The Perception and Cognition of Racialized Voices

by

Tedra James
Class of 2018

A thesis submitted to the
faculty of Wesleyan University
in partial fulfillment of the requirements for the
Degree of Bachelor of Arts
with Departmental Honors in Neuroscience & Behavior and African American Studies

Middletown, Connecticut  April 17, 2018
Abstract

The Perception and Cognition of Racialized Voices

Human listeners are able to determine racial and/or ethnic background from a voice based on acoustic and linguistic cues alone (Perrachione et al, 2010; Reaser et al, 2004). However, the mechanisms that underlie this determination remain poorly understood. Here we report a series of three experiments that each contribute to understanding what properties of the voice guide a listener to decide whether a talker is black or white. A speech corpus of continuous, read sentences is used in each experiment as experimental stimuli. The first experiment involves a norming study conducted online whereby subjects (n = 200) listened to each sample and completed two-alternative forced choice (2AFC) tests to identify the race of the speaker. The stimuli identified correctly above chance level were then used in the second experiment, where we conduct a series of comparisons between the acoustic qualities of black and white speakers. Finally, the third experiment involves changing the salient cues identified in experiment 2. After applying manipulations to the audio, we test whether or not listener categorization changes in another two sets of online studies. Results indicate that the ability of listeners to identify speaker race is affected after manipulating acoustic properties of duration, F0, and the harmonics of the sound. These features are likely just some of the cues that influence a listener’s determination of race from the voice. Taken together, we observe that physical qualities in a voice factors into discerning speaker race. This study has implications for understanding and reducing racial prejudice.

Keywords: Voice Perception, Person Perception/Identification, Social Cognition, Race/Ethnicity
Acknowledgements

A sincere thank you to my readers for choosing to spend your time reading this work! I truly appreciate it.

Thank you to all the MIND Labbies who have supported me and the project from its conception. A special shout out to Max Bouvagnet, who signed on to the project early and played an instrumental role in its various stages of development.

To my friendship crew-- you not only tolerated my craziness during this challenging process and continued to encourage me, helped edit and listened to me whine and complain, you drove me forward. You also chose to incorporate me into your in-group, and for that I will be forever grateful.

To Professor Loui-- you have taught me a wider array of knowledge and skill that I would never have imagined applying to your lab my sophomore year. Thank you for your guidance, patience, and mentorship and helping to foster my personal and intellectual growth.

Finally, and most importantly: To my parents, from whom I adopted the intellectual curiosity and tenacity to pursue and complete this thesis. I love you so much.
# Table of Contents

## CHAPTER 1: GENERAL INTRODUCTION

1.1 Social Identity Extraction from Diverse Sources ............................. 5  
1.2 Racial Identity Extraction from the Voice ..................................... 7  
1.3 Purpose of Study ................................................................. 9

## CHAPTER 2: EXPERIMENT 1: ASSESSING THE PERCEPTION OF RACIALIZATION IN THE VOICE ........................................ 14  
2.1 Introduction .............................................................................. 14  
2.2 Methods ................................................................................. 14  
2.2.1 Procedure ............................................................................ 16  
2.3 Subjects .................................................................................. 17  
2.4 Results .................................................................................... 18  
2.5 Conclusions ............................................................................ 21

## CHAPTER 3: EXPERIMENT 2: ACOUSTIC ANALYSES OF RACIALIZATION ........................................ 22  
3.1 Introduction .............................................................................. 22  
3.2 Methods ................................................................................. 23  
3.3 Results .................................................................................... 24  
3.4 Discussion and Conclusion ......................................................... 26

## CHAPTER 4: EXPERIMENT 3: “REVERSING” (OR ELIMINATING) RACIALIZATION ........................................................................... 27  
4.1 Introduction .............................................................................. 27  
4.2 Methods ................................................................................. 27  
4.3 Subjects .................................................................................. 28  
4.4 Procedure .............................................................................. 29  
4.4 Results .................................................................................... 29  
4.5 Conclusion .............................................................................. 31

## CHAPTER 5: GENERAL DISCUSSION .................................................. 31  
5.1 Future Directions .................................................................... 33  
5.2 Implications ........................................................................... 34

## REFERENCES .................................................................................. 35

## APPENDIX ....................................................................................... 37
CHAPTER 1: General Introduction

1.1 Social Identity Extraction from Diverse Sources

People have a remarkable ability that allows them to reliably extract social identity information from another person’s voice. Stimuli as short as a single word (e.g., “hello”) can provide one with a general idea of a speaker’s age, sex, affect, and race (Ambady & Rosenthal, 1992; Ambady & Rosenthal, 1997; Purnell, Baugh, & Idsardi, 1999; Willis & Todorov, 2006). Strong evidence for this idea comes from Nalini Ambady’s studies on “thin slices” of behavior that demonstrate that stimuli as short as 30 seconds are sufficient for the accurate and/or highly-agreed upon derivation of social meaning (1992; 1997). The high consistency and accuracy with which raters judge a talker’s affect, competence, bias from teachers and judges, and how well a patient will respond to therapy from thin slices of behavior in the face, voice, and body movement suggests that there may be essential properties in many social behaviors that help guide certain perceptive social assessments.

