on the left side, associated with the `e` key, say "Asian" (アジア人) and "Japanese" (日本の), the labels on the right side, associated with the `i` key, say "Caucasian" (白人) and "American" (アメリカの). Here the participant is supposed to press the `e` key associated with the label "Caucasian."

(4) Reversed target-concept discrimination: Participants had to categorize a set of pictures belonging to one of two target concepts, "Caucasian" or "Asian" (see Figure 5.4). This task was essentially *the same* as task (1), however, the keys associated with target-concepts were reversed (the labels appeared on reversed sides of the screen). This task was used to ensure that the participant will not simply associate one key (or one side) with one concept and another key (or another side) with the other one. Therefore the results from this part were *not* used in the final computations.

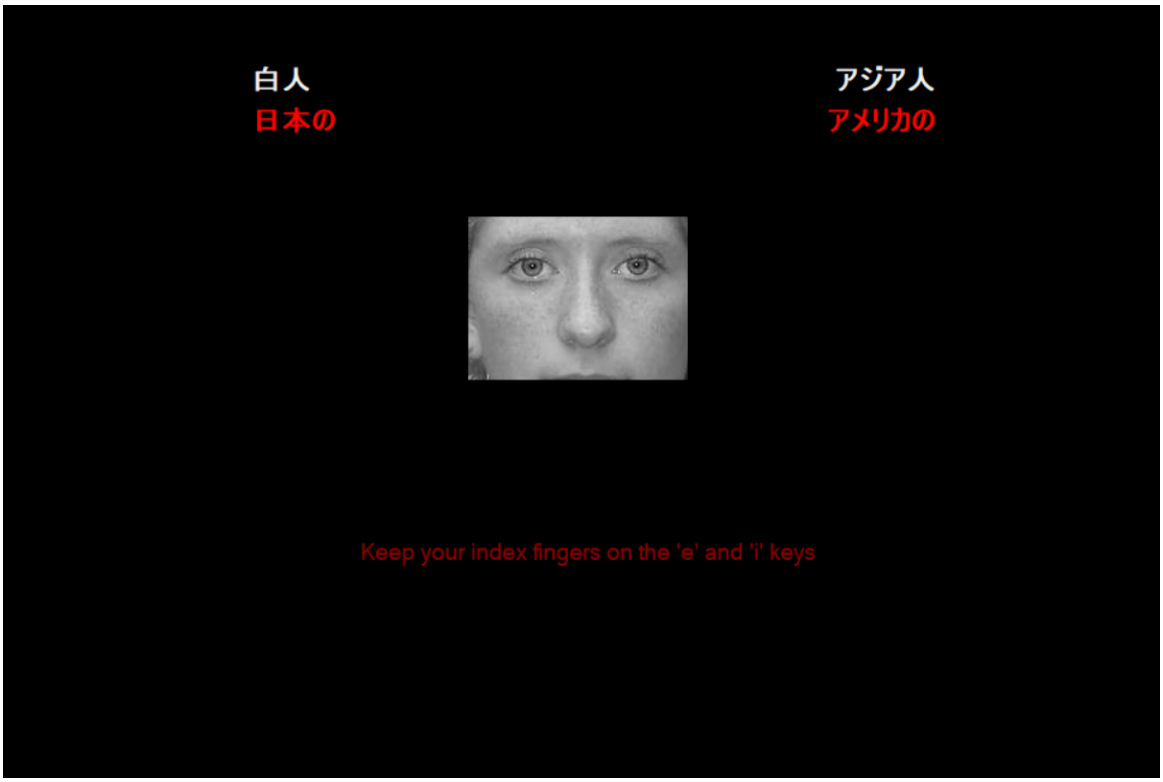
(5) Reversed combined task: Participants have to categorize a set of pictures or

Figure 5.4: IAT Step 4: Reversed Target-Concept Discrimination.

Note: The label on the left side, associated with the `e` key, says "Caucasian" (白人), the label on the right side, associated with the `i` key, says "Asian" (アジア人). Here the participant is supposed to press the `e` key associated with the label "Caucasian."

words presented one at a time, which belong to one of two target concepts such as "Caucasian" or "Asian" or one of two attributes such as "American" or "Japanese" (see Figure 5.5). This task was very similar to task (3), however, the combination of concepts and attributes was reversed. The same key on the keyboard was linked to an incongruent concept-attribute pair (i.e., "Asian" and "American"). Similarly, the other incongruent pair (i.e., "Caucasian" and "Japanese") shared one response key. Results from this part *were* used when computing the final score. Just as in (3) Initial combined task, the idea is that when closely related items share the same key (congruent pairing) participants will answer more quickly than in the incongruent pairing like in this task. The core of IAT is the *difference* in the response time between

Figure 5.5: IAT Step 5: Reversed Combined Task.



Note: The labels on the left side, associated with the `e` key, say "Caucasian" (白人) and "Japanese" (日本の), the labels on the right side, associated with the `i` key, say "Asian" (アジアの) and "American" (アメリカの). Here the participant is supposed to press the `e` key associated with the label "Caucasian."

task 3 and task 5.

The strength of implicit association, or the D score, is measured using the response time from task 3 (initial combination task) and task 5 (reversed combination task) (Greenwald et al., 2003). Tasks 1, 2, and 4, where *only* concepts (or *only* attributes) are displayed, are treated as distractors which main purpose is to familiarize participants with the stimuli and the tasks. Thus, they are *not* used to compute the D score. The D score can range from -2 to 2, where a positive score implies preferences for the congruent pairing presented in task 3 while the negative score indicates preferences for the incongruent pairing presented in task 5. The main

idea behind IAT is that participants will respond faster when items which are for them closely related share the same response key (congruent condition). This effect will not be present (D score = 0) if there is no bias (no preference) towards either of pairings. Nosek et al. (2002) proposed an absolute D score of 0.15, 0.35, and 0.65 as lower boundaries for weak, moderate and strong preferences respectively.

5.1.2 Methods

Participants

The (non-native) participants in this experiment were 80 native speakers of Japanese (40 males and 40 females) recruited mainly, but not exclusively, among the students of the University of Tokyo. They ranged in age between 18 and 35 years old, with a mean of 22.5 (SD = 4.42). All but four participants self-assessed their overall English level as lower intermediate (CEFR B1) or higher. The four participants who assessed their English proficiency as beginner all passed Eiken Level 2 or got at least 550 points on the TOEIC test; thus, should be considered at an intermediate level. None of the participants reported any hearing or vision impairments. Moreover, none of the participants stayed or lived abroad for a period longer than 1 year, with a mean of 2.8 months (SD = 4.12 months). This research was approved by the ethics committee of the University of Tokyo. All participants received monetary compensation for participating in the experiment.

Stimuli

The current experiment was presented using FreeIAT 1.3.3 software (Meade, 2009). One specification of the FreeIAT is that it does not allow for both concepts and attributes to be presented as pictures. The software allows only for one of these two categories, either concepts or attributes, to be displayed as pictures, while the

other one has to be presented as words. While this may be viewed as a possible limitation, FreeIAT, unlike other available IAT tests, runs on a local machine (not on a server). Therefore it does not rely on the internet speed or the stability of internet connection. This can be crucial, given that the IAT score is being computed, taking into consideration the response times (RT) as measured in milliseconds. However, this particular feature of FreeIAT meant that either concepts or attributes had to be presented as words. Since encoding ethnicity (concepts) as words-only stimuli would be far from feasible, the concepts were presented as pictures ("Asian" faces and "Caucasian" faces), while the attributes were displayed as words related to places or symbols that were either "Japanese" or "American."

Pictures Used for Concepts

Similarly as in Devos and Banaji (2005), 20 pictures (10 Caucasian faces and 10 East Asian faces) were chosen to represent the two ethnic groups (full set used for the concepts can be found in Appendix A). Half of the pictures in each group were female faces. The pictures used in this experiment were chosen from The Chicago Face Database (Ma et al., 2015) as well as from the internet by filtering the search results for the Creative Commons (CC) license. All pictures were black-and-white with a neutral expression. The size of each picture was adjusted to about 190 x 140 pixels (the actual size differed slightly depending on the face shape). Moreover, in order to possibly minimize the influence of other variables, each face was cropped so that the hair was not fully visible - a practice common in IAT research (Figure 5.6).

Words Used for Attributes

To represent the concept of "American" 8 symbols and places were chosen: *$*, *the U.S.*, *The Statue of Liberty*, *The White House*, *Thanksgiving*, *Washington D.C.*, *Los*

Figure 5.6: Examples of male and female faces (Caucasian) after preprocessing.

Angeles, *New York*. These were matched with 8 symbols and places related to the concept of "Japanese": *Meiji Shrine*, *Asakusa*, *Tokyo Tower*, *Hello Kitty*, *Kyoto*, ¥, *Obon*, *Itsukushima*. All the attributes listed here were presented in written form (not as pictures). They were all written in Latin alphabet in order to avoid mixing two different syllabaries (*hiragana* and *katakana*) and Chinese characters — all commonly used in Japanese writing system[2]. Although the attributes were presented as words not pictures, the total number of items as well as their type roughly matched these used in Devos and Banaji (2005).

Procedure

As stated in the previous section the experiment was designed using the FreeIAT (Meade, 2009), an open source software which computes the D score based on an improved IAT algorithm described in details in Greenwald et al. (2003). FreeIAT is highly customizable, allowing the user to control the number of stimuli along with the number of trials in each task. In addition, FreeIAT provides feedback - a red X - which appears on the screen if the wrong key was chosen (Figure 5.7). Participants need then to press the correct key before proceeding to the next trial.

[2]Using Japanese writing system would mean that all "American" symbols would have been written in *katakana* while all "Japanese" symbols would have been written in *hiragana* and Chinese characters.

Figure 5.7: Example of IAT "Asian" vs "Caucasian" category with a feedback (picture on the right).

The RT is recorded for *each* trial. If the wrong key is chosen the RT denotes the time from when the stimuli was presented to the time when the right answer was provided. This means that making a mistake does not "stop" the timer. The software then eliminates trials for which the RT was higher than 10,000 ms and does not compute the D score for those participants whose RT was less than 300 ms for more than 10% of the trials. This is a standard procedure as extremely slow responses may indicate momentary inattention while extremely fast responses are usually initiated prior to perceiving the stimulus (Greenwald et al., 2003).

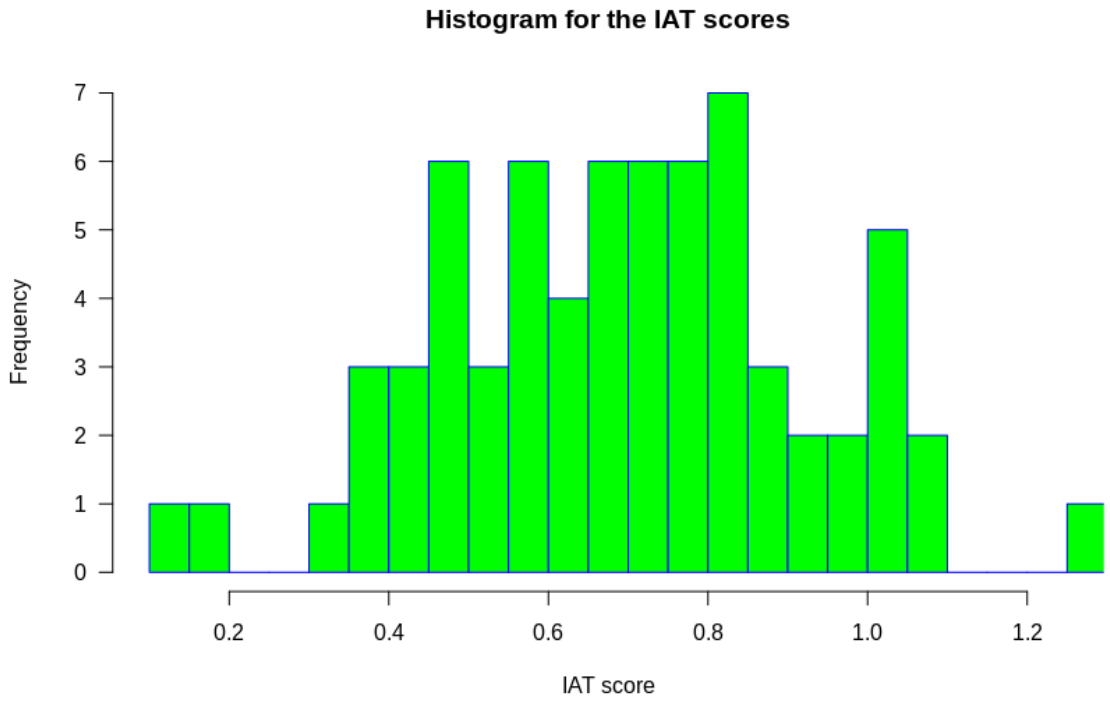
Each participant received written instructions (see Appendix B) accompanied with a detailed oral explanation delivered in Japanese and was given a chance to ask questions. They were also presented with the stimuli prior to the experiment. All of the participants stated that they had no problem identifying the items in each category (i.e., they could easily distinguish between Asian and Caucasian faces, as well as Japanese and American symbols or landmarks). They were instructed to look directly at the screen and to answer as fast as possible without sacrificing accuracy. The answers were provided by pressing either the `e` key if the item belonged to the category on the left or the `i` key if the item belonged to the category on the right. The stimuli were presented in 5 blocks (tasks), each of which including 42 trials. Only RTs from tasks 3 and 5 were used to compute the D score. Trials in tasks 1, 2, and 4

were considered practice sessions and were the only trials excluded from the analyses. The experiment lasted about 5 minutes per participant.

5.1.3 Results and Discussion

The D score (IAT effect) for each participant was automatically computed by the FreeIAT software using the scoring algorithm introduced in Greenwald, McGhee, and Schwartz (1998). All trials with the response time higher than 10,000 ms were eliminated. Two participants were excluded from the analysis due to having latency lower than 300 ms for more than 10% of the trials.

Figure 5.8: Histogram of the D scores (IAT) for 78 participants. Absolute values of 0.65, 0.35, and 0.15 are usually treated as cutoff points for "strong," "moderate," and "weak" association (Nosek et al., 2002).



One sample t-test indicated that the D scores were significantly different from 0

($t(77) = 27.41$, $p < 0.001$, M=0.69, SD=0.22). Moreover, the IAT results revealed that a majority of Japanese participants responded faster in the congruent block than in the incongruent block. This means that participants associated more strongly the concept of being American with a Caucasian face than with an Asian face. This is also consistent with the results presented in Devos and Banaji (2005) where both, Asian and Caucasian native English speakers from North America demonstrated similar bias.

Figure 5.8 shows the distribution of D scores for 78 participants (data of 2 participants was excluded due to too fast response time). While all participants demonstrated preference towards the congruent pairing (all D scores > 0) for 46 participants this preference was strong (D score >= 0.65 with a max D score = 1.29), for 30 participants it was moderate (D score between 0.35 and 0.65), for 1 participant it was weak (D score between 0.15 and 0.35) and only 1 participant received a very low D score = 0.11.

The results described above suggest that 76 out of 78 of the native Japanese speakers who participated in this experiment have moderate to strong bias towards the "American = Caucasian" pairing. This means that they associate the concept of being American with being Caucasian, suggesting that they will be more likely to be affected by the speaker's ethnicity in the perception experiment. It is, without a doubt, much more difficult to measure this kind of bias for non-native speakers simply because labels used in Devos and Banaji (2005), like "foreigner," mean something different to native speakers and had to be replaced tentatively by "Japanese." However, the fact that *all* 78 participants received a positive D score with 46 participants showing very strong (D score > 0.65) preference towards the congruent pairing suggests that there is a tendency to associate a Caucasian face with the concept of being American. This also appears to be consistent with the claim made by Kubota and Fujimoto

(2013) that in Japan being a native English teacher is often conflated with being Caucasian. However, one important thing to remember is that while the Japanese listeners in this experiment may associated being Caucasian with being American it does not necessarily mean that they associate being *native English speaker* with only Caucasian face. It is possible, for instance, that they have some experience interacting with Asian-looking native speakers from countries other than the United States. This possibility will be discussed further in chapter 6.

Having confirmed that the Japanese participants in this study generally exhibit a moderate to strong bias toward the "Caucasian = American" pairing, it is questionable if this bias may also impact their speech perception. In order to evaluate this, a series of perceptual tasks were conducted and are reported in the next section. In these perceptual tasks, the same native Japanese speakers had to evaluate native American English utterances paired with either (1) a picture of an East Asian face, (2) a picture of a Caucasian face, (3) a video of an East Asian speaker, (4) a video of a Caucasian speaker, or (5) audio-only stimuli with no visual cues. In perceptual tasks participants were asked to rate the speakers for accentedness and comprehensibility. They were also asked to transcribe each utterance as a measure of intelligibility. These ratings of accentedness and comprehensibility, as well as intelligibility score, were then used in order to investigate whether there is a correlation between the listener's IAT scores and the three measures for Asian guise and Caucasian guise separately. For instance, it is possible that participants with higher D scores may rate the Asian guise as more accented or lesson comprehensible compared to participants with lower D scores. Similarly, they may even exhibit lower intelligibility than participants with lower D scores.

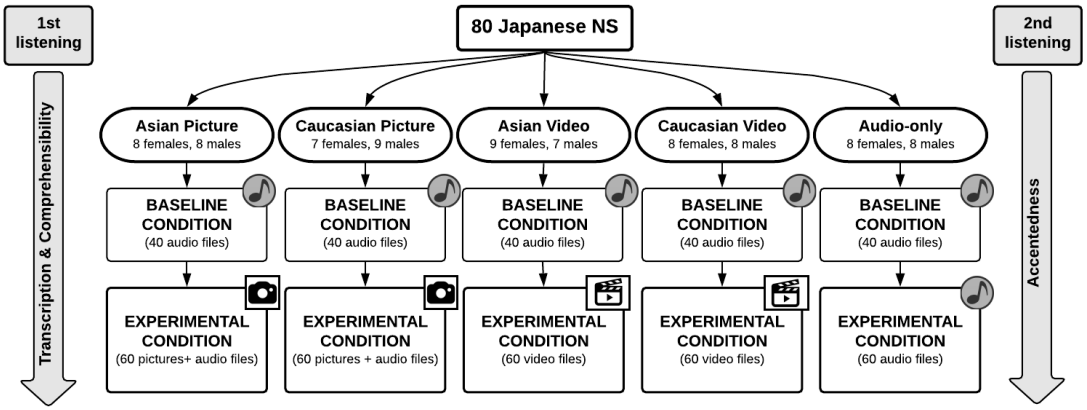
5.2 Perception Experiment: Accentedness and Comprehensibility Rating Task and Measurement of Intelligibility

5.2.1 Overview

As it was discussed in Chapter 4, there is a growing body of research exploring the role that socioindexical cues, such as ethnicity, gender, or age, play in speech perception. Previous studies indicate that seeing a picture of the speaker may affect speech perception for native English listeners. However, this effect may be smaller if, instead of a picture, the listeners will be presented with a video of the speaker (Zheng & Samuel, 2017). In order to investigate how the ethnicity of the speaker, as perceived from both pictures and videos, affects *non-native* English listeners, the same 80 native speakers of Japanese as in Section 5.1 were randomly assigned to one of five groups (between-subject design) where they completed three tasks in two different conditions (within-subject design). Figure 5.9 shows the general flow of the experiment. In the first condition, referred to as the *baseline* condition, participants in all groups listened to the same audio-only stimuli (40 native English utterances). In the second condition, referred to as the *experimental* condition, all participants were presented with the same audio stimuli (60 native English utterances) but combined with different visual cues for different groups (matched-guise design): pictures (Asian Picture group and Caucasian Picture group), videos (Asian Video group and Caucasian Video group), or no visual cues (Audio-only group). In order to complete all three tasks, participants in each group listened to the stimuli twice. During the first listening, they were asked (1) to perform a transcription task and (2) to rate the perceived comprehensibility of each

utterance on a 9-point Likert scale that ranged from 1 (totally incomprehensible) to 9 (fully comprehensible), as described in Derwing and Munro (2009). Once they had completed these tasks for all the stimuli, they were asked to listen again and (3) to rate each utterance for the accentedness on a 9-point Likert scale, going from 1 (strong foreign accent) to 9 (native accent) (Derwing & Munro, 2009).

Figure 5.9: The general design of the experiment. Eighty Japanese native speakers were divided into five groups. Each group completed two rating tasks and a transcription task in two conditions - the *baseline* condition and *experimental* condition).



