THE PSYCHOLOGY AND PSYCHOPHYSICS OF VOICE RECOGNITION

by

Bryan Shilowich

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(PSYCHOLOGY)
December 2019

Acknowledgments

First and foremost, I would like to give infinite gratitude and credit to my parents, Barbara and Walter Shilowich. They gave me a nurturing, cherished childhood and always encouraged my wildly varying intellectual, artistic, and athletic pursuits. This unconditional encouragement fostered a curiosity and passion for living that I genuinely take for granted in my adulthood. In this most recent chapter of my life, they have given me unparalleled emotional, financial, and all-encompassing parental support, which is to say that they have been there whenever I needed them, in whatever capacity. And for that, I can never thank them enough.

USC has provided me a remarkable doctoral education, and my dissertation would likely not have reached fruition if not for the wonderful people I've met here, both faculty and students. A very special thanks is owed to Dr. Irving Biederman, my advisor and the chair of my dissertation committee. He graciously accepted me into his lab with minimal research experience and insistently instilled in me a fine appreciation of clean, simple, and powerfully parsimonious experimental design. For that and a multitude of other things, including, above all, the immense patience, I cannot sufficiently express my gratitude. Likewise, the rest of my committee deserves recognition for being invaluable contributors to this project as well as my general well-being and success throughout the process. Dr. Jason Zevin, Dr. Toby Mintz, and Dr. Louis Goldstein each provided their own unique research perspective on my work, remarkable cooperation in keeping up with some rather urgent deadlines I unfairly imposed on them, and genuine, compassionate support and encouragement to both finish this on time and end up being quite proud of it.

My degree was funded by the Department of Psychology - a combination of the McGuigan Award (which provided two fellowship years), four summer research stipends, and ten teaching assistantships. Without these, not only my degree but also my entire life supporting it would simply not have been possible. Dr. Stan Huey and Dr. Jo Ann Farver, the department chairs, deserve special mention and appreciation for their always-available guidance and support. I'd like to thank Dr. Xiaokun Xu and Dr. Ori Amir, who were the senior members of my lab when I arrived. Their early mentorship laid the foundation for the vast majority of the technical skills I have developed as a research scientist. In fact, some lines of the analysis code used for this dissertation were directly copied from that very first code Xiaokun helped me write back in 2012, which he graciously did while busy working on his own dissertation. Dr. Leslie Berntsen gets a special shout out for her relentless positivity; it's uncanny, and those optimistic vibes will insistently reverberate in my brain any time I ever doubt myself, for the rest of my life. Everyone on this list I have already given my thanks to, many times, in person - except for one. The last but certainly not least USC person I'd like to thank is Dr. Bosco Tjan, whose intellectual prowess and cheerful, hilarious demeanor will be missed by many more than myself. I just wish I could have told him how much I admired him.

Outside the academic sphere of USC, the deep friendships I've made during my time in California are too numerous to go through individually. If we're close enough for you to be reading this, by all means you already know that our friendship means the world to me, and the past seven years of my life would not have been the same without you. That said, I'd like to thank a handful of people who have been particularly fundamental in either leading me in some way to pursue my PhD to begin with, guiding

me throughout, or a combination thereof. The top honors go to a certain Derya Kadipasaoglu, whose brilliance inspires me to do everything that I do, and to do it to the fullest; her passion for all topics in art, science, philosophy, and their infinite to-be-found connections reverberates with my own in a way that I have simply never encountered in another person. She deserves a dissertation-length ode to her being, but these measly two sentences will just have to do. My brother, Craig Shilowich, deserves similar praise. He always has been and always will be my greatest role model, in a way that he could never possibly understand. Hermanos. On the topic of role models, I'd like to thank Paul Rogers, my earliest and most profound combination of coach, teacher, mentor, and friend. He's just genuinely one of the coolest dudes I've ever met, but he also instilled in me a sense of dedication, persistence, and inspiration to be the best version of myself I can be. Much more recently and in a similar vein, I want to thank Ranjiv Perera, for always pushing me, challenging my every thought and action, and forcing me to understand myself in ways that only he could see.