Social impressions gathered from thin slices of behavior are not only highly-agreed upon between subjects, but they are also unlikely to change following greater exposure to the behavior (Willis & Todorov, 2006). Alexander Todorov’s studies reveal that exposure to faces lasting as short as 100ms elicits unwavering convictions about a person’s trustworthiness, competence, attractiveness, likability, and aggressiveness (2006). Todorov and Willis’s joint study led them to conclude that greater length of exposure does not affect social judgement; it only increases one’s confidence in one’s judgement. The fact that social comprehension is achieved so rapidly suggests that humans have a highly efficient neural mechanism to achieve it.
The capacity to gather certain types of social information is likely an adaptive mechanism. Evolutionary benefits for judging another’s level of trustworthiness, aggressiveness, and even competence includes helping one to decide who to work with and who to avoid to increase the likelihood of their survival. The tendency to make social decisions about which people we prefer occurs at a very young age (Kinzler, Dupoux, & Spelke, 2007). Infants (5-6 months), 10 month-olds, and five year-olds prefer to interact with speakers of their native language. They choose to play with social beings that look like their mothers (Kinzler et al., 2007). Early onset of social discrimination between familiar and unfamiliar voices suggests that this tendency is more likely to be innate than learned. Innate behaviors and thought processes compel us to think about the adaptive advantages of extracting social information.

Our social discrimination skills are likely utilized to construct in- and out-groups so that we can make informed judgements about who to include and who to avoid in our social networks. One neural structure that has consistently been implicated to have a role in this in- out-group construction process is the amygdala (Man, Ames, Todorov, & Cunningham, 2016). Neural imaging and lesion studies suggest that the amygdala is responsible for both threat detection and appetitive behaviors (Man, et al 2016). Authors of the chapter, “Amygdala Tuning Towards Self and Others”, synthesize evidence and theory to convincingly posit that the amygdala processes emotionally-charged and motivationally relevant stimuli to guide our decision-making brain regions on what to do (Man, et al 2016). Our brain
structures encode who may be friendly versus who may be threatening to help us survive.

1.2 Racial Identity Extraction from the Voice

Here we ask how and why racial identity might similarly be deduced from thin slices of the voice; especially because race is usually thought of as a social construct that depends on visual attributes. What might be the essential acoustic properties in certain voices that guide a listener’s perceptual delineation of race?

Understanding how race might be discerned from the voice requires a rudimentary understanding of speech production. Sound energy that makes up speech sounds come from modifications of air (Belin, Fecteau, & Bedard, 2004). The lungs serve as the primary source of a stream of air that gets sent up through the different parts of the vocal tract before exiting the mouth (Belin et al., 2004; Latinus & Belin, 2011). Within the vocal tract are several anatomical structures that open and close, change shape, and compress the air, effectively causing the air stream to behave and be heard in unique ways (Latinus & Belin, 2011). The manner by which the vocal tract modifies sound is periodic, producing a sinusoidal-like waveform with a characteristic structure (Latinus & Belin, 2011). This characteristic frequency of sound energy is referred to as fundamental frequency (F0), and is the largest component of the sound wave (in amplitude). F0 is often perceived as pitch (Latinus & Belin, 2011). Other components of the sound energy are referred to as the harmonics of the F0, which are concentrations of acoustical energy that can be viewed in the spectral decomposition of sound. Harmonics are also called formants; they are produced from the vocal tract structures that cause the sound to resonate at
certain frequencies (Latinus & Belin, 2011). The frequency of formants and their proximity to one another correspond to what we perceive as vowels and vowel quality (Peterson & Barney, 1952). Now that we know what features a voice has, we return to the question of race. Despite personal experience testifying to the ability to predict race from voice alone, the crucial factors for determining race from voice are not well studied.

Thomas and Reaser review a series of studies that investigate the auditory cues that may be used to categorize the race of a voice (2004). Many of the studies they analyze demonstrate that subjects are able to discriminate race with high accuracy (Thomas & Reaser, 2004). Notwithstanding, studies that compare isolated black populations (e.g., Hyde County, North Carolina and Appalachia (also N.C.)) and white speakers demonstrate that listeners do not always have high success (Thomas & Reaser, 2004). This indicates that sounding black as a distinct vocal quality from sounding white is not necessarily universal nor absolute. Further, the cues that seem to significantly distinguish white and black speakers are inconsistent across the studies.

Other studies that attempt to resolve racialized vocal features include Julie Walton’s 1992 dissertation and a study by Tyler Perrachione, Joan Y. Chiao, and Patrick C. M. Wong (2010). Walton’s dissertation indicates that for black and white speakers who were correctly categorized at a rate above 58%, the harmonic-to-noise ratio (dynamics of formants) and changes in the sound amplitude (also called shimmer) distinguish black and white speakers (1992). Perrachione, Chiao, and Wong demonstrate high consistency and accuracy rates for race categorization from voice.
Their results indicate that a unique formant vowel space (F1 frequency versus F2), longer voicing onset time (the duration of the beginning part of a word), and other linguistic features (i.e., devoicing final consonants and not pronouncing r’s) are the distinctive features of their black speakers (2010). A lack of shared cues with their group of typically-sounding black speakers led listeners to incorrectly categorize a group of black speakers as white, suggesting that the presence or absence of certain cues influenced their categorization (Perrachione, Chiao, & Wong, 2010).