5.2.2 Methods

Participants

Participants in this experiment were the same 80 native speakers of Japanese as described in section 5.1.2. Their responses were recorded on the same day as the experiment outlined in section 5.1. A short break was offered between the two experiments.

Stimuli

Stimulus materials for the current experiment consisted of 100 sentences or phrases produced by 10 native speakers of American English (5 males and 5 females). Four of the speakers were from California, one was from Washington, one from Virginia, one from Wisconsin, one from New Mexico, one from Massachusetts, and one from Texas (see Table 5.1), however, many of the speakers did not have the typical accent of their native region anymore. Forty of the recorded utterances (2 male and 2 female voices) were presented as audio-only stimuli in the *baseline* condition. The order of presentation was randomized for each participant, but the exact same stimuli were presented in all 5 groups. The remaining sixty utterances (3 male and 3 female voices) were presented aurally along with: (1) picture featuring an Asian face (Asian picture group), (2) picture featuring a Caucasian face (Caucasian picture group), (3) video featuring an Asian face (Asian video group), (4) video featuring a Caucasian face (Caucasian video group), (5) no visual cues (Audio-only group - control group). The following sections describe the preparation of the audio files (Auditory Stimuli), video files (Video Stimuli), and picture files (Picture Stimuli).

Table 5.1: List of native English speakers with the states they came from.

Speaker	**State**	**Speaker**	**State**
Female01	California	Male01	New Mexico
Female02	Seattle	Male02	California
Female03	Wisconsin	Male03	Virginia
Female04	Massachusetts	Male04	California
Female05	California	Male05	Texas

Auditory Stimuli

Ten native speakers of English from the United States (5 males and 5 females) were each presented with one of ten short picture stories. All the stories were similar to the

"Suitcase Story" (Derwing et al., 2004). The speakers were given time to familiarize themselves with their story and to ask questions about its content. Each of the native speakers was instructed to introduce himself or herself, speak about his or her day and then tell the story from the pictures. They were asked to speak naturally but to avoid complicated words whenever possible. All the samples were recorded in a soundproof booth using SONY ECM-MS957 microphone on a Lenovo IdealPad Y580 computer. About 5 minutes of audio was recorded by each speaker.

Ten sentences or phrases were extracted from each of the recordings using Praat (Boersma & Weenink, 2010). Each sample consisted of about 10-18 words, including prepositions and articles. Table 5.2 shows examples of 10 sentences (one for each talker). The transcription of all sentences can be found in Appendix E. The intensity was adjusted to 70 dB across all samples. Since the utterances were recorded in a soundproof booth, an echo effect was added using Adobe Premiere Pro CC 7.0 in order to make them sound more natural, especially when combined with video.

The final audio material consisted of 100 utterances, 10 for each of the recorded native speakers. Half of these were female voices. Two male and two female voices (40 utterances total) were chosen for the *baseline* condition, where the utterances were presented only as audio stimuli to all five groups. The remaining 60 utterances (3 female and 3 male voices) were presented in the *experimental* condition as audio-only stimuli to the control group and accompanied with visual cues (videos or pictures) to the remaining four groups.

Video Stimuli

Video stimuli were prepared using the 60 utterances chosen for the *experimental* condition. Each utterance was used to prepare two videos, one with an Asian-looking

Table 5.2: Examples of sentences recorded by individual talkers.

Talker	Transcription	Length
Female 01	*But then she gets home and I guess there is a sudden rainstorm.*	13 words
Female 02	*I think it's very clean and safe and everyone is very friendly.*	13 words
Female 03	*We have a car and it's racing down the road very very fast.*	14 words
Female 04	*I got a laptop from my mom and then I started doing some video editing.*	15 words
Female 05	*But it seems that everything she says just makes things worse.*	11 words
Male 01	*I see a man and a woman sitting at the table, talking.*	12 words
Male 02	*The man and the woman are now swimming inside the ocean and they seemed to have spotted something.*	18 words
Male 03	*This guy is not really in control of his dog.*	10 words
Male 04	*And the boy is standing next to him with his jacket over his head.*	14 words
Male 05	*He is a really nice man, so he's what we call gentleman.*	13 words

actor (for Asian video group) and one with a Caucasian-looking actor (for Caucasian video group). In order to keep the conditions as similar as possible for both video groups, neither of the actors was the original speaker.

All the recordings were made in front of a white wall with actors wearing white t-shirts. The videos were recorded using the LG G5 built-in camera in Australia or Sony HDR-CX405 camera in Japan. All but two actors were native speakers of English from either North America or Australia. The remaining two actors, one male and one female, were from Hong Kong and Vietnam, respectively. They both reported using English every day at school or work.

Before recording each sentence, speakers were given the corresponding script and were asked to practice it for a few minutes. They were also asked to listen a few

times to the original audio and to try saying it, in the same way, paying attention to all the pauses and intonation. Finally, the video recordings were made with the original audio running in the background. All speakers were asked to repeat the given utterance so that it was as close as possible to the original audio. On average, each of the speakers repeated each utterance about 40 to 50 times during one recording. The best matching attempt was then chosen, and the audio from the video recording was replaced with the given utterance from auditory stimuli using Adobe Premiere Pro CC 7.0. In this process, special attention was given to lip movements in order to ensure that all of the videos look as natural as possible. The 60 videos recorded with Asian looking actors were used in the *experimental* condition for the Asian video group while the 60 videos with Caucasian looking actors were used in the *experimental* condition for the Caucasian video group.

Picture Stimuli

Pictures of the same people recorded for the video stimuli were used to prepare the stimuli for Asian picture group and Caucasian picture group (3 male and 3 female faces per each group). All pictures were taken in the same setting as the videos, in front of a white wall and with the person wearing a white t-shirt (see Appendix D). The pictures were then combined together with the 60 audio files selected for the *experimental* condition using Adobe Premiere Pro CC 7.0 in a way to ensure that each picture will be shown on the screen at the same time as the audio stimuli. Each audio file was used twice: once with an Asian and once with a Caucasian face to create stimuli for two groups. The same face was matched with the same voice as for the video stimuli. The 60 stimuli with Asian faces were presented in *experimental* condition in the Asian picture group while the 60 stimuli with the Caucasian faces were used in *experimental* condition in the Caucasian picture group.

Procedure

The data in this experiment were collected at The University of Tokyo, Komaba Campus, in a soundproof booth (SoundLab) or in a quiet room. Prior to their arrival, participants were asked to fill in a questionnaire regarding their linguistic background and travel experience. After completing the questionnaire each of the participants was randomly assigned to one of 5 experimental groups. The gender variable was controlled for in order to ensure comparable numbers of males and females in each group.

Upon their arrival, the participants were asked to sign a consent form. All the instructions (verbal, in writing, and on the screen) were provided in Japanese (see Appendix C for the written instructions). Participants were instructed to always look at the computer screen when listening to the stimuli. They were also asked to take a break between the *baseline* and *experimental* conditions. Additional opportunities for breaks were provided between the utterances, which were separated by a screen with a "NEXT" button. Participants had to press this button in order to proceed to the next trial. They were encouraged to use these opportunities to take breaks whenever they felt like it. Prior to the task, each participant was presented with a short list of the most difficult words, which appear in the recordings and were selected based on author's teaching experience (*rainstorm, dozen, to go off, ingredients, confident, to pour, to sweat, damage, clown, tons of, stove, curled hair, mixing bowl, to drag, to signal, fortunate, underneath, pond, to spot something, environmental, encourage, to be concerned, to drop into*). They were given time to read the list and to ask questions about words they did not understand. They were also given a chance to adjust the sound to a comfortable level prior to the experiment. Participants listened to the stimuli on Toshiba Dynabook T752, Lenovo IdeaPad Y580, or Lenovo Yoga 910s laptop computers using BOSE AE2 headphones.

The experiment was designed using a simple PHP script embedded in an HTML file. Collected data were stored in a MySQL database on a server provided by FastComet. The experiment was preceded by a practice block of five audio-only utterances produced by a female native English speaker whose voice was not used in the experiment. Responses to the practice block were excluded from the analysis. After completing the practice block, participants moved on to the main experiment, where they were presented with 100 utterances, 40 in the *baseline* and 60 in the *experimental* condition. The utterances were presented in a randomized order in both conditions. A cross was displayed on the screen for a short time with an audio prompt (a "beep" sound) before each utterance to direct the listener's attention to the screen. Half of the participants in each group (4 males and 4 females) were presented with the *baseline* condition prior to the *experimental* condition. The participants listened to all utterances twice in order to complete the transcription and rating tasks.

FIRST LISTENING. The first time participants listened to the utterances; they performed the comprehensibility rating task and the transcription task for both the *baseline* and *experimental* condition. They were instructed to rate the perceived comprehensibility on a 9-point Likert scale by simply choosing a number between 1 (very easy to understand) and 9 (very difficult to understand), and to transcribe as much of the utterance as they could by typing the words into a textbox on the screen. The transcription task was then used to compute the intelligibility score based on criteria explained in section 5.2.3. After listening to all 100 utterances and performing the first grading task and transcription task, the participants were asked to take a short break.

SECOND LISTENING. The second time participants listened to the utterances; they performed the accentedness rating task. Similarly to

comprehensibility, accentedness was rated on a 9-point Likert scale by choosing a number between 1 (strong foreign accent) and 9 (native speaker). The whole experiment, including the instructions, filling in the questionnaire, and breaks, lasted for about 2 hours.

5.2.3 Results and Discussion

Prior to the analysis, the ratings were checked for internal consistency using Cronbach's alpha. Cronbach's alpha is widely used and reported as evidence of reliable scoring in numerous pronunciation studies (Isaacs & Thomson, 2013). The Cronbach's alpha was computed with the *alpha* function (Revelle, 2018). The computations were carried out separately for (1) the *baseline* condition (all groups together) and (2) for each group in the *experimental* condition. The internal consistency for the *baseline* condition was checked for all groups collectively since participants in each group performed the exact same task with the same stimuli and with no additional cues. On the other hand, in the *experimental* condition, participants were rating the same audio stimuli but with different visual cues, therefore the values were computed separately for each group.

Cronbach's alpha obtained for accentedness ratings ranged from 0.96 to 0.98, which is comparable to native English raters (Derwing et al., 2004; Isaacs & Trofimovich, 2011). The values for comprehensibility were even higher ranging between 0.97 and 0.99 and were also comparable to the values obtained for native English raters. These results suggest that the ratings were highly consistent and that non-native listeners in this experiment were rating accentedness and comprehensibility of native English utterances with a consistency similar to that of native English listeners.

Accentedness

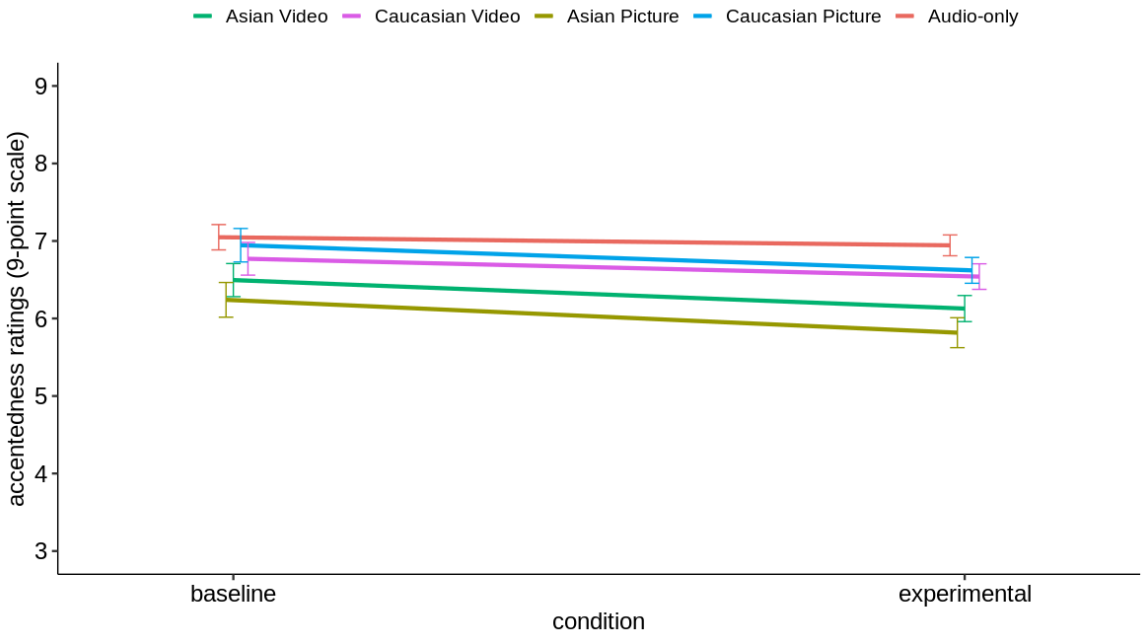
The accentedness ratings were collected on a 9-point Likert scale ranging from 1 - very strong non-native accent to 9 - native speaker's accent. The 9-point Likert scale was demonstrated to be the most appropriate for evaluating the accentedness of an utterance (Derwing & Munro, 2009). In order to investigate the effect of speaker's perceived ethnicity on these ratings, the raw data were first plotted for visual inspection. Figure 5.10 does not indicate any apparent differences between the groups in neither the *baseline* nor the *experimental* condition. In fact, the means of individual groups differ only by 0.1 to 0.8 points in the *baseline* condition and 0.08 to 1.13 points in the *experimental* condition (see Table 5.3). Overall, participants in all groups assigned slightly lower ratings to the speakers in the *experimental* condition compared to the *baseline* condition (about 0.1-0.42 points difference). This, however, could be due to some characteristics of individual speakers, i.e., some speakers just sound *less* native-like compared to other speakers.

Table 5.3: Mean, median and standard deviation of the accentedness ratings.

Condition	Group	N	Mean	St. Dev.	Median
Baseline	Asian Picture	640	6.24	2.30	7
	Asian Video	640	6.50	2.17	7
	Caucasian Picture	640	6.95	2.27	8
	Caucasian Video	640	6.77	2.09	7
	Audio Only	640	7.05	1.93	7
Experimental	Asian Picture	960	5.82	2.499	6
	Asian Video	960	6.13	2.28	7
	Caucasian Picture	960	6.62	2.14	7
	Caucasian Video	960	6.54	2.19	7
	Audio Only	960	6.94	1.92	7

The data were also plotted by the gender of the native English speaker to evaluate the tendencies for male and female speakers separately (see Figure 5.11).

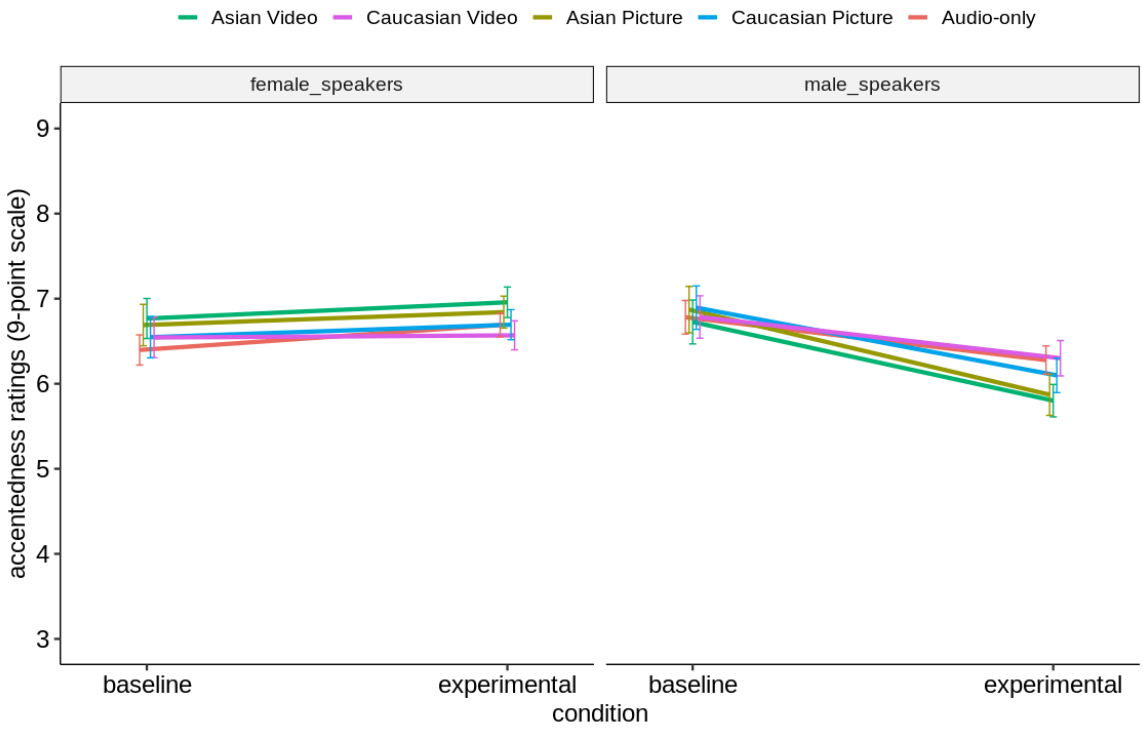
Figure 5.10: The mean accentedness ratings for each group in the *baseline* and *experimental* condition with 95% confidence interval. Lower ratings indicate that the utterance was perceived as more accented.



The ratings in the *baseline* condition seem to be comparable across the groups for both male and female speakers (about 0.08-0.71 points difference between the groups for female speakers and 0.12-0.91 points between the groups for male speakers). Likewise, ratings in the *experimental* condition do not seem to be very far apart (roughly 0.03-0.85 points difference for female speakers and 0.08-0.41 points for male speakers). Overall, male speakers were rated as more accented in the *experimental* condition while female speakers were rated slightly less accented in the *experimental* condition when compared to the *baseline* condition.

Looking further at individual speakers revealed some other patterns in the accentedness ratings (Figure 5.12). While female speakers seem to be receiving relatively similar ratings with a greater variance in the *experimental* condition (especially for `female04`), ratings of individual male speakers differ to a greater

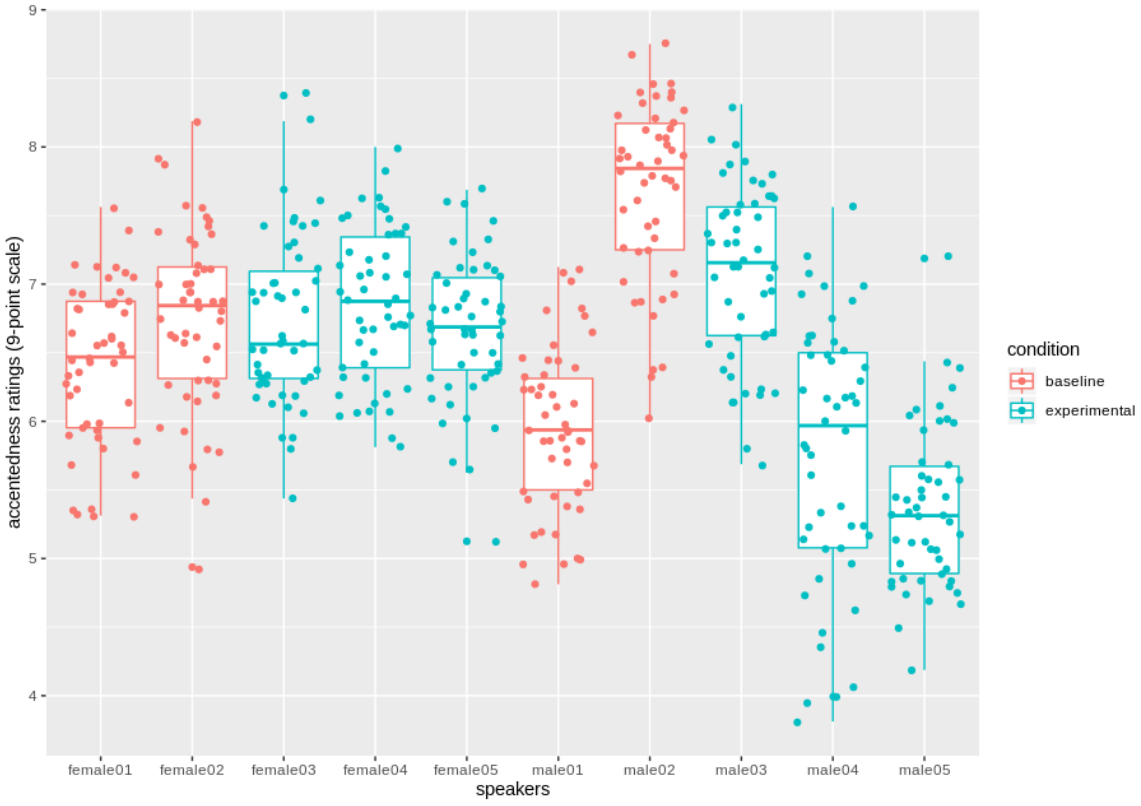
Figure 5.11: The mean accentedness ratings for each group in the *baseline* and *experimental* condition by speaker's gender with 95% confidence interval. Lower ratings indicate that the utterance was perceived as more accented.



extent. For instance, `male02` was rated exceptionally high (closer to a native speaker), which accounted for the high mean ratings for males in the *baseline* condition compared to the *experimental* condition. This remains true when we separate the ratings by group, although in the Audio-only group male speakers seem to be rated in a more comparable way than in the other groups.