Because the cues that aid race determination from voice are neither well studied nor highly agreed-upon, we should assume that it is a combination of several different cues that is responsible for leading a listener to categorize a speaker as one race or another. Previous literature demonstrates that there are multiple acoustic cues that the brain integrates to form its impression of race from voice, but of current interest is determining the ones that are the most influential of race perception. Of the different properties of sound picked up from a stimulus, which make it to the higher-order neural processes involved in person perception?

1.3 Purpose of Study

Understanding the mechanisms of race perception has the potential to combat the racist practices that continue to characterize the U.S. We argue that determining race from voice may often be harmless, but it can also be used to discriminate against racial and ethnic minority groups; either implicitly or consciously. Anti-blackness is a pervasive factor in the U.S. Despite the apparent success of the Civil Rights Movement in the 1960s, stark racial inequality characterizes the state. Black Americans continue to be disproportionately affected by poverty, live in segregated
communities, and be killed by health disparities and law enforcement (PewResearchCenter, 2016). All this in spite of the enactment of protective legislature such as the Fair Housing Act, Title IX, and the Civil Rights Act of 1964. One may ask how inequality is able to persist despite such legislative and social impediments.

John Dovidio and Samuel Gaertner offer a description of a modern, unconscious, yet harmful type of racism that they call “aversive racism” (2005). Unlike overt racism, aversive racism tends to be enacted by well-meaning white liberals who claim to have no bias against black people, but harbor unconscious anti-black sentiment (Gaertner & Dovidio, 2005). Dovidio and Gartner’s studies expose aversive racist tendencies in a number of ways: white liberals offered help to black motorists (phone actors) less often than white motorists and ended the call prematurely in response to a black voice on the phone more frequently (1973). Other studies evidence implicit bias via increased eye-blinks towards black people, looking away more often, and a decreased willingness to help black people during scenarios that do not directly implicate the participant (Hodson, Gaertner, & Dovidio 2002; Gaertner & Dovidio 1977). Aversive racism is particularly harmful because of its implicit nature. As Dovidio and Gaertner acknowledge, people are likely unaware of the subtle ways they enact anti-blackness. Therefore, it is somewhat more difficult to addressed.

A more specific type of aversive racism is “linguistic profiling”, a term coined by John Baugh in 2002 to describe the process in which individuals determine the race of a speaker from their voice alone (e.g., over the phone) and proceed to
discriminate against that individual based on their judgement. Baugh’s theory is based on a set of four experiments comparing levels of discrimination against individuals phoning landlords for housing opportunities (Baugh, 2002). One experiment featured Baugh himself acting as different speakers by mimicking “Standard English”, African American Vernacular English, and Chicano (Mexican)-American English accents (Purnell et al., 1999). Another of his experiments featured hired actors of each of those races and confirmed the previous results (Purnell et al., 1999). In both cases, African American and Chicano voices are called back at a lower rate solely based on the dialect in their voice. Subsequently, Baugh conducted acoustic analyses on the voice’s harmonics and harmonic-to-noise ratio when actors uttered the word “hello”—this word alone was sufficient for categorizing speaker race above chance(Purnell et al., 1999). Despite the failure to find any statistically significantly different measures for any cues except the second formant on the phoneme /e/, Baugh’s studies corroborate that there is indeed an ability to categorize race from only the voice and that this ability may have negative consequences for racial and ethnic minorities. Again, much investigation is left to be done regarding confirming the cues that influence process of race categorization.

John Baugh’s study suggests that labeling a voice a certain race may be possible with different ethnic groups. Given the U.S.’s current political climate, evaluating prejudice and discrimination from multiple modalities, angles, and frameworks has the potential to foster greater understanding and cooperation between the diverse groups that constitute America.
The aim of the current study is to examine how the voice is racialized. Fostering greater understanding of the acoustic features potentially involved in the categorization process will help to combat practices such as aversive racism and linguistic profiling.

1.4 General Framework

Our theoretical framework is based off that used by Perrachione, et al (2010). Their model integrates existing models of face perception—consistent with previous literature on voice perception—to understand the process of person perception. The hypothesis is based off their 2007 study and Pascal Belin’s model of voice perception (Belin et al., 2004; Perrachoine & Wong, 2007). It posits that “accurate identification/categorization of a person relies [both] on a listeners’ knowledge of
linguistic patterns (and idiosyncrasies) used by the speaker, [as well as] the perception of the physical acoustic properties of the voice” (Perrachoine & Wong, 2007). In other words, our brains integrate the structural acoustic properties of a voice with our stored memories of linguistic patterns to facilitate voice recognition and subsequently person identification. Importantly, Perrachione et al’s model rejects the idea that racialized voice perception is dependent on structural characteristics of the voice alone. Social factors including exposure to speech and dialect are thought to influence one’s ability to recognize an individual and/or categorize their race.