In order to investigate whether the small differences between the groups are indeed not significant (i.e., there is no effect of speaker's perceived ethnicity) the data were analyzed in `R language` (R Core Team, 2019) using a linear mixed-effects model implemented with the *lmer* function from *lme4* package (Bates et al., 2015). Linear mixed-effects models were demonstrated to be more robust for this kind of data compared to the traditional ANOVA since they allow for random intercept for

Figure 5.12: Boxplots of the accentedness ratings received by individual speakers. Higher ratings indicate that the speaker sounded more like a native speaker.



each item (here each audio file) and each subject (Jaeger, 2008; Sheppard et al., 2017).

In order to fit a linear mixed-effects model, the response variable should be continuous rather than discrete. In the current research, accentedness ratings were collected on a 9-point Likert scale. However, a previous study comparing accentedness ratings performed on a 7-point Likert scale and the DME (Direct Magnitude Estimation) successfully demonstrated that listeners treat accentedness as a continuous scale (M. Helen Southwood, James E. Flege, 1999). The same results were then reproduced using the DME and a 9-point Likert scale (Munro, 2018). Furthermore, linear mixed effect models were shown to work well with a 9-point Likert scale (Sheppard et al., 2017). Therefore, the model was fitted and

evaluated using the raw data.

The full model included (1) `group` (a factor with 5 levels), (2) `condition` (*baseline* and *experimental*), (3) `gender` (of the native English speaker) and the interaction between these variables all modeled as fixed effects (see Appendix F). Self-reported English level (a factor with 6 levels) was also added to the model as a fixed effect in order to assess whether the accentedness ratings were affected by the English proficiency of the Japanese participants. Intercepts for `participant` and `item`[3] (each audio file) were included as random effects. Furthermore, following the guidelines presented in Barr (2013) and in Heisig and Schaeffer (2018), by-participant random slopes for the effect of `condition`, `gender` (of the speaker) and their interaction were also added to the model. The maximal model was justified by model comparison. Including random intercepts for `participant` and `item` allowed to model some individual differences between the subjects and between the items. Similarly, including by-participant random slopes for the effect of `gender` (of the speaker) and `condition`, allowed to account for the degree to which these two variables affected individual listeners and to avoid type 1 errors with intercept-only models (Barr, 2013). Finally, a vector containing the number of words in each sentence was set as weights.

In order to investigate whether the `English_level` variable is an important predictor of accentedness ratings, the full model was compared to the same model but *without* the self-reported English level using the Likelihood Ratio Test. The results of this comparison indicated that the English level did not affect the accentedness ratings, therefore, it was excluded from the further analysis ($X^2(5) = 6.31, p = 0.28$)

Visual inspection of the residual plots of the final model (without the

[3]The `item` was embedded in the `speaker`.

`English_level`) did not reveal any apparent heteroscedasticity or deviation from normality. The model was also checked for the potential influential data points, which may affect the outcome in the same way as outliers in the traditional ANOVA (Winter, 2013). These points were identified using the *influence* function. Cook's distance was computed using the *cooks.distance* function from *influence.ME* package (Nieuwenhuis et al., 2012). A new data set was created using 3 times the overall mean as the cutoff value, which resulted in eliminating 565 out of 8000 data points. Further analysis was conducted for both models - with and without the influential data points. There was no difference between these two models regarding the significance of fixed effects. Therefore the p-values reported below are the values obtained from the original model fitted with the full dataset.

One important characteristic of the linear mixed effect models, as implemented with *lmer* function, is that they do not provide p-values. Although it is possible to obtain p-values using only the Likelihood Ratio Test, with triple interaction, this would require fitting multiple models. Moreover, the Likelihood Ratio Test tends to be quite anti-conservative and sometimes yields very small p-values requiring some other additional forms of validation (Luke, 2016). Therefore, the p-values were obtained using the *anova* function from *lmerTest* package (Kuznetsova et al., 2017), which provides the Satterthwaite approximation of degrees of freedom.

Table 5.4 shows the results of the *anova* function applied to the fitted model. The effect size was calculated using the *r.squaredGLMM* function from the *MuMIn* package (Barton, 2018). Since mixed effects models include random effects two R^2 were calculated following the recommendation in Nakagawa et al. (2017). The *marginal* R^2, which includes only variance of fixed effects, was R^2=0.02 while the *conditional* R^2, which includes variance of both random and fixed effects, was R^2=0.14. The `group` x `condition` interaction was not significant ($F(4, 75) = 0.37$,

Table 5.4: Type III Analysis of Variance Table with Satterthwaite's method for the effect of speaker's perceived ethnicity on the accentedness ratings.

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
group	150.20	37.55	4	70.16	2.14	0.09
condition	6.33	6.33	1	6.43	0.36	0.57
gender	3.73	3.73	1	6.41	0.21	0.66
English_level	105.26	21.05	5	70	1.20	0.32
group:condition	26.03	6.51	4	75	0.37	0.83
group:gender	110.21	27.55	4	75	1.57	0.19
condition:gender	16.06	16.06	1	6.33	0.92	0.37
group:condition:gender	29.06	7.26	4	75	0.41	0.80

$p > 0.05$) suggesting that speaker's perceived ethnicity had no effect on the accentedness ratings. That was true even after looking at male and female speakers separately meaning that the perceived ethnicity had no effect on accentedness ratings of both male and female speakers ($F(4,\ 75) = 0.41,\ p > 0.05$). No other effect was significant (all $p > 0.05$).

Comprehensibility

The comprehensibility, similarly to the accentedness, was rated on a 9-point Likert scale ranging from 1 - easy to understand, to 9 - difficult to understand. Visual inspection of the plotted data did not indicate any apparent differences between the groups (Figure 5.13). On the contrary, participants in all groups were, on average, rating the comprehensibility of presented utterances in a similar way in both conditions with only 0.1-0.81 points difference between the groups in the *baseline* condition and about 0.08-0.13 points difference between the groups in the *experimental* condition (Table 5.3). Moreover, Figure 5.13 reveals no changes in the ratings between the *baseline* and *experimental* condition for none of the groups. This suggests that there was no effect of the speaker's perceived ethnicity on the comprehensibility ratings.

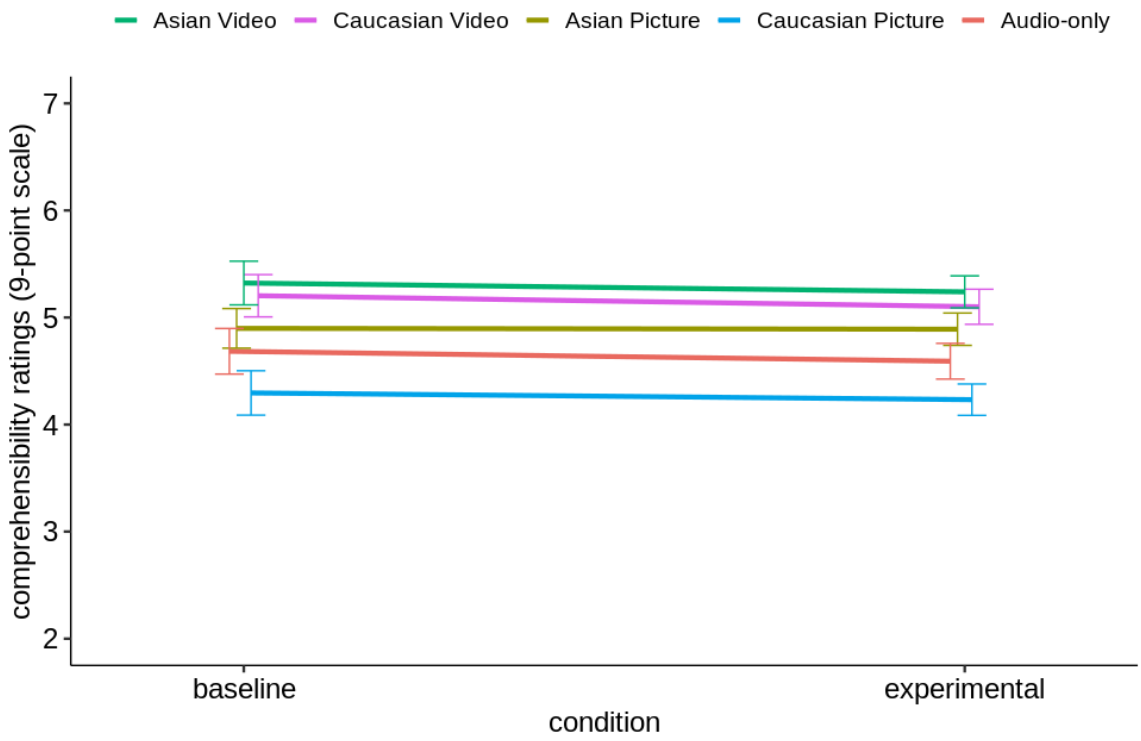
Table 5.5: Mean, median and standard deviation for the comprehensibility ratings.

Condition	Group	N	Mean	St. Dev.	Median
Baseline	Asian Picture	640	4.90	2.49	5
	Asian Video	640	5.32	2.27	5
	Caucasian Picture	640	4.30	2.40	4
	Caucasian Video	640	5.20	2.47	5
	Audio Only	640	4.68	2.08	5
Experimental	Asian Picture	960	4.89	2.44	5
	Asian Video	960	5.24	2.19	5
	Caucasian Picture	960	4.23	2.23	4
	Caucasian Video	960	5.10	2.38	5
	Audio Only	960	4.59	2.15	4

Similarly to the accentedness ratings, the comprehensibility ratings were inspected for male and female speakers separately. The ratings plotted by the gender of the native English speaker revealed that female speakers were, on average, rated as slightly more comprehensible in the *experimental* condition than in the *baseline* condition (Figure 5.14). This tendency is reversed for male speakers who were, on average, rated as *less* comprehensible in the *experimental* condition than in the *baseline* condition. However, these rating patterns are present in all groups regardless of the visual cue employed, which indicates that there was no effect of the speaker‘s perceived ethnicity for neither male nor female speakers.

Figure 5.15 shows comprehensibility scores assigned to individual speakers across all groups. There seems to be some variance among both male and female speakers. Both male speakers in the *baseline* condition were rated as more comprehensible (lower ratings) than all male speakers in the *experimental* condition. A reverse pattern can be observed for the female speakers, where the two female speakers in the *baseline* condition were rated as less comprehensible (higher ratings) than female speakers in the *experimental* condition. Further examination of the same data divided by groups did not show any apparent differences in rating patterns among the groups, that is,

Figure 5.13: The mean comprehensibility ratings for each group in the *baseline* and *experimental* condition with 95% confidence interval. Lower ratings indicate that the utterance was perceived as easier to understand.

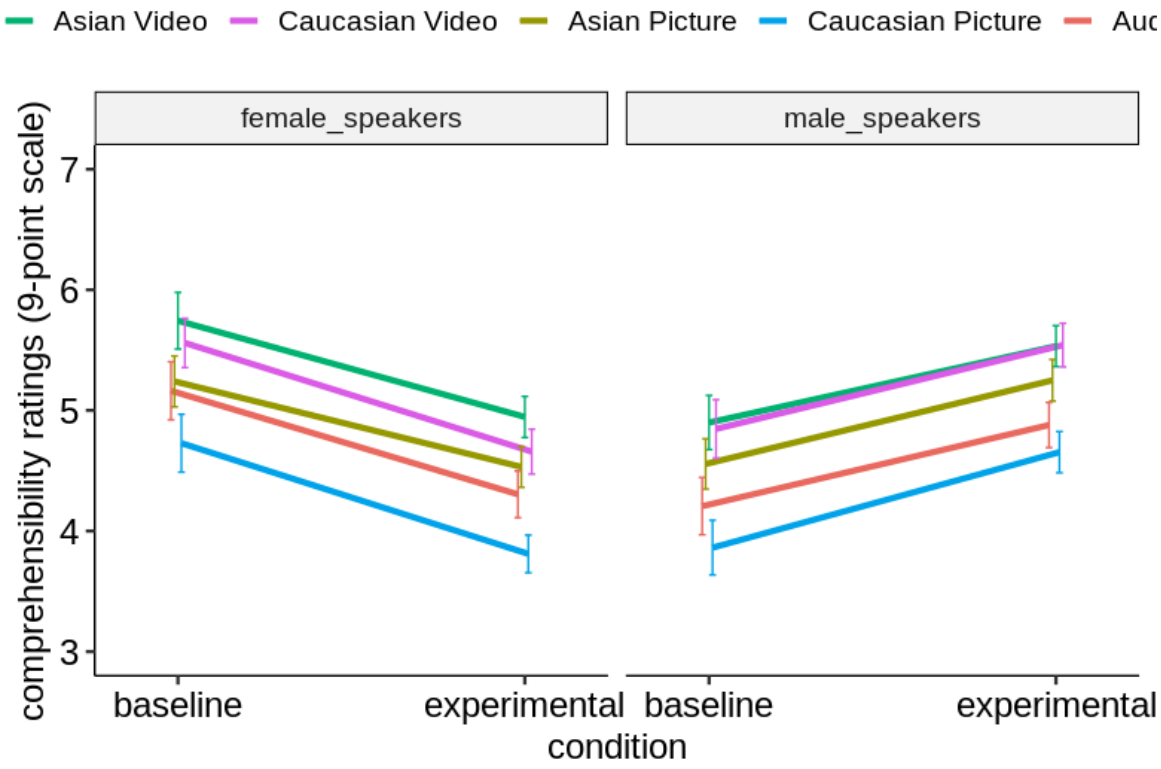


the participants generally agreed on how difficult a given speaker sounded like.

In order to investigate whether there was, indeed, no effect of the speaker's perceived ethnicity on the comprehensibility ratings, a linear mixed-effects model was used, similarly as for the accentedness ratings. The full model included (1) `group` (a factor with 5 levels), (2) `condition` (*baseline* and *experimental*), (3) `gender` (of the native English speaker) and the interaction between these variables (see Appendix F). The self-reported English level (a factor with 6 levels) was also added to the model in order to investigate whether it influenced the comprehensibility ratings. Random intercepts for `participant` and `item`[4] (each audio file) were included to the model along with by-participant random slopes for

[4]The `item` was embedded in the speaker.

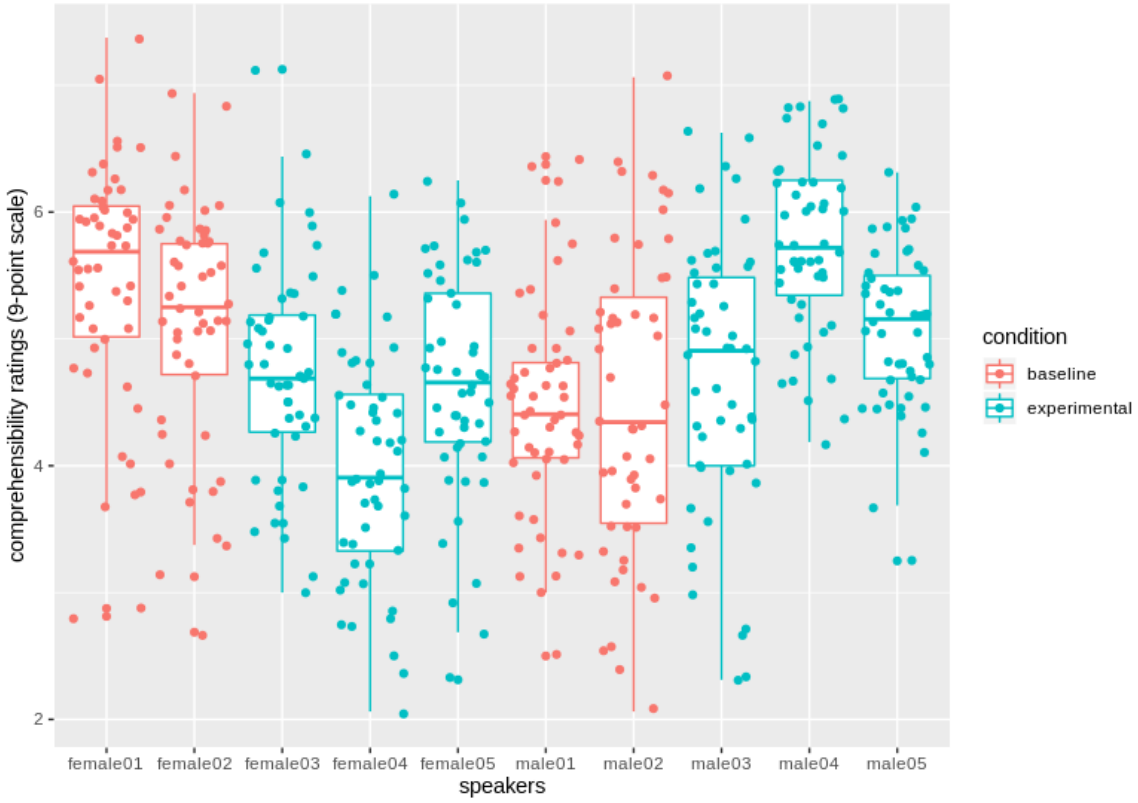
Figure 5.14: The mean comprehensibility ratings for each group in the *baseline* and *experimental* condition by gender of the native English speaker with 95% confidence interval. Lower ratings indicate that the utterance was perceived as easier to understand.



the effect of `condition` and `gender` (of the native English speaker), and their interaction. The maximal model was used to avoid type 1 errors (Barr, 2013) and was justified by model selection. Finally, a vector containing the number of words in each sentence was set as weights.

Models with and without the `English_level` were compared using the Likelihood Ratio Test in order to determine whether the self-reported English level was an important predictor of comprehensibility ratings. The Likelihood Ratio Test indicated that the two models are significantly different ($X^2(5) = 28.62, p < 0.001$) suggesting that the self-reported English level is an important predictor of

Figure 5.15: Boxplots of comprehensibility scores assigned to individual speakers. Higher score indicates that the speaker was more difficult to understand.



comprehensibility ratings. This is rather to be expected since both English level and comprehensibility are self-assessed values and both evaluate some kind of English proficiency. The amount of utterance one thinks he or she understood (comprehensibility) should be naturally related to the self-evaluation of one's English skills. Since the self-assessed English level was a significant predictor, it was kept in the comprehensibility model.

Visual inspection of the residual plots of the full fitted models revealed no apparent deviations from normality or homoscedasticity. Similarly to the accentedness analysis, influential data points were identified using the *influence* function from the *influence.ME* package (Nieuwenhuis et al., 2012) and Cook's

distance was computed for each observation. The new data set was generated using 3 times the mean as a cutoff value and it included 7595 out of the original 8000 points. There was no difference between the model fitted with the full dataset and the model fitted with the reduced dataset in terms of the significance of the fixed effects; therefore the results reported below are the values computed using the model fitted with the full dataset.

Table 5.6: Type III Analysis of Variance Table with Satterthwaite's method for the effect of speaker's perceived ethnicity on the comprehensibility ratings.

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
group	68.74	17.18	4	70.20	1.12	0.35
condition	1.04	1.04	1.	6.76	0.07	0.80
gender	0.46	0.46	1	6.21	0.03	0.87
english_level	473.09	94.62	5	70	6.17	**<0.001**
group:condition	5.51	1.38	4	75	0.09	0.98
group:gender	54.68	13.67	4	75	0.89	0.47
condition:gender	132.84	132.84	1	6.27	8.67	**0.02**
group:condition:gender	20.34	5.09	4	75	0.33	0.86

As mentioned before, the *lmer* function by itself does not provide the p-values. Hence, p-values were obtained using the *anova* function (Kuznetsova et al., 2017) with Satterthwaite's approximation of degrees of freedom (Table 5.6). The effect size was calculated using the *r.squaredGLMM* function from the *MuMIn* package (Barton, 2018). Since mixed effects models also include random effects, two R^2 were calculated following the recommendation in Nakagawa et al. (2017). The *marginal* R^2, which includes only variance of fixed effects, was R^2=0.05 while the *conditional* R^2, which includes variance of both random and fixed effects, was R^2=0.19. The `group` x `condition` interaction was *not* significant ($F(4,75)$=0.09, p=0.98) indicating no effect of the speaker's perceived ethnicity on the comprehensibility ratings. The self-reported English level was significant ($F(5, 70)$= 6.17, $p < 0.001$),

which confirms the results of the Likelihood Ratio Test. Unsurprisingly, participants who assessed their English skills as more native-like evaluated the utterances as relatively easy to understand while the participants whose English level was lower found the utterances less comprehensible. Moreover, the `condition` x `gender` (of the native English speaker) interaction was significant ($F(1, 6.27) = 8.67$, $p < 0.05$), which seems to be in line with the tendencies visible in Figure 5.14, that is female speakers were rated as more difficult to understand in the *baseline* condition than male speakers in the *baseline* condition. Similarly, male speakers were rated as more difficult to understand in the *experimental* condition than female speakers in the *experimental* condition. No other effect was significant (all $p > 0.05$).

Intelligibility

Unlike accentedness and comprehensibility, intelligibility is not a rating on its own. The intelligibility score described and used in the current study is a ratio of correctly transcribed words over the total number of words in the given sentence. Therefore, before proceeding to the analysis, it is necessary to outline the coding procedure used in order to obtain the intelligibility score.

Coding

The entire set of utterances from the audio stimuli was primarily transcribed by the researcher. These transcriptions were then verified during the recording of the video stimuli by each of the actors. Since two of the actors were non-native English speakers, a native English speaker from North America was asked to verify their sets (20 utterances). Finally, a total word count was assigned to each sentence based on the number of *content words*, i.e., nouns, verbs (excluding the copula), adjectives, and adverbs. This procedure is similar to the one described in Kennedy and Trofimovich

(2008).

In order to compute the intelligibility score for each sentence transcribed during the transcription task, all sentences were coded by the number of correctly transcribed *content words* (i.e., nouns, verbs, adjectives, and adverbs). Since participants in this experiment were not native English speakers, performing a simple word count of the exact word matches or even applying some simple regularization techniques, like in Munro and Derwing (1995a), might not have reflected their actual understanding of *non-native* English listeners. Therefore a scoring algorithm was created based on several scoring techniques described and used in Kennedy and Trofimovich (2008) and Sheppard et al. (2017). The scoring procedure was then adjusted to take into consideration the L1 of listeners in this study, in this case Japanese (e.g., difficulties with the distinction between *l* and *r* sounds). The final scoring process was conducted according to the following rules:

(1) The words were regularized, for example, if the target word was *presents* the answer *present* was also counted as correct.

(2) Spelling mistakes common for native speakers of Japanese, like writing *b* instead of *v*, were marked as correct.

(3) Misspelled, yet recognizable words, like *minuets* instead of *minutes*, were counted as correct.

(4) Homophonous, like *too* instead of *two* were counted as correct. This includes lexemes with mistakes common for Japanese native speakers (e.g., *bitch* /bɪtʃ/ instead of *beach* /bitʃ/, where the only difference is the contrast between vowels /ɪ/ and /i/ that is difficult for the native speakers of Japanese to distinguish).

(5) Equivalent forms, like *a lot* instead of *lots* were counted as correct.

(6) Omissions of the third person *s*, like *want* instead of *wants* were counted as correct.

Finally, the intelligibility score was computed by taking the ratio of correctly

transcribed content words over the total number of content words in a given sentence.

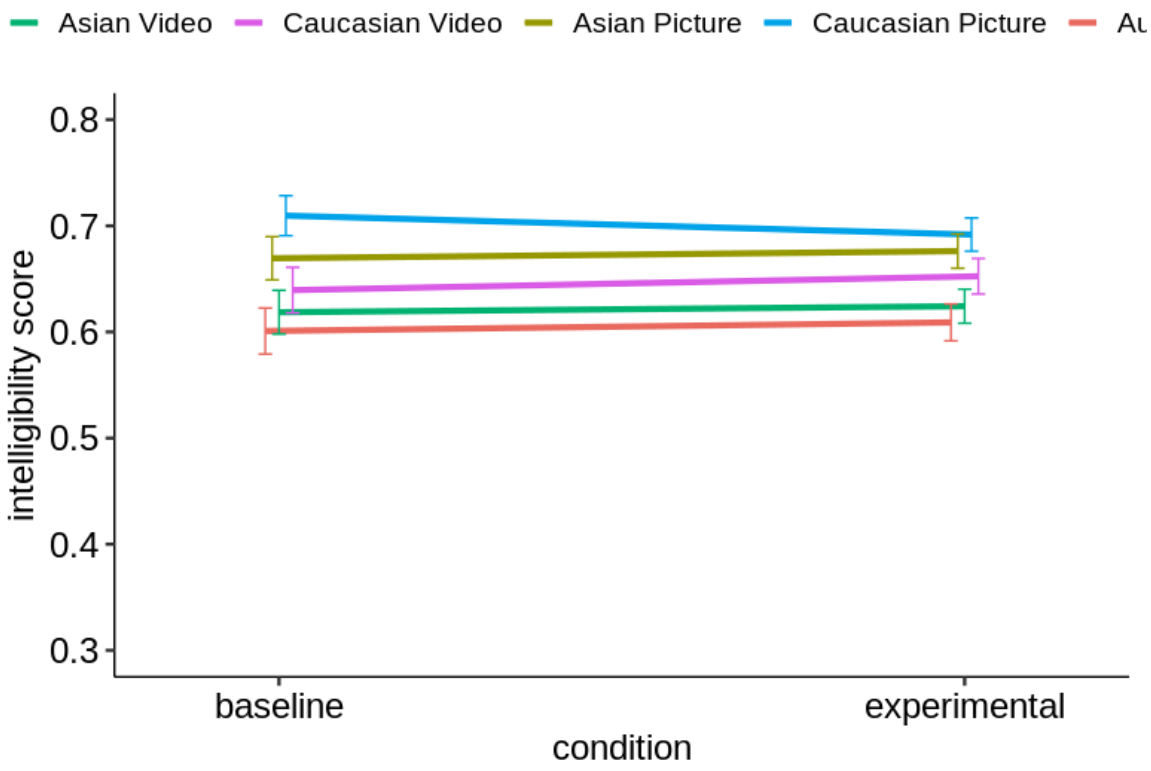
In order to explore the general tendencies in the data, the mean intelligibility score was first plotted for each group and each condition. Figure 5.16 indicates no effect of the perceived ethnicity of the speaker on the intelligibility with the mean intelligibility scores being comparable between the groups in both the *baseline* and *experimental* conditions. Moreover, there seem to be no apparent change in the intelligibility between the *baseline* and *experimental* condition for none of the groups (Table 5.7).

Table 5.7: Mean, median and standard deviation for the intelligibility score.

Condition	Group	N	Mean	St. Dev.	Median
Baseline	Asian Picture	640	0.67	0.26	0.71
	Asian Video	640	0.62	0.27	0.63
	Caucasian Picture	640	0.71	0.24	0.75
	Caucasian Video	640	0.64	0.28	0.67
	Audio Only	640	0.60	0.28	0.60
Experimental	Asian Picture	960	0.68	0.26	0.71
	Asian Video	960	0.62	0.25	0.63
	Caucasian Picture	960	0.69	0.25	0.71
	Caucasian Video	960	0.65	0.26	0.67
	Audio Only	960	0.61	0.28	0.60

Figure 5.17 shows the mean intelligibility scores for each group and condition plotted for male and female speakers separately. Female speakers in the *experimental* condition were generally more intelligible than the female speakers in the *baseline* condition. Conversely, male speakers in the *experimental* condition were, on average, less intelligible than the male speakers in the *experimental* condition. This tendency seems to be comparable with the comprehensibility ratings. The level of difficulty that participants *perceived* for a given utterance appears to be in line with how difficult

Figure 5.16: The mean intelligibility scores for each group in the *baseline* and *experimental* condition with 95% confidence interval.



it really was for them to transcribe that utterance. However, there seems to be no difference between the groups. Hence, it seems that the speaker's perceived ethnicity had no effect on the intelligibility scores even after considering the data of male and female speakers separately.

Looking at the intelligibility scores for each speaker revealed that `female01` was less intelligible than other female speakers or even other male speakers. This resulted in an overall lower intelligibility score in the *baseline* condition for female speakers. On the other hand, both male speakers in the *baseline* condition got higher intelligibility scores than any of the male speakers in the *experimental* condition (see Figure 5.18).

A similar pattern can also be observed after looking at the intelligibility scores plotted by the speaker per group.

Figure 5.17: The mean intelligibility scores for each group in the *baseline* and *experimental* condition by gender of the speaker with 95% confidence interval.

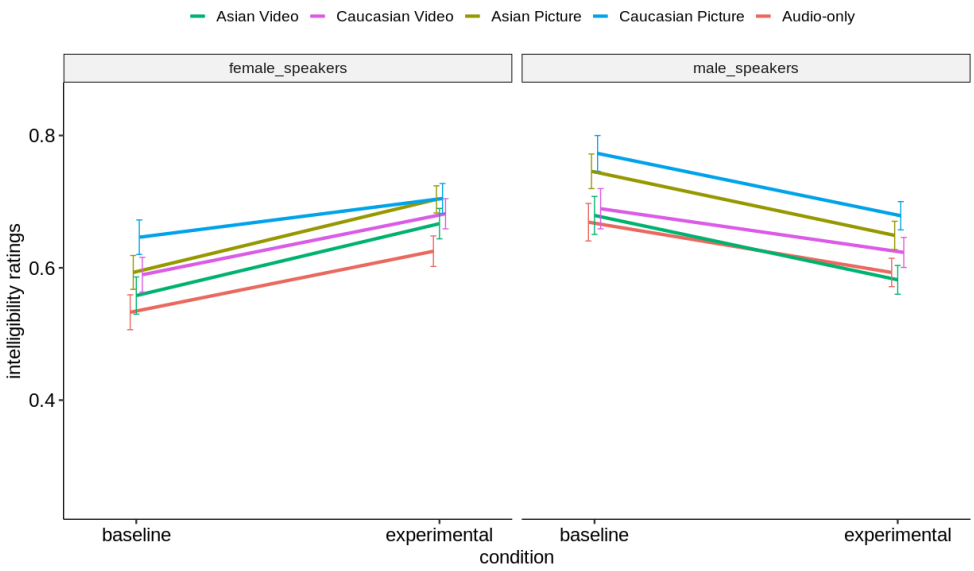
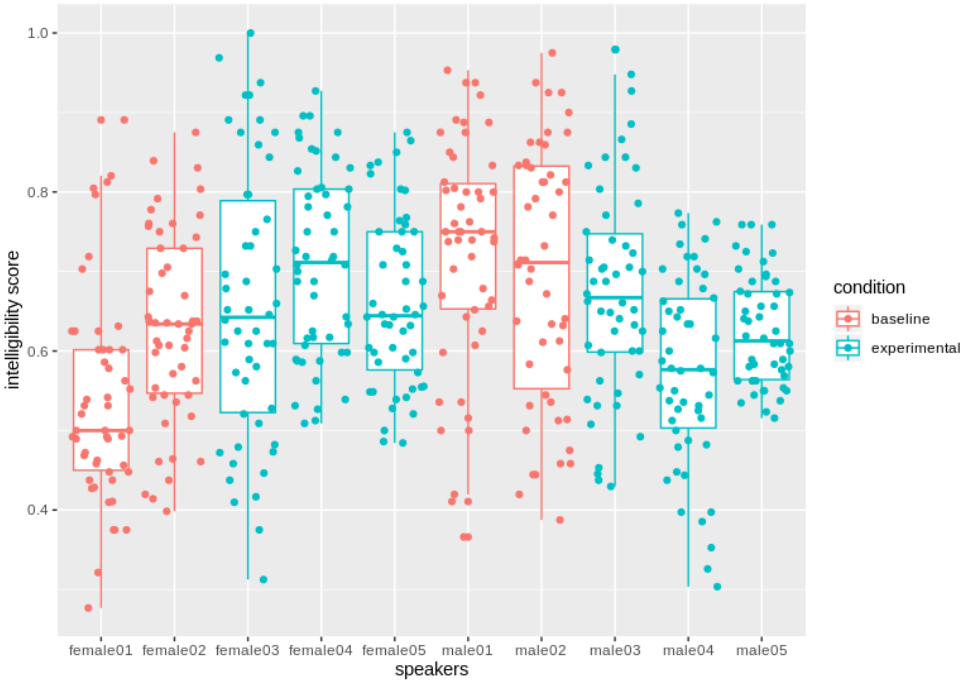


Figure 5.18: Boxplots of the intelligibility scores for each speaker.



The intelligibility scores are quite different from accentedness and comprehensibility ratings. Therefore they require a slightly different statistical approach. Each score itself is a *ratio* of correctly transcribed words and the total number of words in a sentence and it is bounded by 0 and 1. Thus, in order to analyze whether there was, indeed, no effect of the speaker's face, a generalized linear mixed model was employed. The generalized linear model can be considered as an extension of the linear mixed effect model, and it allows to build a model using a *ratio* response variable.

A generalized linear mixed model was fitted using the *glmer* function from the *lme4* package (Bates et al., 2015). The full model included (1) `group` (a factor with 5 levels), (2) `condition` (*baseline* and *experimental*), (3) `gender` (of the native English speaker) and their interaction all modeled as fixed effects (see Appendix F). The self-reported English level (a factor with 6 levels) was also added to the model as a fixed effect. Furthermore, random intercepts for `participant` and `item` along with by-participant random slopes for the effect of `condition` were included to the model with binomial errors and a logit link function. By-participant random slopes for the effect of `gender` (of the native English speaker) were not included, as in the accentedness or comprehensibility models, in order to prevent overfitting that could affect the overall results. The inclusion of random slopes was also justified by model comparison (Barr, 2013). The choice of random slopes was also justified by model selection. Moreover, a vector containing the total number of words in each sentence was set as weights of the model to account for the ambiguity of ratio-like scores (e.g., 5 out of 10 and 10 out of 20 result in the same 0.5 ratio).

Visual inspection of the residuals plots did not indicate any apparent deviation from homoscedasticity or normality. Influential data points were identified using the *influence* function from *influence.Me* package (Nieuwenhuis et al., 2012) similarly as

for the accentedness and comprehensibility ratings. There was no difference between the model with the points identified as influential and the model without these points in terms of the significance of fixed effects. Therefore, the results reported below were obtained using the model fitted with the full dataset.

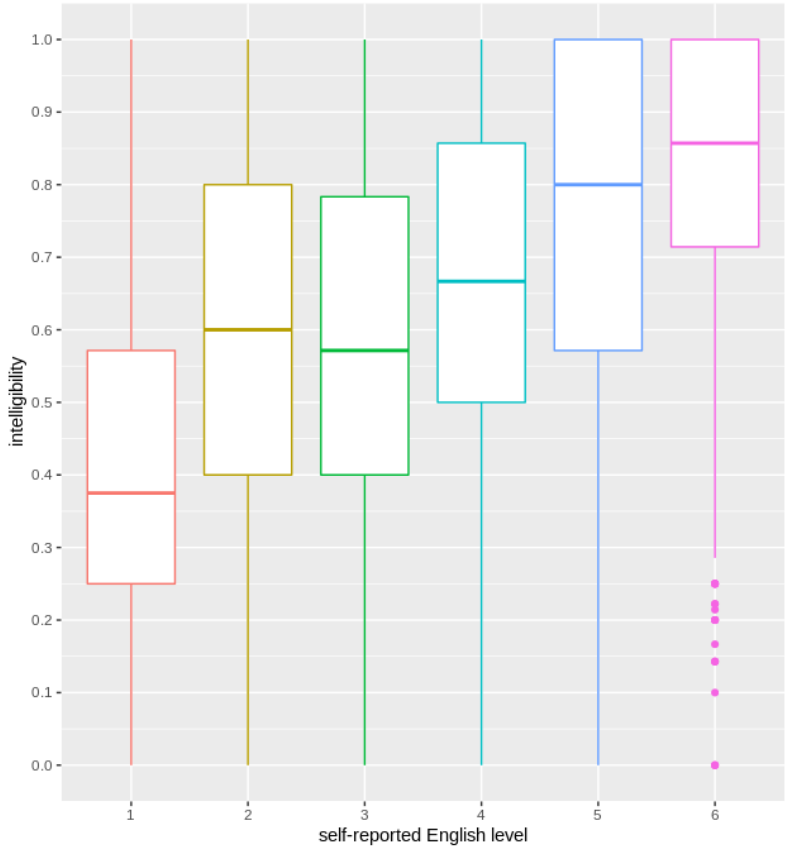
Table 5.8: Analysis of Deviance Table (Type III Wald chisquare tests) for the effect of speaker's perceived ethnicity on the intelligibility of English utterances.

	Chisq	Df	Pr(>Chisq)
`(Intercept)`	40.48	1	$<$ **0.001**
`group`	0.39	4	0.98
`condition`	0.01	1	0.93
`gender`	2.17	1	0.14
`English_level`	43.46	5	$<$ **0.001**
`group:condition`	4.20	4	0.38
`group:gender`	18.91	4	$<$ **0.001**
`condition:gender`	12.70	1	$<$ **0.001**
`group:condition:gender`	9.19	4	0.06

The p-values were acquired using the *Anova* function from the *car* package (Fox & Weisberg, 2011). The effect size was calculated using the *r.squaredGLMM* function from the *MuMIn* package (Barton, 2018). Since mixed effects models also include random effects, two R^2 were calculated following the recommendation in Nakagawa et al. (2017). The *marginal* R^2, which includes only variance of fixed effects, was R^2=0.04 while the *conditional* R^2, which includes variance of both random and fixed effects, was R^2=0.17. The intercept was significant ($p < 0.001$). The `group` x `condition` interaction was *not* significant ($X^2(4) = 4.20, p = 0.38$) indicating that the speaker's perceived ethnicity had no effect on the intelligibility of the utterances delivered by a native English speaker. There was also no effect of speaker's perceived ethnicity when looking at male speakers and female speakers separately ($X^2(4) = 9.19, p = 0.06$). The small p-value observed here seems to reflect the tendencies shown in Figure 5.17. While in general the intelligibility ratio descends for male speakers and ascends for

female speakers when comparing the *baseline* and *experimental* conditions, one can also observe an interaction between the Caucasian Video group and Asian Picture group for female speakers and an interaction between the Asian Video group and Audio-only group for the male speakers. However, these interactions seem to be too small to yield significant results.

Figure 5.19: Boxplots of the intelligibility scores by the self-reported English level ranging from 1 - lower intermediate to 6 - native-like.



The self-reported English level was significant ($X^2(6) = 49.82, p < 0.001$), which is to be expected since intelligibility is one of the measures for English proficiency and therefore it would be affected by the listener's English level. Figure 5.19 indicates that indeed participants who evaluated their English level as being close to native

speakers have a very high overall intelligibility scores, while participants whose English level oscillated around lower intermediate have low intelligibility scores with a mean slightly below 0.4. The `group` x `gender` (of the native English speaker) interaction on intelligibility scores was significant ($X^2(4) = 18.91, p < 0.001$) indicating that there was a difference in intelligibility of male and female native English speakers depending on group. Finally, the `condition` x `gender` (of the native English speaker) interaction was significant ($X^2(1) = 12.70, p < 0.001$). This tendency is visible in Figure 5.18, where female speakers in the *baseline* condition are less intelligible, regardless of the group, than female speakers in the *experimental* condition, while male speakers in the *baseline* condition are more intelligible than male speakers in the *experimental* condition. No other effects were significant (all $p > 0.05$).

5.2.4 Correlation between the IAT effect and the Accentedness, Comprehensibility, and Intelligibility

The data for accentedness, comprehensibility, and intelligibility were further analyzed in order to investigate if there was any correlation between the listener‘s D score in Implicit Association Test and their performance in Perception Experiment. More precisely, the correlation between each of the three measures and the D scores was computed for Asian face data and Caucasian face data separately. Before applying any statistical measures, the data were preprocessed in the following way:

(1) Mean values of accentedness ratings, comprehensibility ratings, and intelligibility scores were computed per participant for the *experimental* condition. Thus, each participant had exactly one mean rating for the accentedness, one mean rating for the comprehensibility, and one mean intelligibility score associated with the *experimental* condition.