These ideas arise from results demonstrating that participant familiarity with the speech and lexical characteristics a speaker uses enhances their ability to recognize and identify that speaker (Perrachoine & Wong, 2007). The 2010 study measured and compared vocal tract length, leading to the conclusion that structural anatomy of the vocal tract had no effect on a listeners’ ability to identify the race of a voice; however, listeners’ exposure to the dialectical features in speech used by racial groups (by being of the same racial group) was positively correlated with their ability to correctly identify the race of a voice, \( p < .002 \) (Perrachione et al., 2010). Final support comes from literature that asserts that hearing a talker’s speech and sinewave reconstructions (just F1, and F2, F3) of a talker is sufficient for voice recognition, implying that the integration of speech knowledge and certain aspects of vocal structure are sufficient for person identification (Sheffert, Pisoni, Fellowes, & Remez, 2002).

The step in the person identification process that involves “tagging” a voice with a certain race remains unclear. The hope is that the following study will
contribute to this understanding. The main goal is to contribute knowledge that makes sense of how linguistic profiling happens, perceptively and cognitively.

*Hypothesis:* Certain acoustic cues used to determine the race of a speaker are so crucial to racial identification that if manipulated, listeners’ perception of the speaker’s race would change.

CHAPTER 2: Experiment 1: Assessing the Perception of Racialization in the Voice

2.1 Introduction
Because race perception is typically dependent on visual attributes, what do people mean when they say that someone “sounds white” or “sounds black”? What acoustic differences are used to guide their perception? Often, people do not explicitly know. Therefore, experiment 1 seeks to establish a widely-socially-agreed upon set of “black-sounding” and “white-sounding” speakers to assess what racialized voice really is. Experiment 1 will also test whether or not a listener’s perception of speaker race is always consistent with the speaker’s self-described racial identity. If it is not, what might our understanding of vocal racialization change?

2.2 Methods
The Texas Instruments & Massachusetts Institute of Technology developed the *Acoustic-phonetic Continuous Speech Corpus* (TIMIT), a large collection of speech samples from 630 different speakers across the United States that was acquired in the late 1980’s and published in 1990 is used in the study as experimental stimuli (Garofolo et al., 1992). The corpus is racially diverse, contains read sentences
(two of which are consistent across all speakers), and includes extensive documentation. In addition to the documentation about each speakers’ hometown/the dialectical region of America in which they grew up, TIMIT documented each speaker’s race, height, sex, age, level of education, and time-aligned transcriptions of the words and phones speakers said in the accompanying audio files. We use TIMIT for the following study because of its large sample size and racial diversity. Further, there was high interest in using the shared sentences across speakers to eliminate distinct grammar and/or colloquialisms that could be factored in listener’s determination. The primary interest remains in the vocal acoustics. TIMIT’s time-aligned transcriptions was predicted to benefit impending acoustic analyses.

After sorting TIMIT’s list of speakers from 8 different dialectical regions by race, the stimulus selection process begins. Consistent dialect region across the speakers is important because of the divergent ways people from different geographic regions of the U.S. speak (Clopper, Pisoni, & de Jong, 2005). To obtain a large set of stimuli, the largest group of black speakers from the same dialectical region was selected (n = 17), which happened to be the Southern Region. All speaker codes are the speaker’s initials and had nothing to do with the date the sentences were acquired or any other potentially confounding variables. After selecting an equal number of white speakers from the same region, and ensuring to match the number of speakers of each gender, we established our speaker stimuli. For testing, we selected the 2 sentences from the corpus that every single speaker reads: 1. “She had your dark suit in greasy wash water all year” and 2. “Don’t ask me to carry an oily rag like that”. The sentences were designed by the Stanford Research Institute to include a diverse
set of phonemes that would highlight interesting features. Figures 2 and 3 depict the breakdown of testing stimuli.

**Figure 2: Schematic of TIMIT Stimuli**

![Diagram showing the breakdown of TIMIT stimuli with 34 speakers, 17 black, 17 white, 8 females, and 8 males.]

1. She had your dark suit in greasy wash water all year
2. Don’t ask me to carry an oily rag like that

**Figure 3: TIMIT Map of 8 Major U.S. Dialectical Regions**

![Map of the United States showing major dialect regions with yellow highlighting the Southern region, home of 34 TIMIT speakers.]

Figure 3: Map from Texas Instruments (Garofo, et al 1990) depicting the major dialect regions in the United States. Note: DR 8 includes speakers who moved all over, which is why it is not depicted. Locations for speakers are points on the map. Highlighted in yellow is the Southern dialect region, home of our 34 TIMIT speakers.

### 2.2.1 Procedure
The workers were first asked to disclose demographic information including their age, sex, the zip code of their hometown, and their ethnic background. They were then tasked to listen to each stimulus in order and select whether they think the voice presented belongs to a black person or white person. If unsure about the race of the speaker, MTURK workers were instructed to give their best guess. The order of the sentences was randomized using a random number generator. Each worker had 15 minutes to categorize the 68 sentence stimuli (34 speakers x 2 sentences) (See Appendix for a screenshot of the experiment).

2.3 Subjects

202 workers from the U.S. were employed on Amazon Mechanical Turk (MTURK) to categorize their impression of the race of our 34 speakers. MTURK is an online platform that allows businesses and researchers access to hundreds of thousands of individuals from 30 different countries to complete studies for pay, effectively providing large amounts of data in short timeframes. We use MTURK to test the TIMIT stimuli to gauge a sense of the social understanding of how white and black people may sound different from one another. We rejected 3 workers for not completing a majority of the survey.