(2) Since there was no effect stemming from the type of stimuli (picture or video), the video and picture data were analyzed as one group for each ethnicity separately. Consequently, the data were analyzed in two groups: (1) Asian Face Group (data from the Asian Picture group and the Asian Video), and (2) Caucasian Face Group (data from the Caucasian Picture group and the Caucasian Video group).

The preprocessed data were analyzed using `R language` (R Core Team, 2019). Pearson correlation between the D scores and mean accentedness ratings, comprehensibility ratings, and intelligibility scores in the *experimental* condition was calculated using the `cor.test` function (included as a default function in the `R Language`) and the `ggscatter` function from the `ggpubr` package (Kassambara, 2018). All the correlations were computed separately for the Caucasian face group and the Asian face group.

No correlation came out significant for the Caucasian face group data (all p's $>$ 0.05) indicating that there was no relationship between the strength of the "American = Caucasian" association and the accentedness ratings, comprehensibility ratings, or intelligibility scores of listeners who saw a Caucasian face. There was also no significant correlation between D scores and the comprehensibility ratings nor the intelligibility scores in the Asian face group (all p's $>$ 0.05), that is, there was no clear relationship between listeners' implicit bias and their comprehensibility ratings and intelligibility scores. There was, however, moderate but *not* significant correlation between listeners' D scores and their ratings of accentedness of the East Asian guise ($r(28)$=-0.34, p=0.06). The negative correlation here suggests that listeners with higher IAT score assigned lower accentedness ratings to the East Asian guise. In other words, listeners who showed stronger "American = Caucasian" association had the tendency to judge native utterances as being somewhat more accented when presented with an Asian face. Furthermore, this tendency was *not* present in the *baseline*

condition, where the same participants rated audio-only stimuli with no visual cues ($r(28)$=-0.09, p=0.62). Thus, this may indicate that the strength of listeners' implicit association "American = Caucasian" affected their perception of the accentedness of native English speakers when they thought that these speakers are East Asian-looking.

Chapter 6

General Discussion

The current research aimed at investigating the effect of speakers‘ perceived ethnicity, either East Asian or Caucasian, on the perception of English utterances (spoken by native speakers of American English) by Japanese listeners. It has been shown previously that native speakers of English can perceive accented English utterances *more* accurately when they believe the speaker to be East Asian (McGowan, 2015). Conversely, native speakers of English may perceive English utterances as spoken by native English speakers *less* accurately when they believe the speaker to be East Asian (Babel & Mellesmoen, 2019; Babel & Russell, 2015; Rubin, 1992). That is, the listeners' expectations impact their perception when they listen to speech in "congruent" (non-native speech with East Asian face) or "incongruent" (native speech with East Asian face) situations. These expectations are presumably shaped by one's experience, as suggested, for instance, by experience-based accounts of speech perception, where the listener may have developed a stronger or weaker "American (English) = Caucasian" association through their personal life experiences.

In the current study, the sum of these experiences was tentatively measured

through the use of an Implicit Association Test (IAT). This test measures the strength of a possible "American = Caucasian" association. I hypothesized that a stronger association is most likely to impact speech perception in some ways. The results of the IAT suggested that, indeed, the Japanese listeners who took part in the current study exhibited a moderate to strong "American = Caucasian" association. These results are to be expected given that the participants were all born and raised in Japan in monolingual Japanese families, and did not spent any considerable time in English-speaking countries like the US, the UK or Australia (average time spent in English-speaking countries = 51 days, SD = 103 days). That is, they were more likely to hold the stereotypical view of an American being Caucasian. It is true that the American society became significantly more diverse in the last two decades with the foreign-born population reaching 31.1 million in 2000 and 44.7 million in 2018 (The Brookings Institution, 2019) with the number of non-Caucasian citizens oscillating around 23.5% of the population (U.S. Census Bureau, 2019). However, the same implicit bias towards "American = Caucasian" pairing was also found for Americans, whether they were Caucasian, Asian, or African Americans (Devos & Banaji, 2005; Yi et al., 2013, 2014).

After confirming that all the Japanese participants had a relatively strong "American = Caucasian" association, the results of the perception tasks were analyzed. In the perception experiment the same participants as in the IAT were asked to rate native American English utterances for their accentedness and comprehensibility on a 9-point Likert scale. They were also asked to transcribe each utterance as a measure of intelligibility.

The participants listened to all native English utterances twice. In the first listening, they were asked to transcribe each sentence for measuring their intelligibility and after the transcription, rate each sentence for its comprehensibility.

The comprehensibility ratings were collected during the first listening, as hearing the same utterance twice could have made it *appear* easier to understand. After a break, listeners were asked to listen to the same stimuli one more time and rate each sentence for accentedness. The order of transcription and accent rating tasks was the same as in Babel and Russell (2015).

The stimuli were divided into two sets, one presented in the *baseline* condition and one presented in the *experimental* condition. In the *baseline* condition, all participants rated and transcribed the same 40 sentences presented as audio-only stimuli. The main purpose of the *baseline* condition was to ensure that any differences between the groups in the *experimental* condition will be due to the visual cue and not due to differences between individual listener or groups of listeners. In the *experimental* condition, participants listened to 60 additional utterances presented with different visual cues using a matched-guise technique. Group 1 saw pictures of Asian-looking speakers, group 2 saw pictures of Caucasian-looking speakers, group 3 saw videos of Asian-looking speakers, group 4 saw videos of Caucasian-looking speakers, while group 5 listened to audio-only stimuli.

The data were analyzed using the mixed-effects model. Contrary to the predictions based on the outcome of the IAT, native Japanese listeners were *not* affected by the speaker's perceived ethnicity when listening to the native (American) English utterances. Specifically, they did not rate the same native English speech as more accented or less comprehensible only because they were led to believe that the speaker is East Asian-looking rather than Caucasian-looking. Moreover, the ethnicity of the speaker did *not* affect the intelligibility scores of non-native English listeners as measured by a transcription task.

The differences between native behavior, as reported in previous studies, and non-native behavior, as reported in the current study, may be explained from a theoretical

point of view, as discussed in the next section. These differences may also stem from methodological limitations of the current study, or different methods used in current study versus previous studies, which will also be discussed in section 6.5.

6.1 Reverse Linguistic Stereotyping vs. The Experience-based Models

Taken together, the lack of significant effect for the general accentedness, comprehensibility, and intelligibility results may be explained by two competing theories, the Reverse Linguistic Stereotyping (RLS) and the experience-based models. In view of the RLS, listeners who have a negative bias towards Asian-looking speakers of English may mistakenly "hear" non-native accent in a native English speech or pay less attention to its content, which would result in lower intelligibility (Lippi-Green, 2012).

Although the results of the IAT indicated that listeners in the current study *implicitly* associated being American with being Caucasian, this does not mean that they had a *negative* bias towards East Asian-looking speakers of English. Hence, under the RLS model, it could be argued that the listeners did not have a negative bias, and that that may have resulted in them not being affected by the speaker's ethnicity.

Whether native Japanese listeners did have a *negative* bias towards East Asian-looking native English speakers could be potentially tested with another IAT. For instance, Babel and Russell (2015) tested that bias pairing unanimously "Asian" and "Caucasian" surnames with positive (e.g., holiday) and negative (e.g., suffering) lexemes. While surnames may not relate directly to East Asian-looking native English speakers, they could be potentially replaced with very short same-length

videos of East-Asian looking and Caucasian-looking native English speaker saying something like "Hi!" or "Hello!" Faster categorization in the task where East Asian face is being paired with the negative lexical item and Caucasian face is being paired with the positive lexical item would then indicate a *negative* implicit bias towards East Asian-looking speakers. In addition to this implicit method, the negative bias could also be tested with an explicit questionnaire, such as the speech evaluation instrument (Zahn & Hopper, 1985). While it seems somewhat unlikely that native Japanese listeners would have a negative bias towards East Asian-looking native English speakers, confirming such a bias would provide evidence against the RLS.

The experience-based models offer an arguably better and more generalizable explanation for the null effect of the speaker's ethnicity observed in the current study. The experience-based approach assumes that speech perception is a socially-weighted process that can be affected by listeners' experiences and societal stereotypes, which usually associate being American with being Caucasian (Devos & Banaji, 2005; Gnevsheva, 2018; Yi et al., 2013). Therefore, when a listener hears native American English speech while seeing an East Asian face (or non-native English speech while looking at a Caucasian face) it creates an *incongruent* condition. This leads to a mismatch effect, which can then affect listener's speech perception (Babel & Mellesmoen, 2019; Babel & Russell, 2015; Gnevsheva, 2018). Conversely, if the listener hears non-native English speech paired with an Asian face (or native American English speech paired with a Caucasian face), this would create a *congruent* condition. In this case, the socioindexical information stored in the brain would match the linguistic information, which could potentially facilitate speech processing (McGowan, 2015).

In the view of this socially-weighted speech perception model, Japanese participants, who also showed moderate to strong preference towards the "American

= Caucasian" pairing, should act accordingly and rate native English utterance as *more* accented and *less* comprehensible when presented with an East Asian guise than when presented with a Caucasian guise. Moreover, the listeners should also experience *more* difficulties with the transcription task in the East Asian guise condition than in the Caucasian guise condition as the former would create a mismatch between the social and linguistic information.

Yet, native Japanese listeners in the current study appeared to be unaffected by the speaker's perceived ethnicity. This could be potentially explained by the experience of the Japanese listeners who took part in the current experiment. The IAT indicated an implicit bias towards the "American = Caucasian" pairing. However, this pairing is not exactly the same as "native English speaker = Caucasian." Hence, the lack of significant effect in the current study may be related to the label used in the IAT, that is "American" and not a "native English speaker." Namely, it is possible that the listeners in the current study *did* associate relatively strongly being American with being Caucasian, but at the same time, they *did not* associate being a native English speaker with being Caucasian. While Kubota and Fujimoto (2013) argued that in Japan being a native English speaker is often conflated with being Caucasian, the constantly changing demographic and increasing trips overseas, including to English-speaking countries in Asia (JTB Tourism Research Consulting Co., 2020) might have been an opportunity for building new experiences, which in the end affected perception of native Japanese speakers. This may be especially true for the educated younger generation living in Tokyo, such as the Japanese listeners in the current experiment who were mostly undergraduate students from the University of Tokyo recruited at the Komaba Campus. Interestingly, in 2015 at Komaba Campus there was about the same number of students from Asian English speaking countries (e.g., the Philippines, Singapore), where the English speakers are more likely to be Asian-looking, as from

the English speaking countries from other parts of the world (e.g., the US, the UK), where the English speakers are more likely to be Caucasian-looking (The University of Tokyo, n.d.).

It is also possible that the results of the current experiment reflect the differences in the nature of the link between social and linguistic information for native and non-native listeners. Sumner et al. (2014) presented a socially-weighted model in which social features and lexical representations are being extracted from speech. Sumner and colleagues explain that this social and linguistic information interacts with each other in the process of social weighting where they are being linked in order to understand the utterance. If the nature of this link is different for non-native listeners, this would mean that a new, adjusted model should be developed in order to account for these differences.

It has been demonstrated that native listeners are able to identify emotions much better than non-native listeners by listening to the voice alone (Nakamichi, Jogan, Usami, & Erickson, 2003). Hence, it is possible that the linking process described in Sumner et al. (2014) were social features are being combined with linguistic information is much *stronger* for native listeners than it is for non-native listeners. While the listeners in the current study most likely perceived that the speaker is Asian/Caucasian, there was *no* interaction between this information and the linguistic information, that is native English speech. Although Japanese participants in this study seem to have a relatively strong bias towards the "American = Caucasian" pairing, the *link* between this social information and the linguistic information, such as native English accent, might have been weaker for them than for native English speakers. Hence, regardless of whether the social features were initially extracted from the speech stream, there was no interaction between the social information and the linguistic information, which resulted in no

social weighting (i.e., no influence of social cues on linguistic information). This led to the lexical representations being probabilistically inferred based entirely on the linguistic information.

In order to account for both native and non-native listeners, a refined model of socially-weighted speech perception should be proposed. In a model like this, the process of social weighting would depend not only on whether social representations were stored in the brain or not but also on the nativeness of the listener. While for native listeners, the social factors extracted from speech interact with linguistic information, for non-native listeners, they are either not being linked with each other or the link is not strong enough for the social factors to influence linguistic features.

In order to confirm this refined model for non-native listeners, the same perception experiment should also be conducted with native speakers of American English. If an effect of the speaker‘s ethnicity would be found (just as in Babel & Mellesmoen, 2019; Babel & Russell, 2015; Rubin, 1992) this would indicate that indeed the social weighting process, at least for speaker's ethnicity, occurs only for native listeners. Conversely, if no effect is found also for native English listeners from the US, this could potentially suggest that a change in demographics, that is an increasing number of foreigners and more diverse society of the US (The Brookings Institution, 2019; U.S. Census Bureau, 2019), led to different experience, which affected the process of speech perception in a different way (i.e., canceled the effect of speaker's ethnicity). While this null effect of the speaker‘s ethnicity could be potentially observed for native English listeners, this would not be in line with the results presented in Babel and Russell (2015) and Babel and Mellesmoen (2019) where both experiments were conducted in even more diverse and multicultural Vancouver. Hence, given these two studies, one could expect that an effect of the speaker‘s perceived ethnicity would be found for native English listeners from the US.

Alternatively, whether an effect of the speaker‘s perceived ethnicity is observed or not may depend strongly on the experimental design. For instance, in Rubin (1992), which remains the most cited study on the effect of ethnic bias with 612 citations at the time of this writing, native English listeners were presented with *only* one mini-lecture in *only* one condition (Asian guise or Caucasian guise), and provided *only* one accentedness rating *after* they listened to the lecture. Since there was no baseline condition with only one rating per participant, the effect of speaker‘s ethnicity on the accentedness ratings might have been due to individual differences between the participants rather than to the effect of ethnicity itself. Furthermore, in the same study, the intelligibility was assessed by a cloze test administrated *after* the listening task. This means the participants would have to *memorize* the entire mini-lecture, which was around 4-minute long. Hence, it may be the participants' memory (how many words they could remember) rather than the effect of the speaker‘s face that led to the lower intelligibility in the Asian guise condition.

Similarly, whether an effect of speaker‘s ethnicity is found seems to depend on the "nativeness" of stimuli included. While most studies incorporating only native voices reported an effect of speaker's ethnicity (Babel & Russell, 2015; Hanulíková, 2018; McGowan, 2011; Rubin, 1992; Rubin et al., 2015), a number of studies incorporating non-native voices did not (de Weers, 2019; Rubin et al., 1997; Rubin & Smith, 1990). Including a continuum of non-native voices might have had created a more apparent distinction between native and non-native English speakers in the current research. Having a wider range in accents would also introduce more variance to the data as 58% of the ratings in the current study were clustered on the higher (more native-like) end of the Likert scale, that is, scores 7, 8, and 9.

Furthermore, non-native English speech in the Asian guise condition (pictures or videos) would possibly be *more* intelligible that the native English speech in the same

condition as listeners may be more likely to *expect* an Asian speaker to speak with a foreign accent (McGowan, 2015). Conversely, non-native English speech presented with the Caucasian guise could suffer from the mismatch effect making the utterances less intelligible than native English speech presented with Caucasian guise.

Finally, it is important to mention that there are several studies, which did not find any effect of the speaker‘s ethnicity on speech perception by native English listeners just as in the current study. For instance, Rubin and Smith (1990) used almost an identical design as in Rubin (1992) but they employed only non-native voices. Rubin and Smith (1990), just like the current study, did not find any effect of ethnic bias on speech perception of native English listeners. Similarly, both Rubin et al. (1997) and de Weers (2019) included a mix of native and non-native English speech and did not report any effect of the speaker‘s ethnicity on the speech perception by native English listeners. These results can be explained here by two factors. One is different experiences of the listeners. As it was already discussed above, Japanese listeners in the current study and potentially native listeners in Rubin and Smith (1990), Rubin et al. (1997), and de Weers (2019) might have just different experience having encountered a great percentage of both East Asian-looking and Caucasian-looking native English speakers. On the other hand, it is also possible that the results were affected by the difference in research design between the current study and the previous studies. These limitations will be further discussed in section 6.5.

To summarize, while the lack of significant effect of the speaker's perceived ethnicity on speech perception by Japanese listeners may be explained by the RLS model (Kang & Rubin, 2009; Rubin, 1992), it can also be explained by the socially-weighted speech perception model (Sumner et al., 2014), which accounts for *all* socioindexical cues making it the model with stronger explanatory power. Given that no effect of ethnic bias on the accentedness, comprehensibility, and

intelligibility was observed in the current study, it is possible that the participant in the current study might just have encountered more Asian-looking native or near-native speakers of English and that their experience led to no effect of speaker's perceived ethnicity. It is also possible that non-native listeners are *less* influenced by the social information, such as the speaker‘s ethnicity, as they don't use social weighting to link social features with linguistic representations in the process of speech perception, that is they rely solemnly on the linguistic information. Finally, it is also possible that given the delicate nature of sociolinguistic experiments, small changes in research design led to different outcomes.

This is not to deny that socioindexical cues may affect speech perception as it is predicted by the socially-weighted model (Sumner et al., 2014). There are numerous studies to account for the interaction of social and linguistic information in the process of speech perception (e.g., Drager, 2010, 2011; Hay & Drager, 2010; Hay, Nolan, & Drager, 2006; Hay, Warren, & Drager, 2006). However, one should be careful to generalize such findings to all everyday life situations since these effects were observed (or not) in a "very constrained laboratory experiments involving rather artificial tasks"(Foulkes & Hay, 2015). Hence, one can not assume that the null effect of the speaker‘s ethnicity observed in the current study will be the same for *all* Japanese listeners as it seems to strongly depend on their *experience*. Investigating the effect of ethnic bias on the perception of native and non-native Japanese speech by native Japanese speakers, especially those who live in places with relatively few foreigners, like the island Shikoku, could potentially provide more evidence for the important role of experience in building, storing, and accessing the social representations.

6.2 Picture Stimuli vs. Video Stimuli

The current research also evaluated whether the effect of the speaker‘s perceived ethnicity would differ depending on whether the listeners were presented with pictures or videos featuring Caucasian or East Asian speakers. Using only accentedness ratings, Zheng and Samuel (2017) demonstrated that simply presenting *pictures* of East Asian or Caucasian faces introduces *demand characteristics*, that is a situation where participants guess the purpose of the experiment and act accordingly. Zheng and Samuel argue that when only pictures are used, listeners are likely to guess the purpose of the experiment and rate Asian guises as accented and Caucasian guises as having native English accent just because they guessed that assessing the role of ethnicity is the purpose of the experiment. Furthermore, in Zheng and Samuel's study the effect of guessing the purpose of the experiment was mostly gone when they replaced static pictures with dubbed videos. This led them to argue that the effect of speaker's perceived ethnicity may take place not on the *perception* level but on the *interpretation* level, that is the listeners did not *perceive* a foreign accent but rather *decided* they heard a foreign accent (see chapter 4 subsection 4.2.3). Hence, Zheng and Samuel's work would suggest that there will be a difference between the effect of video stimuli and the effect of picture stimuli such that in the picture condition, participants will guess the purpose of the experiment and evaluate the East Asian guise as less native-like.

Although Zheng and Samuel (2017) demonstrated that the type of stimuli (pictures or videos) may change the effect of the speaker‘s perceived ethnicity at least on the accentedness ratings performed by native English listeners, this effect could not be replicated in the current study for non-native English listeners. On the contrary, no interaction between the type of stimuli and the speaker‘s ethnicity was found. Japanese participants rated both accentedness and comprehensibility in a similar way

regardless of whether they were presented with pictures or videos of East Asian and Caucasian guises. Furthermore, they also performed similarly on the transcription task suggesting that the type of stimuli was also irrelevant for their intelligibility.

It is possible that the lack of effect was created by the nature of the link between the social and linguistic knowledge for non-native English listeners, which led to the general lack of difference between the groups. While Japanese participants did show a moderate to strong "American = Caucasian" association, this association did not affect their perception of native English utterances. It appears as the socioindexical factors, whether retrieved or not, were not linked with the linguistic information for both the picture stimuli and the video stimuli. This resulted in a null effect of the speaker's perceived ethnicity on accentedness, comprehensibility, and intelligibility. Conversely, it also possible that while the Japanese listeners associated being American with being Caucasian, this is not the same as associating *all* native English speakers with being Caucasians. Native Japanese listeners in the current study were undergraduate students at the University of Tokyo, where they could encounter many native or near-native Asian-looking English speakers. Hence, their personal experience may lead to them just not being affected by the speaker‘s ethnicity despite the type of stimuli.