Most workers ($n = 124$) self-identified as white. There were 90 females, 106 males, and two workers who indicated their biological sex as “other”. The average age of the sample is on the young side ($M = 34.01, SD = 10.56$). Most of the workers come from the Western Dialectical Region ($n = 39$), followed by the Northern and South Midland Regions ($n = 35$, each), proceeded by the North Midland region ($n = 34$), the South ($n = 26$), New England ($n = 6$), and finally, New York City ($n = 5$).
(See appendix for map of worker hometown locations and a pie chart of listener ethnicity).

2.4 Results

The workers who completed the task were generally very accurate at labeling the speaker with the race that was consistent with TIMIT’s documentation. A binomial test conducted for each of the 68 sentences revealed that listener categorization of speaker race was consistently above .5, \( p < .05 \). This indicates that more than half of listeners had similar judgements of speaker race. One exception was a black woman speaker (GDP0) whose sentence 2 had a listener-categorization rate of 50%, \( p = .777 \).

Although GDP0 was the only speaker who failed to have strong listener agreement, three other black speakers: two black males (HIT0 and DRB0)’s first sentences were categorized as white by more than half the listeners, \( p = .003 \). Interestingly, both sentences provided by a black woman (DTD0) were categorized as sounding white by more than 50% of listeners, \( p = .000 \). That there is strong agreement about black speakers sounding white is significant, but unsurprising.

White speakers tended to be categorized correctly at a higher rate (\( M = 87.7\% \), \( SD = .0681 \)) than black speakers (\( M = 72.5\% \), \( SD = .1804 \)). An independent samples t-test (equal variances not assumed) revealed that this accuracy difference between black and white speakers is statistically significant \( t(198) = -4.604, p < .001 \). Greater accuracy for identifying white voices over black voices is consistent with previous literature that implicates “white bias”; Namely, because whiteness as a racial category
is an overrepresented population in America, it is presumed that listeners, in cases of uncertainty, will categorize a speaker as “white” by default (Perrachione et al 2010).

Still, the vast majority of sentences ($n = 64$) were categorized as their documented race at a rate above chance, indicating that most listeners were not guessing, but could actually decipher speaker race. A plot depicting the average categorization (either black=1 or white=2) of each sentence is shown below. Error bars represent the 95% confidence intervals of listener ratings for each speaker sentence.

The average accuracy of 73% for categorizing black speakers was particularly affected by the aforementioned group of three black speakers that were frequently misclassified as white more than 50% of the time. These speakers evidence that just because one is black, doesn’t mean that they necessarily sound black. This group of

Figure 4: Categorization of Speaker Race by MTURK Listeners

The blue shaded region encloses speakers in Group 1 (black-sounding black speakers), the yellow shaded region encloses Group 2 speakers (white-sounding white speakers). The unshaded region includes Group 3 speakers (black speakers misidentified as white). Error bars = 95% confidence interval.
speakers might get called back for housing at a rate more similar to white speakers compared to their “black-sounding” peers because of the increased likelihood of their being mistaken as white due to the sound of their voice. Based off MTURK listeners’ categorization of HIT0, DRB0, and DTD0’s voices, they should not be used to try to assess what constitutes sounding black.

A univariate ANOVA was conducted to investigate the effect of listener ethnicity on race categorization accuracy. The data revealed a significant effect of listener race on their accuracy $F(7, 198)= 5.278, p < .001$. Though ethnic groups performed equally well when categorizing white speakers, Asian Americans performed worse while categorizing black speakers. The average accuracy for Pacific Islanders and Indigenous listeners was also lower than other ethnic groups, but their accuracy difference failed to be statistically significant. This is likely due to being too

**Figure 5: Average Accuracy While Categorizing Black Speakers, by Ethnicity**
small a sample size.

2.5 Conclusions

Based on results provided by MTURK workers, we can begin to sort the 34 TIMIT speakers into three unique groups: Group 1: Black speakers that sound black, Group 2: White speakers that sound white, and Group 3: Black speakers that don’t sound black. To clarify, Groups 1 and 2 were the speakers that were consistently categorized as sounding like their documented race more than 50% of the time, and Group 3 includes black speakers that were not consistently categorized as sounding black. Group three criterion permits black speakers categorized as sounding white above chance, even if only one of their sentences is categorized that way. Therefore, GDP0, who almost qualified for inclusion in Group 3, does not qualify for inclusion
in group three because one sentence was categorized at chance level, while the other sentence was consistently rated as sounding black. This led to the inclusion of GDP0 in Group 1. The distinction made between being a race versus being perceived as a member of a certain race (Group 1 versus 3) is crucial. Establishing speaker groups using this methodology limits subsequent analyses to voices that were consistently rated/perceived as sounding black versus white by a general population at a rate above chance. Distinguishing speakers this way will allow for the proper determination of acoustics that distinguish black-sounding vs. white-sounding voices.

CHAPTER 3: Experiment 2: Acoustic Analyses of Racialization

3.1 Introduction

Here we analyze the acoustic differences between our established groups from Experiment 1. Using a phonetic analysis software, Praat, we extract acoustic features that previous literature has indicated might be relied on for race categorization. The goal is to resolve the acoustic differences between “black-sounding” and “white-sounding” speakers. Below we show a representative waveform and the corresponding spectrogram of two male speakers.

Figure 6: Representative Spectrogram and Waveform for Representative Male Speakers Saying “Suit”
3.2 Methods

We first extract the fundamental frequency (F0) for each sentence and use the command “kill octave jumps” to eliminate pitch outliers from the resulting pitch contour. Separate pitch extraction parameters for male and female speakers are imposed; namely, male speakers have a pitch ceiling of 300 Hz and females 500 Hz due to well-documented vocal tract differences (Belin 2004). Praat used time step analysis windows of .01 seconds to obtain the respective F0 values.

Next, we extract the first 5 formants (F1-F5) of each sentence. Maximum formant values have distinct ceilings for men versus women. Therefore, formant extraction parameters were constrained to have a maximum value of 5000 Hz for men and 5500 Hz for women. The formant analysis window was .0125 seconds.

Subsequently, average duration of the sentences (in seconds) is compared between groups. Sentences were then normalized to the same length to compare their
formant frequencies over time for Groups 1 (Black) and 2 (White), by sex. Finally, TIMIT’s phoneme documentation was used to extract the average formant values for specific vowels. We validated the cardinal vowels /a/, /i/, and /u/ using the International Phonetic Association’s vowel chart website. Each speaker provided 11 vowels for analysis across the two sentences (except for DHL0, who did not pronounce the “iy” in carry).

Figure 7: Words Selected to Contribute to Vowel Space Analysis

<table>
<thead>
<tr>
<th>/u/</th>
<th>/i/</th>
<th>/a/</th>
</tr>
</thead>
<tbody>
<tr>
<td>suit</td>
<td>She</td>
<td>dark</td>
</tr>
<tr>
<td>oily</td>
<td>greasy</td>
<td>wash</td>
</tr>
<tr>
<td>year</td>
<td>me</td>
<td>water</td>
</tr>
<tr>
<td>carry</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.3 Results

The fundamental frequency of the sentences does not differ between Group 1 (Black) and Group 2 (White) speakers at a statistically significant level, as revealed by a one-way ANOVA on F0 and group $F(2,65) = .425, p = .665$. This eliminates the possibility that F0 alone is used to categorize race.

Despite the trend of the average white speaker speaking faster than the average black speaker, duration failed to be significantly different between groups at
the .05 level, as demonstrated by a one-way ANOVA on sentence duration $F(2,65) = 1.806, p = .172$.

Formant frequencies seem to differ across black and white speakers, but their differences are not consistent across gender. In general, black speakers have a lower F1 than white speakers. This different is less consistent for men, as sometimes black men have a higher F1 than white men. The pattern is less obvious for F2. Black speakers tend to have higher F2 than white speakers for some parts of Sentence 1, but there are a couple of instances where white female speakers have a higher F2, and even more instances where white male speakers have a higher F2 than black males. These inconsistencies reflect the dynamic nature of speech, how sex modulates vocal difference, and how different phonemes are more distinct across race than others. (See appendix for formants plotted over time)

The distinct patterns of formant frequency are better understood when focusing on individual vowels. Plotting vowel space in an F2-F1 axis is a common practice amongst linguistic scholars due to the rough correlation with the positionality of the vocal tract (source). The results plotted in the vowel space chart illustrate acoustic differences across vowels; Namely, black female speakers have a lower F1 and F2 for /u/ and /a/ than white females, while their articulation of /i/ shows very slight difference. For males, black speakers have a higher F1 and lower F2 when pronouncing /u/, a lower F1 and F2 when pronouncing /i/, and a very slight difference when pronouncing /a/ compared to white male speakers.

Figure 8: Average formants for the cardinal vowels depicted in figure 7 are plotted on an F2-F1 axis with F1 on the Y and F2 as vertical axis. Smaller frequencies for both are towards the upper right coorners. Black circles represent black speakers and white circles represent white speakers.
3.4 Discussion and Conclusion

The results from Experiment two suggest that there is no a simple trend that explicates racial acoustic difference. Instead, there seem to be nuanced differences elicited during distinct parts of a sentence (e.g., certain vowels over others). For example, black and white women pronounce /i/ more similarly than the way they pronounce /u/. Listeners may be relying on the acoustic properties of /u/ over /i/ to predict race from a sound source. The distinct pattern of difference across gender implies that listeners may need to switch the cues they attend to for males and females to make accurate racial judgments. The most generalized difference is the difference in duration, as our sample of black speakers tend to speak more slowly than our white speakers, despite failing to reach significance.

Results show that the speakers categorized by MTURK workers as sounding black versus white have slightly different acoustic features in their voices despite saying the same exact sentences. Our results suggest that there’s more than one acoustic cue being used when people are engaging in labeling race from voice, and further, that the cues used are likely not consistent across gender, nor across different
phonemes. The next phase of the study attempts to confirm that the salient features that define black-sounding and white-sounding speakers are truly crucial for racial categorization that is consistent with speaker race.