The lack of a significant difference between the video and picture stimuli could also be attributed to different choices in the research design. While Zheng and Samuel (2017) employed single words pronounced with a certain degree of foreign accent, the current study used whole sentences spoken in natural native American English voice. It is, therefore, possible that using single words made it easier for participants to guess the purpose of the experiment and adjust their accentedness ratings accordingly. Conversely, longer utterances spoken in natural speech, which were used in the current study, might have made that "guessing" process more challenging, especially as the

non-native participants would possibly be more concentrated on listening to the native English speech rather than on trying to figure out the purpose of the experiment.

Finally, employing native English voices instead of a "nativeness" continuum (like in Zheng & Samuel, 2017) might have also affected the results. While with the native and non-native stimuli, listeners would be more likely to use the whole scale equally, with only native English voices a high proportion of samples (about 58%) were rated as 7, 8, or 9, that is as more native-like. It is possible that including non-native speech would introduce more variance to the data and lead to a significant interaction between the speaker's ethnicity and the type of stimuli.

Unlike in (Zheng & Samuel, 2017), the participants in the current study did not show any significant differences between the ratings of accentedness and comprehensibility for picture and video stimuli. Similarly, the type of stimuli did not affect their intelligibility scores. This potentially provides further evidence for the lack of the effect of ethnic bias on speech perception by Japanese listeners.

6.3 Speaker's Gender and the Effect of Speaker's Ethnicity

Furthermore, the results were also evaluated dividing the data by the gender of the speaker. Speaker's gender is usually an important socioindexical cue, which accounts for the difference, for instance, in speech production and perception (Strand, 1999). As it was presented in section 4.1, listeners were found to adjust their perceptual boundaries depending on whether they believed the speaker to be a male or a female (Johnson et al., 1999; Strand & Johnson, 1996). Yet, a great number of previous studies evaluating speech perception by native English listeners employed only one gender, usually female (Babel & Mellesmoen, 2019; McGowan, 2011, 2015; Rubin,

1992; Rubin et al., 1999, 1997).

Previous studies on gender as a socioindexical cue indicated that looking at male and female speakers separately could potentially yield different results (Johnson et al., 1999; Strand & Johnson, 1996). However, the results of the current research did not indicate any effect of the speaker‘s perceived ethnicity on the accentedness, comprehensibility, and intelligibility for neither male nor female speakers. This means that non-native English listeners were equally unaffected by the speaker's ethnicity regardless of the gender of the speaker.

It is possible that gender does not interact with the speaker‘s ethnicity, and hence it does not affect the process of speech perception differently when the speaker is, for instance, an Asian-looking female rather than an Asian-looking male or a Caucasian-looking female. While it might have appeared like the specific choice of female speakers could have attributed to the effect of ethnic bias reported in the previous research, the results of the current research challenge this idea.

6.4 The Correlation between the IAT effect and Accentedness, Comprehensibility, and Intelligibility

Finally, this research also evaluated the relation between the listeners‘ D scores (IAT effect) and their ratings of accentedness, comprehensibility as well as their performance on the transcription task (intelligibility). The data were analyzed separately for the East Asian guises and for the Caucasian guises. No significant correlation was found between the D score and the accentedness, comprehensibility, and intelligibility for the Caucasian face data. Similarly, there was no significant

correlation between the D scores and accentedness, comprehensibility, and intelligibility for the Asian face data. However, it should be mentioned that the correlation between the listeners' D scores and accentedness in the *experimental* condition was close to the significance level ($r(28)$=-0.34, $p = 0.06$). It was a moderate negative correlation, which suggests that listeners who showed stronger "American = Caucasian" association tended to rate native English voice presented with an Asian face as more accented.

The fact that the strength of the "American = Caucasian" association for listeners in the Asian face condition was moderately (but not significantly) correlated with their ratings of the Asian guise may suggest again that the link between the social and linguistic information does exist even for non-native listeners, however, it is not as strong as for native listeners. Therefore the social information did not affect the linguistic information to a greater extent. In order to confirm this hypothesis, one would also have to investigate whether native English listeners *will* be affected by the speaker's perceived ethnicity in the same research design, when rating the same native English utterances. If native English listeners were affected by the speaker's perceived ethnicity this would provide evidence for the different nature of the link between social and linguistic information. Alternatively, one should also take into consideration that a large number of comparisons might have led to a nearly significant effect.

6.5 Limitations of the Current Study

One of the limitations of the current research are listeners-related factors. The participants in the current study were recruited mostly from the undergraduate students at the University of Tokyo, which makes them all young and well educated. They were also more likely to attend language schools in order to prepare for the

entrance exams. Moreover, about 83% of the participants have travelled abroad on a least one occasion. While recruiting participants from a specific group may help to control for certain factors, such as their background or general experience, it also makes the results of the current experiment less generalizable to a larger population.

In addition to the above-mentioned, the results of this experiment might have been influenced by the research design. This study followed the design employed in Babel and Russell (2015), which reported an effect of ethnic bias on both accentedness and comprehensibility. However, unlike in Babel and Russell, the stimuli for the current study were clear speech and not speech in noise as such might have been even more challenging for native Japanese listeners. Furthermore, while Babel and Russell employed a within-subject design, the current study implemented a between-subject design. While the within-subject design has certain limitations, such as the participants noticing the purpose of the experiment (or the fact that the same voice has been presented with different faces), a between-subject design makes it more difficult to compare between the groups (as participants rating Asian guises and participants rating Caucasian guises are *not* the same participants). This problem was partially addressed by including the *baseline* condition, which was the same for all participants. However, in order to confirm that non-native listeners are indeed less influenced by socioindexical cues than native listeners seem to be in the process of speech perception, one should also conduct the exact same study with native English listeners. Should an effect of ethnicity be found for native English listeners in the same design, this could potentially provide stronger evidence for the theory that while native listeners are affected by social information in the socially-weighted speech perception process non-native listeners are not.

Finally, one more limitation of the current study is the lack of non-native English speech. Including a "nativeness" continuum might have brought more variance to

the accentedness ratings with listeners using the whole scale more evenly. It is also possible, for instance, that Japanese listeners may transcribe non-native English utterances more accurately when presented with an Asian guise than when presented with a Caucasian guise. On the other hand, it is possible that they would have more problems when transcribing native English utterances when presented with an Asian face than when presented with a Caucasian face. If a difference like this was observed, this could provide additional evidence for the experience-based model of speech perception.

6.6 Future Directions

One possible future study would be to include both native and non-native listeners. This study is the first to address the effect of ethnic bias on speech perception by non-native listeners. However, in order to better understand the role which socioindexical cues play in the process of speech perception and the link between social and linguistic information, both native and non-native listeners should be included into the same research design. Including both native and non-native listeners would not only lead to better generalizability of the findings but also to a deeper understanding of how socioindexicality affects speech perception in the view of a socially-weighted speech perception model.

Furthermore, to the best of my knowledge, almost all previous studies investigating the effect of ethnicity (East Asian vs. Caucasian) on the speech perception, concentrated on the speech perception of native and non-native *English* utterances (Babel & Mellesmoen, 2019; Babel & Russell, 2015; de Weers, 2019; Gnevsheva, 2018; McGowan, 2011, 2015; Rubin, 1992; Rubin & Smith, 1990). Hence, a question arises whether the same effect would also be present for a

language like Japanese when evaluated by the Japanese native speakers. The number of foreigners in Japan has been gradually growing (Osumi, 2019) while the Japanese government embraced foreign workforce (Toshihiro, 2019). Moreover, more and more foreigners each year takes the Japanese Language Proficiency Test (Japanese Language Proficiency Test, 2018). With this ongoing change in Japanese demographics, it seems important to investigate whether the ethnicity of the speaker may affect the accentedness, comprehensibility, or intelligibility of their Japanese speech. Thus, it would be interesting to evaluate whether the same effect of ethnic bias will be present for native Japanese listeners presented with native and non-native Japanese utterances accompanied by either East Asian or Caucasian guises. If such an effect would also be found for native Japanese listeners evaluating native Japanese utterances, then this could provide additional support for the socially-weighted model and yet again increase the generalizability of the findings. Furthermore, if presenting non-native Japanese speech with a Caucasian face would have led to *enhanced* intelligibility, this could potentially provide new evidence against the RLS and in favor of an experience-based models in a context broader than simply native English speech, in particular native American English speech.

Finally, because of the potential *demand characteristic* effect, it seems also important to measure comprehensibility in an *online* rather than *offline* task. Using a Likert scale is a practice employed in many research (Munro, 2018; Munro & Derwing, 1995a) which does not depend on, for instance, the speed of the computer or the internet which may cause some delays that are crucial for other types of measurement such as the response time. However, it is an *offline* task, where the listener gets a reasonable time to process the utterance and to *think* about their answer. Hence, the outcome may be what the listener *decided* he or she heard, not what he or she actually *perceived*. An *online* task offers an alternative for

comprehensibility ratings by presenting the listeners with a True/False statement and measuring the response time (as executed in de Weers, 2019). Hence, using an *online* task in a design incorporating both native and non-native listeners may give us a better understanding of how difficult it was for the listeners to process the utterance in the given condition where longer response time would mean more difficulties with processing and hence worse comprehensibility.

Chapter 7

Conclusion

Speech perception is, by its nature, a multimodal process involving the integration of both linguistic and social information that are being linked together in order to encode the intended message along with information pertaining to the person who uttered it. In other words, listeners use both these factors in order to map a highly variable acoustic input onto a mental representation. When the social information matches the linguistic information as per listener's expectations (e.g., Caucasian face is paired with a native English utterance) the speech can be processed more easily (Babel & Mellesmoen, 2019; McGowan, 2015). However, any mismatch between these two factors can alter the listener‘s perception of the utterance (Babel & Russell, 2015; Gnevsheva, 2018).

The current study contributes to this discussion by evaluating how social information, in particular, the speaker‘s perceived ethnicity, may affect speech perception by *non-native* English listeners. More specifically, it investigated whether Japanese native speakers, who come from monolingual families, will be affected by the speaker‘s ethnicity when rating accentedness and comprehensibility of native English utterances. Furthermore, the current study also evaluated whether ethnic

bias will affect the actual intelligibility of the native English stimuli.

In addition, the effect of ethnic bias was also investigated in relation to the type of stimuli, pictures of the speakers presented with audio files vs. videos of the speakers. Finally, the gender of the speaker was also taken into consideration as it seems to be an informative socioindexical cue (Johnson et al., 1999; Strand & Johnson, 1996), yet most previous research on the effect of ethnic bias on speech perception by native English listeners employed only female speakers (e.g., de Weers, 2019; McGowan, 2011, 2015; Rubin, 1992; Rubin & Smith, 1990).

This link between social and linguistic information appears to be strong for native English listeners listening to native and non-native English speech. For instance, native English listeners may have *less* difficulties in understanding Chinese-accented English utterances when they are presented with an East Asian face (McGowan, 2015). Here, social information provides additional information about the speaker and their speech, which facilitates the processing of the message. On the other hand, native English listeners may have *more* trouble understanding native English utterances when presented with an East Asian guise (Babel & Mellesmoen, 2019; Babel & Russell, 2015). Additionally, they may also rate native English speech as more accented when presented with an East Asian face than when presented with a Caucasian face (Babel & Russell, 2015; Rubin, 1992).

Contrary to what was demonstrated for the native English listeners, Japanese listener's in the current study were *not* affected by the speaker's perceived ethnicity when rating the accentedness and comprehensibility of native English speech. Moreover, speaker's ethnicity did *not* have any effect on listeners' intelligibility. These results indicate that non-native English listeners in the current study were not affected by speaker's ethnicity when listening to native English speech.

Furthermore, there was no significant interaction between the type of stimuli

(pictures of the speakers presented with audio files vs. videos) and speaker's ethnicity. This suggests that native Japanese listeners rated accentedness and comprehensibility of native English utterances as well as transcribed these utterances in a comparable way regardless of whether the ethnicity of the speaker was operationalized with picture or video stimuli.

In addition, an interaction between the speaker's ethnicity and speaker's gender was also investigated as previous studies tended to include mainly female speakers (de Weers, 2019; McGowan, 2015; Rubin, 1992; Rubin & Smith, 1990). However, this interaction was also not significant, indicating that there was no effect of speaker's perceived ethnicity on speech perception for neither female nor male speakers.

Two competing theories endeavored to explain the effect of ethnicity on speech perception by native English listeners: the Reverse Linguistic Stereotyping (RLS) and the experience-based models. The RLS is a theoretical framework developed by Rubin and his colleagues and first introduced in detail in Kang and Rubin (2009). The RLS assumes that listeners hold a negative bias against Asian-looking English speakers. Because of this bias listeners may "hear" a non-native accent in native English speech. Furthermore, the bias could also cause poorer intelligibility of native English utterances when presented with an Asian face than when presented with a Caucasian face as listeners who have a *negative* bias towards East Asian-looking English speakers may intentionally *choose* to pay *less* attention to the utterance, which will lead to a communicative breakdown (Kang & Rubin, 2009; Lippi-Green, 2012; Rubin, 1992).

On the other hand, the experience-based approach, which is derived from exemplar theory (Foulkes, 2010; Foulkes & Hay, 2015), assumes that episodic traces are being stored in memory in order to be activated when presented with a consistent social category (e.g., speaker's ethnicity). This idea was incorporated by

Sumner and colleagues into the socially-weighted speech perception model introduced in Sumner et al. (2014). In a model like this, an utterance is being parsed into multiple social and linguistic information, which interact with each other in the process of social weighting. This means that listeners‘ perception of speech is being shaped by their past experiences and societal stereotypes.

Since native English listeners usually associate being American (and hence possibly a native English speaker) with being Caucasian (Babel & Russell, 2015; Devos & Banaji, 2005; Gnevsheva, 2018; Yi et al., 2013, 2014) presenting native English speech with an Asian face may for some listeners create an incongruent condition. This would lead to a mismatch effect, which would then affect speech perception. The socially-weighted speech perception model goes even further providing also explanation for the positive effects in the literature, such as native English listeners having *less* problems understanding Chinese-accented English speech when it is presented with an Asian face (McGowan, 2015). The reason for this effect is the congruent condition created by matching Chinese-accented English utterances with an Asian face. Hence, when listeners‘ expectations match the actual linguistic signal, it is easier for them to process the spoken utterance.

While this research does not rule out the RLS, I argued that experience-based models, or more specifically the socially-weighted speech perception model, may offer a possibly better explanation that can account for both the results of the current experiment and the results of previous studies. The experience-based approach suggests that listeners will rely on their personal experience when incorporating social information to the process of speech perception. Hence, the null effect of speaker‘s ethnicity observed in the current study could suggest that the Japanese participants simply had different experiences, that is, they had experience interacting with both East Asian-looking and Caucasian-looking native English

speakers. This idea can be partially supported by the fact that almost all listeners in the current study were recruited at the University of Tokyo, Komaba Campus, where in 2015 there was about the same number of students from English speaking countries in Asia, for instance the Philippines or Singapore, as from the English speaking countries from other parts of the world, for instance, the US or the UK (The University of Tokyo, n.d.). This idea that experience would shape the speech perception of the Japanese listeners in the current experiment is supported by the socially-weighted model of speech perception (Sumner et al., 2014).

Alternatively, the lack of significant results in terms of the effect of the speaker‘s ethnicity in the current study could suggest that non-native speakers may rely on the socioindexical information, in particular, the ethnicity of the speaker much *less* than native speakers do. This would create a need to refine the socially-weighted model of speech perception as if the socioindexical information is linked together with linguistic information in the socially-weighted process of speech perception, then it is possible that this link may not be strong enough for non-native listeners, at least listeners in the current study. Hence, they may not make any assumptions about the speaker and the way he or she would speak, that is, they may not expect native English speech from a Caucasian speaker and non-native English speech from an Asian speaker. To confirm this theory, more research is required with native English speakers included in the same research design in order to assure that the null effect of the speaker‘s ethnicity observed in the current study was due to a difference between native and non-native listeners and not due to the research design. Should an effect of speaker's ethnicity be observed for native English listeners, this would suggest that non-native listeners are not affected by social information, at least by the speaker's ethnicity, to the same extent as native listeners. A finding like this would also suggest a need to develop a new *refined* model of socially-weighted speech perception, which would

account equally for both native and non-native listeners.

There are almost three decades of research regarding the effect of the speaker‘s perceived ethnicity on speech perception by native listeners. Hence, it would then be unwise to generalize the results of this study to *all* non-native listeners and certainly research including non-native listeners from other countries is needed in order to gain a deeper understanding of how and when socioindexical cues may affect the speech perception by non-native listeners. Moreover, it would be equally interesting to see whether native Japanese listeners would act in the same way as native English listeners seem to when evaluating native Japanese utterances presented with an Asian guise and a Caucasian guise. Should there be an effect of the speaker‘s ethnicity for native Japanese speech evaluated by native Japanese listeners, it would suggest that native Japanese listeners rely on the socioindexical cues in the same way as native English listeners do when listening to their native language. This would potentially provide evidence in favor of the socially-weighted speech perception model and also provide more generalizability to it as the same effect of social information would be observed for a language other than English.

References

Babel, M., & Mellesmoen, G. (2019). Perceptual adaptation to stereotyped accents in audio-visual speech. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th international congress of phonetic sciences* (pp. 1044–1048). Melbourne, Australia: Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Babel, M., & Russell, J. (2015). Expectations and speech intelligibility. *The Journal of the Acoustical Society of America*, *137*(5), 2823–2833. Retrieved from `https://doi.org/10.1121/1.4919317`

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*. Retrieved from `https://doi.org/10.3389/fpsyg.2013.00328`

Barton, K. (2018). Mumin: Multi-model inference [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=MuMIn` (R package version 1.42.1)

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bayard, D., & Green, J. A. (2005). Evaluating English accents worldwide. *Te Reo*, *48*, 21–28.

Bayard, D., Weatherall, A., Gallois, C., & Pittam, J. (2001). Pax Americana?: Accent

attitudinal evaluations in New Zealand, Australia, and America. *Journal of Sociolinguistics*, *5*(1), 22–49.

Bell, A. (1997). The phonetics of fish and chips in New Zealand: Marking national and ethnic identities. *English World-Wide*, *18*(2), 243–270.

Boersma, P., & Weenink, D. (2010). *Praat: Doing phonetics by computer* [Computer Program]. Version 6.0.16. Retrieved from `http://www.praat.org/`

Bouavichith, D. A. (2019). The role of socioindexical expectation in the perception of gay male speech. In *Proceedings of the 19th international congress of phonetic sciences* (p. 1029-1033).

Bundgaard-Nielsen, R. L., Best, C. T., Kroos, C., & Tyler, M. D. (2011). Second language learners vocabulary expansion is associated with improved second language vowel intelligibility. *Applied Psycholinguistics*, *33*(3), 643–664.

Clopper, C. G., & Pisoni, D. B. (2006). The nationwide speech project: A new corpus of american english dialects. *Speech Communication*, *48*(6), 633–644. Retrieved from `https://doi.org/10.1016/j.specom.2005.09.010`

Crowther, D., Trofimovich, P., & Isaacs, T. (2016). Linguistic dimensions of second language accent and comprehensibility. *Journal of Second Language Pronunciation*, *2*(2), 160–182.

Davis, D. R. (2010). Standardized English: The history of the earlier circles. In A. Kirkpatrick (Ed.), *The routledge handbook of world Englishes* (pp. 17–36). New York: Routledge.

Denes, P. B., & Pinson, E. N. (1973). *The speech chain: The physics and biology of spoken language*. Anchor. Retrieved from `https://www.xarg.org/ref/a/0385042388/`

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *19*(1), 1—16.

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, *42*(04), 476.

Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning*, *48*(3), 393–410.

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*(4), 655–679.

Devos, T., & Banaji, R. M. (2005). American = White? *Journal of Personality and Social Psychology*, *88*(3), 447–466.

de Weers, N. (2019, August). The interaction between (un)expected speaker ethnicity and accent combinations on response times. Paper presented at New Sounds 2019, Tokyo.

Drager, K. (2010). Sociophonetic variation in speech perception. *Language and Linguistics Compass*, *4*(7), 473–480.

Drager, K. (2011). Speaker age and vowel perception. *Language and Speech*, *54*(1), 99–121.

Flege, J. E. (1984). The detection of French accent by American listeners. *The Journal of the Acoustical Society of America*, *76*(3), 692–707.

Foulkes, P. (2010). Exploring social-indexical knowledge: A long past but a short history. *Laboratory Phonology*, *1*(1), 5–39.