CHAPTER 4: Experiment 3: “Reversing” (or Eliminating) Racialization

4.1 Introduction

This final experiment seeks to validate the unique acoustic cues associated with black and white-sounding speakers that were identified in Experiment 2. Are these essential factors that guide a listener’s perception of speaker race? We attempt to test whether or not changing the cues will impact listeners’ accuracy. If eliminating or changing the cues fails to change listener accuracy, there are likely more salient cues used for race identification that we failed to identify in our acoustic analysis. If changing the cues decreases listener accuracy, it suggests we successfully identified at least some of the acoustic cues that are essential for discerning the race from voice alone.

4.2 Methods

Using MaxMSP, a visual programming language used for music and sound, filters and acoustic manipulations were created and applied to each speech sample in groups 1 and 2. First, the duration of the speech was changed by adjusting playback speed to be either 10% faster (for black voices) or 10% slower (for white voices). The rate of 10% was chosen with the following logics: the timing difference had to be noticeable enough to affect a listener’s perception of the same voice, but not so different that the speech would no longer sound naturalistic. Changing the playback speed of each sentence effectively increased the fundamental frequency of each
sentence, causing black voices to have a slightly higher F0 and white voices to have a slightly lower F0.

Next, filters that either up- or down-regulate the formant 1 and 2 frequencies for the speakers depending on their group are created and applied to the speed-adjusted sentences (show filters). Careful listening and tuning during the creation of the filters had to be a factor to prevent the voices from sounding unnatural. Notably, the filters are different for men versus women, while filters changing a voice from white to black versus changing black to white are the inverse of one another.

![Figure 9: Filter Changing Acoustics for Female Speakers](image)

After applying each filter to the original group 1 and 2 samples, they were recorded as .wav files and sent out in two new, distinct batches for MTURK testing. One batch included only the manipulated female voices, the other included only male voices. We hypothesized that the applied manipulations would yield different levels of accuracy for each speaker compared to the original batch from Experiment 1.

### 4.3 Subjects

Forty-nine workers on MTURK completed the second (female-only) batch. Of those participants, 32 identified as male and they were slightly younger than our
first cohort of workers \((M = 30.65, SD = 7.21)\). For the third (male-only) batch, 50 different workers completed the MTURK survey. These workers were in an age range more similar to the first batch of workers \((M = 33.92, SD = 10.52)\). Thirty-seven of them identified as male. Both of these batches had only one listener indicate their biological sex as “other”. Listener hometowns for these batches are plotted on the MTURK hometown map as purple squares (batch 2), and stars (batch 3), respectively (see appendix).

4.4 Procedure

The second and third batches were designed similarly to the first batch. They included the same demographic survey questions at the beginning, the worker pool was again limited to people currently living in the U.S., and stimuli were presented in a random order. Again, workers were tasked with categorizing a voice as either black or white and instructed to use their best guess in cases of uncertainty. The only significant difference is the amount of sentence stimuli they had to categorize \((n = 34\) for female voices) and \((n = 28\) for male voices).

4.4 Results

Our manipulations successfully impacted listener accuracy of race categorization for the second batch with female-only voices, and less successfully for the batch of male-only voices. T-tests for independent samples were conducted on each sentence before and after audio manipulations to compare the distribution of accuracy; most sentences by female speakers had been differently categorized, \(p < .05\). The comparison chart below includes asterix over sentences that were not categorized differently.
Manipulated female voices elicited categorizations of each speaker’s race that were brought closer to the level of chance. White speakers were categorized as sounding black more often, and vice versa. (see figure) Conversely, only one white male was judged as sounding black more often (DAS0’s first sentence) after manipulation $t(59.327) = 3.122, p = .003$, and only one black male (AHH0’s first manipulated sentence) was categorized as sounding white more often than the original $t(65.873) = 5.129, p < .001$. The rest of the listener categorizations for white speakers were largely unchanged. One particular white male, DSJ0, had a second sentence that was actually categorized as white at a higher rate post-manipulation compared to the original. Surprisingly, many black men were categorized as sounding black more often than before, including both WEM0’s sentences, LIH0’s second sentence, and EWM0’s first sentence. Further investigation is necessary to begin to understand why the data showed this result.

Figure 11
4.5 Conclusion

The purpose of the manipulations implemented is not necessarily to make a black speaker sound white or vice versa; the changes are designed to decrease the perceived racialization of each speaker’s voice. This is measured by the rate at which each speaker is categorized as either black or white by MTURK listeners. Thus, the real goal is to make the speakers sound more racially ambiguous. For women, the changes seem to have successfully done that. For men, the manipulations were less successful. The methods described in this experiment demonstrate that by manipulating the cues identified as being particularly important for racialization, people’s perception of the race of a female voice can be changed. These results have promising implications for the ability to de-racialize female voices, potentially in real-time.

CHAPTER 5: General Discussion

This study describes a set of three separate experiments that each contribute to understanding acoustic vocal difference across racial lines. Recall that the purpose of
the study was to examine how humans extract acoustic information from the voice to make categorical judgements about the race of a speaker. The first experiment confirmed that people have this ability, the second investigated the auditory features this ability might rely on, and the third tested how that ability is affected when the features are changed. Results indicate that the racialization of voice is in part defined by the first two formants of certain vowels, possibly in conjunction with F0 and duration for women, and that speakers who are less consistently categorized as being their self-identified race may share features of another racial group and/or lack features of their own. Notably, these results are applicable for Southern speakers only; These differences may likely diverge if speakers from a different dialectical region of the U.S. are the subject of study.

Potential sources of error or any possible issues with the data include:

1. Because the TIMIT Corpus includes read speech rather than conversational speech, some differences we have identified may have arisen due to reading aloud rather than natural speaking differences. Different groups of people are not necessarily used to reading speech in the same way, which could affect the manner of speaking.

2. Despite the instruction to make a best guess, some MTURK users neglected to categorize certain sentences.

3. Though different parts of the sentence contain distinct acoustic difference (e.g., different vowels have different F2-F1 space), the filters created are non-discriminatory in that they are applied over the whole sentence. This may explain the failure to see significant differences in the categorization of manipulated male
audio compared to their original files—male speakers had more nuanced differences between their F1, and especially F2 values over time.

4. Because experiment 3 does not necessarily demonstrate a reversal in race categorization, applied filters may be impacting accuracy simply due to acoustic signal being changed in general; This may not be the case, however, given the increased racialization effects for male speakers.

5.1 Future Directions

Because the project is so new, there are many directions it could take. In addition to addressing the aforementioned sources of error, we would like to perform more acoustic analyses. The next steps that come to mind are comparing change in frequency and amplitude over time (jitter and shimmer, respectively), and comparing the harmonic-to-noise ratios of each sentence between groups 1 (black speakers) and 2 (white speakers). Another possibility is to make direct comparisons between the acoustic properties of group 3 speakers (black speakers miscategorized as white) with groups 1 (black speakers consistently categorized as black) and 2 (white speakers consistently categorized as white). Importantly, we would like to tailor the filter to target specific vowels for up- or down- regulation of formant frequencies rather than applying it to the entire sentence. Testing the impact of a more targeted filter may increase success with listener categorization of male voices.

Finally, we would like to combine the study with neuroimaging studies (fMRI and EEG) to assess the neural correlates of race perception and categorization processes in real-time. The timing of racial encoding may be resolved using EEG testing methodology. It would also be interesting to assess whether or not the
structural properties of the voices might differentially affect a listener’s EEG waveform depending on the race of the speaker.

5.2 Implications

This study confirms that people can and do accurately discern race from the voice alone. Vocal acoustic features are able to be instrumentalized to discriminate against ethnic groups, supporting the theory of linguistic profiling. Our study contributes to this understanding. Recall from the introduction that social psychological theories on race are frequently linked to in- and out-group formation theory. This may make sense given the way humans may be adapted to preferentially recognize and prefer shared language. Shared language is regionally defined, as demonstrated by the map of dialectical regions of the United States (Clopper et al., 2005). The former geographic link with race likely made more sense for our ancestors to encode as threatening. Those that were around you were part of your family, community, and survival in-group. Those that came from afar were likely harder to trust. The modern age of imperialism, globalization, and colonization causes this skill to likely be overused in situations that are no longer evolutionarily beneficial or relevant.

If our study suggests that initial steps of linguistic profiling is an implicit phenomenon, it does not give racial discrimination a pass. Though the amygdala is part of the more implicit impression-forming processes of social interaction, it still recruits other functional areas of the brain when given the time and/or direction. For example, the PFC is recruited during extended exposure to racialized faces, causing a decrease in amygdala activity for other-race faces (Cunningham et al., 2004). Further,
educating people on their implicit tendencies to discriminate may influence their readiness to do it—one can only fix a problem when one knows what the problem is.

By drawing attention to the social issue of racial discrimination and proposing possible mechanisms for how it functions, we can begin to brainstorm solutions and accountability measures to combat it. One potential solution is creating a real-time filtering device with our design in mind, that could be attached to the phone so that prospective landlords and employers are unable to profile applicant voices. Another is to use the wealth of information on voice acting and singing to modulate these differences on one’s own.

Continuous increase in understanding about how we encode racial difference in different modalities is crucial to the sustained battle for equality in America. There is potential to narrow the employment, income, and housing gap between ethnic minority groups and white Americans, resulting in a more equitable society. Especially at the political moment that we find ourselves in in 2018, fostering greater tolerance in our society continues to inspire and guide this work.

References


Appendix

MTURK Listener Hometowns (Zip Codes)

Map created using ESRI ArcGIS. Batch 1 is the pool of 199 MTURK workers. Batch 2 includes listeners to female voices, and Batch 3 includes listeners of male voices.

Breakdown of the First Batch of MTURK Listeners’ Self-Indicated Ethnicity

Pie chart for ethnic identity of workers who participated in the norming study (Experiment 1) only
Average Duration of Each Sentence, by Group and Sex

Average duration depicted by sex. Black boxes represent black (Group 1) speakers and white boxes represent white (Group 2) speakers.

Average F0 for Male and Female Speakers

Average F0 depicted by sex. Black boxes represent black (Group 1) speakers, white boxes represent white (Group 2) speakers.
F2 of Male Speakers for SA1

F2 of Male Speakers for SA2

Time (seconds)

Frequency (Hz)