Foulkes, P., & Hay, J. B. (2015). The emergence of sociophonetic structure. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence* (pp. 292–313). New York: John Wiley & Sons, Inc. Retrieved from `https://doi.org/10.1002/9781118346136.ch13`

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (Second ed.). Thousand Oaks CA: Sage. Retrieved from `http://socserv.socsci.mcmaster`

.ca/jfox/Books/Companion

Gnevsheva, K. (2018). The expectation mismatch effect in accentedness perception of Asian and Caucasian non-native speakers of English. *Linguistics*, *56*(3), 581–598. Retrieved from https://doi.org/10.1515/ling-2018-0006

Greenwald, A. G., Banaji, M., & Nosek, B. (1998). *Project implicit.* Retrieved 2019-01-30, from https://implicit.harvard.edu/implicit/

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216.

Hansen Edwards, J. G., Zampini, M. L., & Cunningham, C. (2018). The accentedness, comprehensibility, and intelligibility of Asian Englishes. *World Englishes*, *37*(4), 538–557. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/weng.12344

Hanulíková, A. (2018). The effect of perceived ethnicity on spoken text comprehension under clear and adverse listening conditions. *Linguistics Vanguard*, *4*(1), 1–9. Retrieved from https://doi.org/10.1515/lingvan-2017-0029

Harrison, G. (2014). Accent and 'othering' in the workplace. In J. M. Levis & A. Moyer (Eds.), *Social dynamics in second language accent* (pp. 255–272). Boston: DE GRUYTER. Retrieved from https://doi.org/10.1515/9781614511762.255

Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics*, *48*(4), 865–892. Retrieved from https://doi.org/10.1515/ling.2010.027

Hay, J., Nolan, A., & Drager, K. (2006). From fush to feesh: Exemplar priming

in speech perception. *The Linguistic Review*, *23*(3), 351–379. Retrieved from https://doi.org/10.1515/tlr.2006.014

Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, *34*(4), 458–484.

Hayes-Harb, R., Smith, B. L., Bent, T., & Bradlow, A. R. (2008). The interlanguage speech intelligibility benefit for native speakers of Mandarin: Production and perception of English word-final voicing contrasts. *Journal of Phonetics*, *36*(4), 664–679.

Heisig, J. P., & Schaeffer, M. (2018). *Why you should always include a random slope for the lower-level variable involved in a cross-level interaction.* OSF. Retrieved from osf.io/mqu7z

Hosoda, M., Nguyen, L. T., & Stone-Romero, E. F. (2012). The effect of Hispanic accents on employment decisions. *Journal of Managerial Psychology*, *27*(4), 347–364. Retrieved from https://doi.org/10.1108/02683941211220162

Hosseinzadeh, N. M., Kambuziya, A. K. Z., & Shariati, M. (2015). British and American phonetic varieties. *Journal of Language Teaching and Research*, *6*(3), 647–655.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*(2), 135–159.

Isaacs, T., & Trofimovich, P. (2011). Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech. *Applied Psycholinguistics*, *32*(01), 113–140.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. Retrieved from https://doi.org/10.1016/j.jml.2007.11

.007

Japanese Language Proficiency Test. (2018). *JLPT in charts.* Retrieved 2020-03-27, from https://jlpt.jp/e/statistics/index.html

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, *34*(4), 485–499. Retrieved from https://doi.org/10.1016/j.wocn.2005.08.004

Johnson, K., Strand, E. A., & DImperio, M. (1999). Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*(4), 359–384. Retrieved from https://doi.org/10.1006/jpho.1999.0100

JTB Tourism Research Consulting Co. (2020). *Japanese outbound tourists statistics.* Retrieved 2020-03-20, from https://www.tourism.jp/en/tourism-database/stats/outbound/

Jułkowska, I. A., & Cebrian, J. (2015). Effects of listener factors and stimulus properties on the intelligibility, comprehensibility and accentedness of L2 speech. *Journal of Second Language Pronunciation*, *1*(2), 211–237.

Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: the english language in the outer circle. In R. Quirk & H. G. Widdowson (Eds.), *English in the world: Teaching and learning the language and literatures* (pp. 11–30). Cambridge: Cambridge University Press.

Kachru, B. B. (1992). Teaching world Englishes. In B. B. Kachru (Ed.), *The other tongue: English across cultures* (pp. 355–365). Urbana: University of Illinois Press.

Kang, O., & Rubin, D. L. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology*, *28*(4), 441–456. Retrieved from https://doi.org/10.1177/0261927x09341950

Kassambara, A. (2018). ggpubr: 'ggplot2' based publication ready plots [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=ggpubr (R package version 0.2)

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, Comprehensibility, and Accentedness of L2 Speech: The Role of Listener Experience and Semantic Context. *Canadian Modern Language Review*, *64*(3), 459–489.

Khan, A., & Alzobidy, S. A. M. (2018). Vowel variation between American English and British English. *International Journal of English Linguistics*, *9*(1), 350–356.

Kiesling, S. F. (2008). English in australia and new zealand. In B. Kachru, Y. Kachru, & C. Nelson (Eds.), *The handbook of world Englishes* (pp. 74–89). John Wiley Sons, Ltd.

Kim, J. (2016). Perceptual associations between words and speaker age. *Laboratory Phonology*, *7*(1), 18. Retrieved from https://doi.org/10.5334/labphon.33

Kim, J., & Drager, K. (2018). Rapid influence of word-talker associations on lexical access. *Topics in Cognitive Science*, *10*(4), 775–786. Retrieved from https://doi.org/10.1111/tops.12351

Kinzler, K. D., & DeJesus, J. M. (2013). Northern = smart and Southern = nice: The development of accent attitudes in the United States. *Quarterly Journal of Experimental Psychology*, *66*(6), 1146–1158. Retrieved from https://doi.org/10.1080/17470218.2012.731695

Kircher, R. (2015). The matched-guise technique. In Z. Hua (Ed.), *Research methods in intercultural communication* (pp. 196–211). John Wiley & Sons, Inc. Retrieved from https://doi.org/10.1002/9781119166283.ch13

Kleinschmidt, D. F., Weatherholtz, K., & Jaeger, T. F. (2018). Sociolinguistic

perception as inference under uncertainty. *Topics in Cognitive Science*, *10*(4), 818–834. Retrieved from https://doi.org/10.1111/tops.12331

Koops, C., Gentry, E., & Pantos, A. (2008). The effect of perceived speaker age on the perception of PIN and PEN vowels in Houston, Texas. *University of Pennsylvania Working Papers in Linguistics*, *14*(2), 12.

Kretzschmar, W. A. J. (2010). The development of Standard American English. In A. Kirkpatrick (Ed.), *The routledge handbook of world Englishes* (pp. 96–112). New York: Routledge.

Kubota, R., & Fujimoto, D. (2013). Racialized native speakers: Voices of Japanese American English language professionals. In S. A. Houghton & D. J. Rivers (Eds.), *Native-speakerism in japan: Intergroup dynamics in foreign language education* (pp. 196–206). Bristol: Multilingual Matters.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.

Labov, W. (1998). *The three dialects of English.* San Diego: Academic Press.

Lambert, W. E., Hodgson, R. C., Gardner, R. C., & Fillenbaum, S. (1960). Evaluational reactions to spoken languages. *The Journal of Abnormal and Social Psychology*, *60*(1), 44–51. Retrieved from https://doi.org/10.1037/h0044430

Lev-Ari, S., & Keysar, B. (2010). Why dont we believe non-native speakers? the influence of accent on credibility. *Journal of Experimental Social Psychology*, *46*(6), 1093–1096. Retrieved from https://doi.org/10.1016/j.jesp.2010.05.025

Levis, J. M. (2006). Pronunciation and the assessment of spoken language. In *Spoken English, tesol and applied linguistics* (pp. 245–270). Palgrave Macmillan UK.

Levon, E. (2007). Sexuality in context: Variation and the sociolinguistic perception of identity. *Language in Society*, *36*(04), 533–554. Retrieved from https://doi.org/10.1017/s0047404507070431

Lippi-Green, R. (2012). *English with an accent: Language, ideology and discrimination in the United States* (2nd ed.). London; New York: Routledge.

Luke, S. G. (2016). Evaluating significance in linear mixed-effects models in r. *Behavior Research Methods*, *49*(4), 1494–1502. Retrieved from https://doi.org/10.3758/s13428-016-0809-y

M. Helen Southwood, James E. Flege. (1999). Scaling foreign accent: Direct magnitude estimation versus interval scaling. *Clinical Linguistics & Phonetics*, *13*(5), 335–349.

Ma, D., Correll, J., & Wittenbrink, B. (2015). The chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*(47), 1122–1135.

Major, R. C. (2007). Identifying a foreign accent in an unfamiliar language. *Studies in Second Language Acquisition*, *29*(04), 539–556. Retrieved from https://doi.org/10.1017/s0272263107070428

Massaro, D. W. (2002). Multimodal speech perception: A paradigm for speech science. In *Text, speech and language technology* (pp. 45–71). Springer Netherlands. Retrieved from https://doi.org/10.1007/978-94-017-2367-1_4

McCambridge, J., de Bruin, M., & Witton, J. (2012). The effects of demand characteristics on research participant behaviours in non-laboratory settings: A systematic review. *PLoS ONE*, *7*(6), 1–6. Retrieved from https://doi.org/10.1371/journal.pone.0039116

McGowan, K. B. (2011). *The Role of Socioindexical Expectation in Speech Perception* (Unpublished doctoral dissertation). The University of Michigan.

McGowan, K. B. (2015). Social Expectation Improves Speech Perception in Noise. *Language and Speech*, *58*(4), 502–521.

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

Meade, A. W. (2009). FreeIAT: An Open-Source Program to Administer the Implicit Association Test. *Applied Psychological Measurement*, *33*(8), 643-643.

Munro, M. J. (2003). A primer on accent discrimination in the canadian context. *TESL Canada Journal*, *20*(2), 38–51. Retrieved from https://doi.org/10.18806/tesl.v20i2.947

Munro, M. J. (2018). Dimensions of pronunciation. In O. Kang, R. Thomson, & J. Murphy (Eds.), *The routledge handbook of contemporary English pronunciation* (pp. 413–431). New York: Routledge.

Munro, M. J., & Derwing, T. M. (1995a). Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Language Learning*, *45*(1), 73–97.

Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, *38*(3), 289–306. Retrieved from https://doi.org/10.1177/002383099503800305

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *49*, 285–310. Retrieved from https://doi.org/10.1111/0023-8333.49.s1.8

Munro, M. J., Derwing, T. M., & Burgess, C. S. (2003). The Detection of Foreign Accent in Backwards Speech. In M. J. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of the 15th international congress of phonetic sciences* (pp. 535–538).

Munson, B., & Babel, M. (2007). Loose lips and silver tongues, or, projecting sexual orientation through speech. *Language and Linguistics Compass*, *1*(5), 416–449. Retrieved from https://doi.org/10.1111/j.1749-818x.2007.00028.x

Munson, B., McDonald, E. C., DeBoe, N. L., & White, A. R. (2006). The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech. *Journal of Phonetics*, *34*(2), 202–240.

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface*, *14*(134), 1–11. Retrieved from https://doi.org/10.1098/rsif.2017.0213

Nakamichi, A., Jogan, A., Usami, M., & Erickson, D. (2003). Perception by native and non-native listeners of vocal emotion in a bilingual movie. *Gifu City Women's College Research Bulletin*, *52*, 87–91. Retrieved from http://www.gifu-cwc.ac.jp/tosyo/kiyo/52/zenbun52/Perception_erickson.pdf

Nelson, C. (1982). Intelligibility and non-native varieties of English. In B. B. Kachru (Ed.), *The other tongue: English across cultures* (pp. 58–73). Urbana: University of Illinois Press.

Niedzielski, N. (1999). The Effect of Social Information on the Perception of Sociolinguistic Variables. *Journal of Language and Social Psychology*, *18*(1), 62–85.

Nieuwenhuis, R., Te Grotenhuis, M., & Pelzer, B. (2012). influence.me: Tools for detecting influential data in mixed effects models. *R Journal*, *4*(2), 38–47.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 101–115.

Osumi, M. (2019, March). Number of foreign residents in japan rose 6.6% in 2018, while number of overstayers grew almost twice as much, government data shows. *The Japan Times*. Retrieved 2020-03-20, from https://www.japantimes.co.jp/news/2019/03/22/national/number-foreign-residents-japan-rose-6-6-2018-number-overstayers-grew-almost-twice-much-government-data-shows/#.XoBx03X7S01

Pantos, A. J., & Perkins, A. W. (2012). Measuring implicit and explicit attitudes toward foreign accented speech. *Journal of Language and Social Psychology*, *32*(1), 3–20. Retrieved from https://doi.org/10.1177/0261927x12463005

Purnell, T., Idsardi, W., & Baugh, J. (1999). Perceptual and Phonetic Experiments on American English Dialect Identification. *Journal of Language and Social Psychology*, *18*(1), 10–30.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Raisler, I. (1976). Differential response to the same message delivered by native and foreign speakers. *Foreign Language Annals*, *9*(3), 256–259. Retrieved from https://doi.org/10.1111/j.1944-9720.1976.tb03221.x

Revelle, W. (2018). psych: Procedures for psychological, psychometric, and personality research [Computer software manual]. Evanston, Illinois. Retrieved from https://CRAN.R-project.org/package=psych (R package version 1.8.12)

Rosenthal, R., & and, R. L. R. (2009). *Artifacts in behavioral research*. Oxford University Press. Retrieved from https://doi.org/10.1093/acprof:oso/9780195385540.001.0001

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of

nonnative English-speaking teaching assistants. *Research in Higher Education*, *33*(4), 511–531.

Rubin, D. L., Ainsworth, S., Cho, E., Turk, D., & Winn, L. (1999). Are greek letter social organizations a factor in undergraduates perceptions of international instructors? *International Journal of Intercultural Relations*, *23*(1), 1–12. Retrieved from https://doi.org/10.1016/s0147-1767(98)00023-6

Rubin, D. L., Coles, V. B., & Barnett, J. T. (2015). Linguistic stereotyping in older adults' perceptions of health care aides. *Health Communication*, *31*(7), 911–916. Retrieved from https://doi.org/10.1080/10410236.2015.1007549

Rubin, D. L., Healy, P., Gardiner, T. C., Zath, R. C., & Moore, C. P. (1997). Nonnative physicians as message sources: Effects of accent and ethnicity on patients responses to AIDS prevention counseling. *Health Communication*, *9*(4), 351–368. Retrieved from https://doi.org/10.1207/s15327027hc0904_4

Rubin, D. L., & Smith, K. A. (1990). Effects of accent, ethnicity, and lecture topic on undergraduates perceptions of nonnative English-speaking teaching assistants. *International Journal of Intercultural Relations*, *14*(3), 337–353. Retrieved from https://doi.org/10.1016/0147-1767(90)90019-s

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, *37*(2), 217—240.

Sheppard, B. E., Elliott, N. C., & Baese-Berk, M. M. (2017). Comprehensibility and intelligibility of international student speech: Comparing perceptions of university EAP instructors and content faculty. *Journal of English for Academic Purposes*, *26*, 42–51.

Shuy, R. W. (1969). The relevance of sociolinguistics for language teaching. *TESOL Quarterly*, *3*(1), 13. Retrieved from https://doi.org/10.2307/3586038

Statistics Bureau of Japan. (2015). *Population and households of Japan.* Retrieved 2020-02-27, from https://www.stat.go.jp/english/data/kokusei/2015/final_en/final_en.html#Summary

Strand, E. A. (1999). Uncovering the Role of Gender Stereotypes in Speech Perception. *Journal of Language and Social Psychology*, *18*(1), 86–100.

Strand, E. A., & Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. In D. Gibbon (Ed.), *Natural language processing and speech technology, results of the 3rd KONVENS conference, Bielefeld, Germany, October 1996* (pp. 14–26). de Gruyter.

Sumner, M., Kim, S. K., King, E., & McGowan, K. B. (2014). The socially weighted encoding of spoken words: a dual-route approach to speech perception. *Frontiers in Psychology*, *4*, 1–13. Retrieved from https://doi.org/10.3389/fpsyg.2013.01015

The Brookings Institution. (2019). *Us foreign-born gains are smallest in a decade, except in Trump states.* Retrieved 2020-02-25, from https://www.brookings.edu/blog/the-avenue/2019/10/01/us-foreign-born-gains-are-smallest-in-a-decade-except-in-trump-states/

The University of Tokyo. (n.d.). *The University of Tokyo, Komaba prospectus 2015-2016.* Retrieved 2020-02-28, from http://www.c.u-tokyo.ac.jp/info/about/booklet-gazette/prospectus/2015/prospectus_2015_E.pdf

Thomas, E. R. (2002). Sociophonetic Applications of Speech Perception Experiments. *American Speech*, *77*(2), 115–147. Retrieved from https://doi.org/10.1215/00031283-77-2-115

Thomson, R. (2017). Measurement of accentedness, intelligibility, and comprehensibility. In *Assessment in second language pronunciation* (pp. 11–29). Routledge. Retrieved from https://doi.org/10.4324/9781315170756-2

Toshihiro, M. (2019). *Japan's historic immigration reform: A work in progress.* Retrieved 2019-02-15, from https://www.nippon.com/en/in-depth/a06004/japan%E2%80%99s-historic-immigration-reform-a-work-in-progress.html

Trent, S. A. (1995). Voice quality: Listener identification of African-American versus caucasian speakers. *The Journal of the Acoustical Society of America*, *98*(5), 2936–2936. Retrieved from https://doi.org/10.1121/1.414099

Tucker, G. R., & Lambert, W. E. (1969). White and negro listeners reactions to various American-English dialects. *Social Forces*, *47*(4), 463–468. Retrieved from https://doi.org/10.2307/2574535

U.S. Census Bureau. (2019). *Quick facts: Population estimates.* Retrieved 2020-03-01, from https://www.census.gov/quickfacts/fact/table/US/PST045219#PST045219

van Bezooijen, R. (1988). The relative importance of pronunciation, prosody, and voice quality for the attribution of social status and personality characteristics. In *Language attitudes in the Dutch language area* (pp. 85–104). De Gruyter Mouton. Retrieved from https://doi.org/10.1515/9783110857856.85

Walker, A., & Hay, J. (2011). Congruence between 'word age' and 'voice age' facilitates lexical access. *Laboratory Phonology*, *2*(1), 219–237. Retrieved from https://doi.org/10.1515/labphon.2011.007

Wells, J. C. (1982). *Accents of English I: An introduction.* Cambridge University Press. Retrieved from https://www.xarg.org/ref/a/0521297192/

Wester, M., & Mayo, C. (2014). Accent rating by native and non-native listeners. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP).* IEEE. Retrieved from https://doi.org/10.1109/icassp.2014.6855098

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. *CoRR*, *abs/1308.5499*. Retrieved from http://arxiv.org/abs/1308.5499

Yi, H.-G., Phelps, J. E. B., Smiljanic, R., & Chandrasekaran, B. (2013). Reduced efficiency of audiovisual integration for nonnative speech. *The Journal of the Acoustical Society of America*, *134*(5), EL387–EL393. Retrieved from https://doi.org/10.1121/1.4822320

Yi, H.-G., Smiljanic, R., & Chandrasekaran, B. (2014). The neural processing of foreign-accented speech and its relationship to listener bias. *Frontiers in Human Neuroscience*, *8*.

Zahn, C. J., & Hopper, R. (1985). Measuring language attitudes: The speech evaluation instrument. *Journal of Language and Social Psychology*, *4*(2), 113–123. Retrieved from https://doi.org/10.1177/0261927x8500400203

Zheng, Y., & Samuel, A. G. (2017). Does seeing an Asian face make speech sound more accented? *Attention, Perception, & Psychophysics*, *79*(6), 1841–1859. Retrieved from https://doi.org/10.3758/s13414-017-1329-2

Appendices

Appendix A

Pictures Used for the IAT

Figure A.1: Asian female 01

Figure A.2: Asian female 02

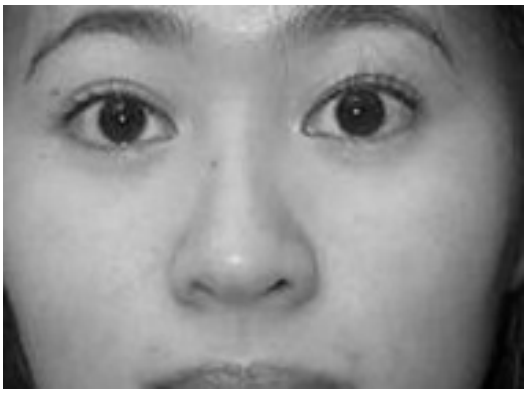


Figure A.3: Asian female 03

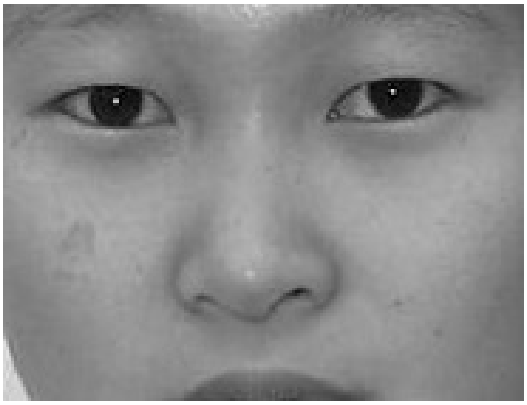


Figure A.4: Asian female 04

Figure A.5: Asian female 05

Figure A.6: Asian male 01

Figure A.7: Asian male 02

Figure A.8: Asian male 03

Figure A.9: Asian male 04

Figure A.10: Asian male 05

Figure A.11: Caucasian female 01

Figure A.12: Caucasian female 02

Figure A.13: Caucasian female 03

Figure A.14: Caucasian female 04

Figure A.15: Caucasian female 05

Figure A.16: Caucasian male 01

Figure A.17: Caucasian male 02

Figure A.18: Caucasian male 03

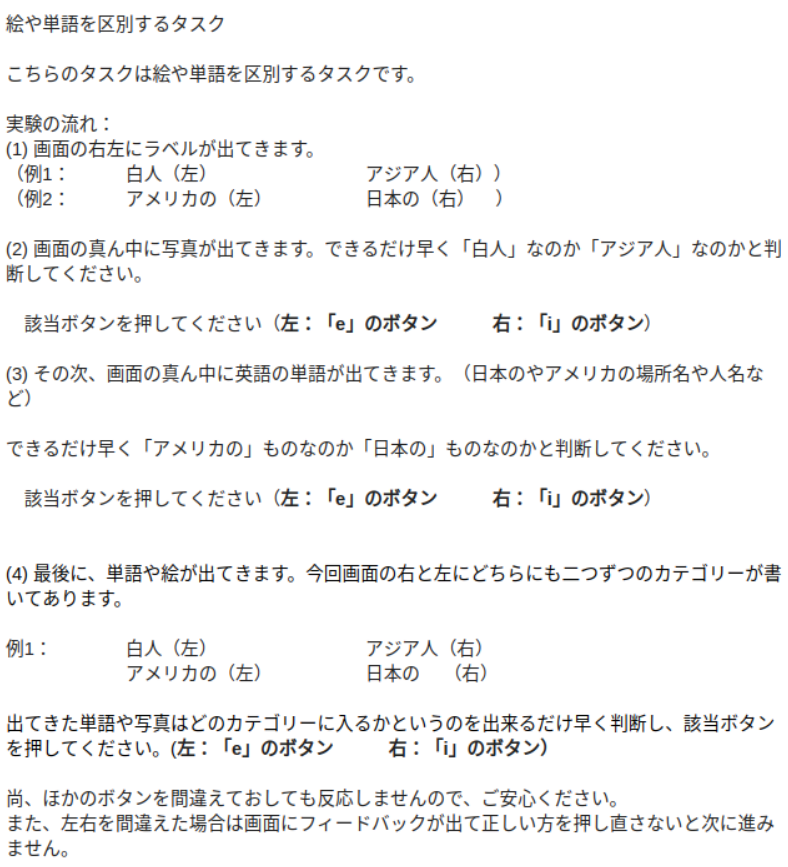
Figure A.19: Caucasian male 04

Figure A.20: Caucasian male 05

Appendix B

Instructions for the IAT

Figure B.1: Japanese instructions for Task 1. Besides the written instructions participants received lengthy oral instructions.



絵や単語を区別するタスク

こちらのタスクは絵や単語を区別するタスクです。

実験の流れ：
(1) 画面の右左にラベルが出てきます。
（例1：　　白人（左）　　アジア人（右））
（例2：　　アメリカの（左）　　日本の（右）　）

(2) 画面の真ん中に写真が出てきます。できるだけ早く「白人」なのか「アジア人」なのかと判断してください。

該当ボタンを押してください（**左：「e」のボタン　　右：「i」のボタン**）

(3) その次、画面の真ん中に英語の単語が出てきます。（日本のやアメリカの場所名や人名など）

できるだけ早く「アメリカの」ものなのか「日本の」ものなのかと判断してください。

該当ボタンを押してください（**左：「e」のボタン　　右：「i」のボタン**）

(4) 最後に、単語や絵が出てきます。今回画面の右と左にどちらにも二つずつのカテゴリーが書いてあります。

例1：　　白人（左）　　アジア人（右）
　　アメリカの（左）　　日本の　（右）

出てきた単語や写真はどのカテゴリーに入るかというのを出来るだけ早く判断し、該当ボタンを押してください。**(左：「e」のボタン　　右：「i」のボタン)**

尚、ほかのボタンを間違えておしても反応しませんので、ご安心ください。
また、左右を間違えた場合は画面にフィードバックが出て正しい方を押し直さないと次に進みません。

Figure B.2: English translation of the Japanese instructions for Task 1.

Classification of Pictures and Words

In this task, you will be asked to classify pictures and words.

Experiment flow:

(1) On both sides of the screen, you will see a label.

（Example 1: Caucasian（on the left)　　Asian (on the right))
（Example 2: American（on the left)　　Japanese (on the right))

(2) You will see a picture in the middle of the screen. Decide as fast as you can whether it was an Asian person or a Caucasian person.

Press the button associated with the correct answer.（**LEFT: "e" button　RIGHT: "i" button**)

(3) Next, you will see a word in English in the middle of the screen. (There will be people or places etc. which will be either American or Japanese)

Try to decide whether the word is something American or Japanese as fast as you can.

Press the button associated with the correct answer.（**LEFT: "e" button　RIGHT: "i" button**)

(4) Finally, you will see both words and pictures appearing. This time on each side of the screen there will be two categories.

Example 1：	Caucasian（on the left)	Asian（on the right)
	American（on the left)	Japanese (on the right)

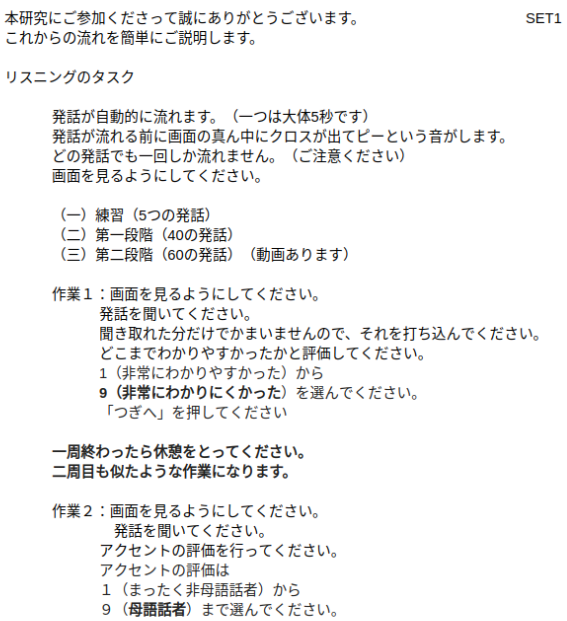
Try to determinate, as fast as possible, to which category belongs the item you see on the screen and press the button associated with the correct answer.（**LEFT: "e" button　RIGHT: "i" button**)

You don't have to worry about pressing any other button by mistake since there will be no reaction to that.
If you press the wrong button feedback will appear on the screen. You will have to press the correct button in order to advance to the next trial.

Appendix C

Instructions for the Perception Experiment

Figure C.1: An example of Japanese instructions for Task 2 for both video groups (baseline first). Besides the written instructions participants received lengthy oral instructions delivered in Japanese.



本研究にご参加くださって誠にありがとうございます。　SET1
これからの流れを簡単にご説明します。

リスニングのタスク

発話が自動的に流れます。（一つは大体5秒です）
発話が流れる前に画面の真ん中にクロスが出てピーという音がします。
どの発話でも一回しか流れません。（ご注意ください）
画面を見るようにしてください。

（一）練習（5つの発話）
（二）第一段階（40の発話）
（三）第二段階（60の発話）（動画あります）

作業１：画面を見るようにしてください。
発話を聞いてください。
聞き取れた分だけでかまいませんので、それを打ち込んでください。
どこまでわかりやすかったかと評価してください。
1（非常にわかりやすかった）から
9（非常にわかりにくかった）を選んでください。
「つぎへ」を押してください

一周終わったら休憩をとってください。
二周目も似たような作業になります。

作業２：画面を見るようにしてください。
発話を聞いてください。
アクセントの評価を行ってください。
アクセントの評価は
1（まったく非母語話者）から
9（**母語話者**）まで選んでください。

Figure C.2: English translation of the example of instruction for Task 2 for both video groups.

Thank you for participating in this research. SET1
Here I will explain the experiment flow.

Listening Task

You will hear audio played automatically（one is about 5 sec）
A cross will appear in the middle of the screen and there will be a beep sound before each utterance.

Every utterance will be played only once.（Pay attention, please）
Make sure you look at the screen.

（1）Practice（5 files）
（2）First Stage（40 files）
（3）Second Stage（60 files） （You will see a video of the speaker）

TASK 1：Look at the screen.
Listen to the speakers.
Type what you hear.
Rate each speaker in terms of how easy it was to understand them.
Choose a number from
1（very easy to understand）to
9（very difficult to understand）
Press “next”

Please, take a break when you finish the first round.
In the second round, you will be asked to complete a similar task.

TASK 2：Look at the screen.
Listen to the speakers.
Rate the accent of each speaker.
Choose from
1（non-native speaker）to
9（**native speaker**）

Appendix D

Pictures of Asian and Caucasian Guises Used in the Perception Experiment

Figure D.1: Asian female 01

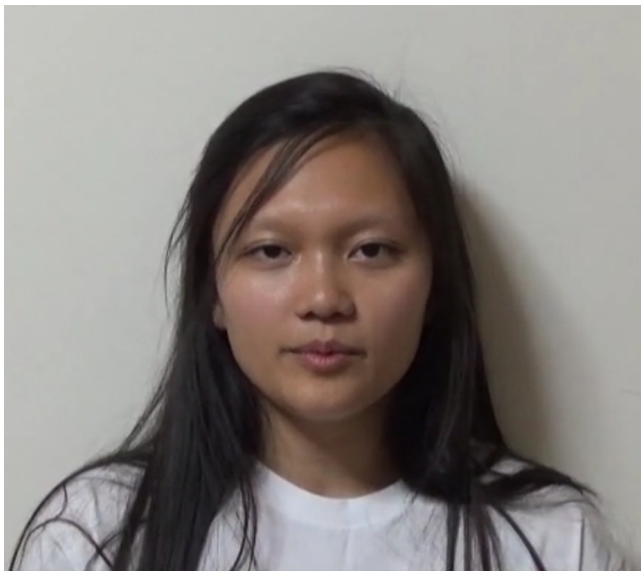


Figure D.2: Asian female 02

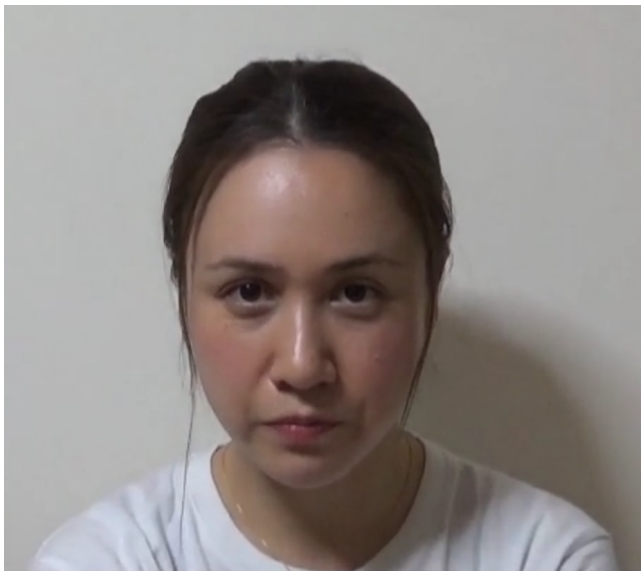


Figure D.3: Asian female 03

Figure D.4: Asian male 01

Figure D.5: Asian male 02

Figure D.6: Asian male 03

Figure D.7: Caucasian female 01

Figure D.8: Caucasian female 02

Figure D.9: Caucasian female 03

Figure D.10: Caucasian male 01

Figure D.11: Caucasian male 02

Figure D.12: Caucasian male 03

Appendix E

Sentences Used in the Perception Experiment

Table E.1: Transcription of recordings for Female 1

Recording number	Transcription
1	I'm still pretty full from eating that last night and it's already almost lunchtime.
2	And she opens the door, the phone starts ringing and her cat is sleeping on the floor.
3	But then she gets home and I guess there is a sudden rainstorm.
4	So she's in the park, she's running and she has a smile on her face.
5	So she gets up and at 7 o'clock she is eating breakfast.
6	Because it's early and she doesn't wanna get out of bed but the sun is up.
7	And her alarm goes off and she doesn't look very happy.
8	She is holding on to a cup of some of the other ingredients.
9	He looks really confident that he knows what he is doing and that he is gonna make a delicious cake.
10	He is pouring a lot of flour or baking soda or something into the bowl.

Table E.2: Transcription of recordings for Female 2

Recording number	Transcription
1	I think it's very clean and safe and everyone is very friendly.
2	Because I don't like sweating when I'm just lying down and doing nothing.
3	I really like my home, the weather is very nice and it's always sunny.
4	And it also rains a lot which is even worse because it's hot rain.
5	He was crying but he admitted to her that he forgot to send out her invitations.
6	She asked her friend why no one was there. Her friend started crying.
7	She was confused. She thought she was very popular but apparently she wasn't.
8	He comes home and has lunch. After eating lunch he decides to take a shower.
9	The next day is a Saturday so he wakes up a little bit later.
10	He watches TV for thirty minutes. At 9 o'clock in the evening he goes to bed.

Table E.3: Transcription of recordings for Female 3

Recording number	Transcription
1	I mean we do that too but we also do a lot of moving and sweating.
2	I've been doing yoga for about 15 years and I do many different styles of yoga.
3	I'm a scientist and this is my day job and I've been doing this for about 20 years.
4	As he approaches the car and he looks inside this tiny little car is full of a dozen clowns.
5	The car has stopped and pulled over and the policeman has got off of his motorcycle.
6	And he's got a radio and he noticed the speeding car and he is calling it in.
7	We have a car and it's racing down the road very very fast.
8	They both look very upset and scared, and I'm sure she is saying that she was right.
9	And he didn't do the recipe correctly. The cake is really coming out of the oven.
10	And the guy is definitely worried, and the girl is definitely worried, and they realised that she was right.

Table E.4: Transcription of recordings for Female 4

Recording number	Transcription
1	And say so much through video and say so much visually.
2	Because she loves presents and her friends gave her a ton of presents, as well.
3	Pictures are worth thousand words but videos are even worth more.
4	I got a laptop from my mom and then I started doing some video editing.
5	Finally her parents give her a cake and she gets so excited.
6	Friends and family are all around her, surrounding her with tons of presents.
7	She finally gets a cake from her family and she blows up the candles.
8	He's super nervous. he goes up on stage and forgets what to say.
9	His teachers are telling him to start, everyone starts laughing.
10	And it was just so much fun and I couldn't stop doing it.

Table E.5: Transcription of recordings for Female 5

Recording number	Transcription
1	And all of the sudden there her grandma is, standing over the stove.
2	But it seems that everything she says just makes things worse.
3	An old man and his wife live on a farm with a happy pig and a happy dog.
4	The grandmother kisses the little girl on the cheek even though it embarrasses her.
5	She runs through the front door and cannot wait to see what's on the stove.
6	Her grandmother is always in the kitchen and she can't wait to see what she is cooking next.
7	She wants the men to just be friends but all they do is fight.
8	He begins to cry he is holding his face in horror.
9	The man with black hair and glasses looks away. He is scared for her.
10	She has beautiful long hair that is curled and very light.

Table E.6: Transcription of recordings for Male 1

Recording number	Transcription
1	In particular probably because I've never owned a TV.
2	I see a man and a woman sitting at the table, talking.
3	While I love video games I hate to admit that I haven't really been a gamer...
4	You can see the man placing the contents of his mixing bowl into an oven.
5	The man is trying to keep up to his dog who is now chasing the cat.
6	While the dog runs around the tree dragging his master behind him.
7	And I came to Japan originally to continue studying Aikido.
8	My current goal is to stay here in Saitama for the next six years.
9	My other hobbies are music, I play guitar, piano, sing...
10	She looks concerned as the man mixes ingredients into a large mixing bowl.

Table E.7: Transcription of recordings for Male 2

Recording number	Transcription
1	There are two people in what looks like a beach.
2	So growing up I was pretty fortunate. I had a lot of activities and friends.
3	He is holding an umbrella and signaling to a boy standing next to him that there's rain.
4	There's a classroom of fishes and octopus. The octopus is a teacher.
5	And octopus is writing the letters A, B, C on a white board.
6	The man has opened his umbrella and looks quite dry underneath the umbrella.
7	The man is inside the pond or ocean and he is signaling to the woman.
8	The man and the woman are now swimming inside the ocean and they seemed to have spotted something.
9	The man and the woman seem a little bit surprised to see this.
10	The boy seems kind of nervous next to the man.

Table E.8: Transcription of recordings for Male 3

Recording number	Transcription
1	This guy is not really in control of his dog.
2	So we shouldn't really be surprised if something happens later on in the story.
3	I'm glad I'm able to walk in front of my master.
4	Why don't you guys go back to the ocean I hear there is a lot of really good tuna out there.
5	This is the way he passes the time while he's riding the bicycle.
6	Because when these balloons go up into the air they cause environmental damage.
7	The powers of cats are not very well understood and so this cat is actually flying into the air.
8	The dog has run around the master and has decided to tie him to the tree.
9	The dog should chase the cat and he is encouraging the dog to chase the cat.
10	To be outside walking my dog because that shows I am a kind of member of society.

Table E.9: Transcription of recordings for Male 4

Recording number	Transcription
1	And now suddenly the older man is crying and he has his hands up to his face.
2	And the, the two men do not look happy. The woman looks OK and she is saying something.
3	One of the men is quite a bit older than her, perhaps even her father.
4	Not sure what's happening or why but these people are not happy even though they are at a party.
5	And the older man has his face in his hands covering his eyes.
6	So he is now kind of flying away and he is saying something.
7	And the boy is standing next to him with his jacket over his head.
8	One person might be, might be a teenager, looks kind of young.
9	Which I imagine is quite funny although neither of them look happy.
10	And it looks like it's beginning to rain. One man is wearing a hat. He has a coat.

Table E.10: Transcription of recordings for Male 5

Recording number	Transcription
1	His life was too much like a made up story.
2	This was a very important day in class and he knew he was going to have to turn the homework in.
3	The poor, young man he was so upset now. He had lost all of his homework.
4	Suddenly a dog, a large dog jumped out and hit him, and attacked him.
5	So she steps out onto his coat and to their surprise...
6	She wants to walk in that direction but she is concerned about getting wet.
7	He is a really nice man, so he's what we call gentleman.
8	So that she can walk across and be dry and not get dirty.
9	He invites her to walk across and she is very pleased.
10	And she suddenly falls through and drops into this.

Appendix F

Code Used for the Analysis

```
model.accent <- lmer(accentedness ~
                     group * condition * gender +
                     English_level +
                     (1+condition*gender|subject) +
                     (1|speaker/item),
                     weights = words_total,
                     control=lmerControl(optimizer="bobyqa",
                         optCtrl=list(maxfun=1e4)),
                     data=data)


model.comprehensibility <- lmer(comprehensibility ~
                     group * condition * gender +
                     English_level+
                     (1+gender*condition|subject) +
                     (1|speaker/item),
                     weights = words_total,
                     control=lmerControl(optimizer="bobyqa",
                         optCtrl=list(maxfun=1e4)),
                     data=data)
```

```
model.intelligibility <- glmer(intelligibility ~
                        group * condition * gender +
                        English_level +
                        (1+condition|subject) +
                        (1|item),
                        family = binomial(link = "logit"),
                        weights=words_total,
                        control=glmerControl(optimizer="bobyqa",
                            optCtrl=list(maxfun=2e4)),
                        data=data,)
```