Finally, I have a few more thanks, scattered throughout my interesting journey to end up here. Dr. Eve Marder, from Brandeis University, was a great neuroscience professor and provided an invaluable recommendation that presumably helped my admission here. But not only that, she was tirelessly supportive in a time when I had been frustrated by several program rejections, and she is directly responsible for my first trip to Japan. I obviously didn't go to that PhD program (OIST), but that trip fundamentally changed me as a person. So, thanks, Eve. I'd also like to thank Alex Trebek, who is not directly responsible but certainly involved in some way in the bizarre twist of fate my life took in 2012 that had me moving to California from Boston within a

week's notice. I wish him the best health, and maybe we'll meet again. I wouldn't be writing this if not for the great Kumotorisan, for showing me how to do anything I put my mind to, and I have Saihoji to thank for the clarity of the path. For similar reasons, my deepest appreciation goes to Gus Kandellis, for not only teaching me to be myself but also showing me how to stay on track, whatever the track may be. Along the way, the creative experiments of D. Laneight-Goldstein & W.M. Joel, Waters et al., and Blackmore et al. each provided inspiration at various points while working through this dissertation, guiding my thinking in their own unique ways. Lastly, I'd like to acknowledge the painting that lives in my room, Sisyphus. Conceived of and completed to fruition within precisely the same time frame, she and this present work will always be conceptual, dizygotic twins; together they serve as highly complementary, tangibly creative products of my existence from this period of my life.

Regarding roommates, my final thanks go to Alisha Brophy. I have lived with her for the past six years; in that time, she has listened to an entire verbal history of my time spent working on this degree -- several times over, I imagine. She's been a great roommate, who deserves a special award for putting up with my Kramer-esque existence. And speaking of *Seinfeld*, I'd like to just go ahead and thank Jerry himself. As you wrap up reading this lengthy Acknowledgments section and prepare to dive into an investigation into "The Psychology and Psychophysics of Voice Recognition," I'd like you to imagine Jerry's voice. Consciously activate whatever voice pattern is stored in your memory -- impose *his* quirky inflection on *your* own internal voice -- as you begin reading, and wonder, "What's the deal with voice recognition?"

Table of Contents

List of Tables

List of Figures

Abstract

Voice recognition is a fundamental pathway to person individuation, although typically overshadowed by its visual counterpart, face recognition. There have been no large scale, parametric studies investigating voice recognition performance as a function of cognitive variables in concert with voice parameters. Using celebrity voice clips of varying lengths, 1-4 sec., paired with similar sounding, unfamiliar voice foils, the present study investigated three key voice parameters distinguishing targets from foils -- fundamental frequency, f0 (pitch), subharmonic-to-harmonic ratio, SHR (creakiness), and syllabic rate--in concert with the cognitive variables of voice familiarity and judged voice distinctiveness as they contributed to recognition accuracy at varying clip lengths. All the variables had robust effects in clips as short as 1 sec. Objective measures of distinctiveness, quantified by the distances of each target voice to that target's sex-based mean for each parameter, showed that sensitivity to distinctiveness increased with familiarity. This effect was most evident on foil trials; at clip lengths of one second and above, f0 and SHR distinctiveness showed no discernible effect on match trials. Speaking rate distinctiveness improved match accuracy, an effect only seen with high familiarity. Recognition accuracy improved with the number of parameters that differed by an amount larger than the median, both in the target-to-foil and target-to-mean voice comparisons. A linear regression model of these three voice parameters, clip length, and subjective measures of distinctiveness and familiarity accounted for 36.7% of the variance in recognition accuracy.

Keywords: *voice recognition, famous voices, voice parameters, voice distinctiveness*

1. Introduction

Humans are social animals. As social animals it is vital that we distinguish other members of our species. The primary route to the identification of individuals is through face recognition and that route has been the subject of extensive research activity over the past 30 years. The most important secondary route to individuation has been through voice, which has received relatively little exploration. The present investigation explores the factors that allow a listener to determine the identity of a familiar speaker from a brief sample of speech. Although in an age of Caller ID and nighttime lighting, identification through voice has assumed a secondary status, it is still of great value for the blind or those with low vision or when a face is simply not in view. It has long occupied a prominent role in forensics. Prosopagnosics report that voices are invaluable in allowing them to identify familiar individuals (Facebook Prosopagnosic Group Entries, 2019). For instance, in a recent thread from August 23, 2019, user Michelle Rhiannon asked, “I was wondering if other people here find that they have really good voice recognition ability to compensate for their inability to recognize faces;” 75% of the 41 replies confirmed reliance on voice recognition, with responses such as, “Yes, I often wait for or get people to speak to figure out who they are,” “That’s the only way I recognise people,” and “Yes, I can sometimes even recognize when people are relatives by the sound of their voices. And I often recognize voices after talking to a person once.”

But what are the factors that allow a listener to identify a known speaker solely from a brief sample of his or her voice? Or to know that the voice is one

that they likely never (or rarely) previously encountered? The voice recognition literature is surprisingly lacking a systematic, parametric study of familiar voices.

That is, of course, not to say that the voice recognition literature as a whole is lacking. There is a large body of voice-specific literature within the more general person recognition literature, which seeks to draw pertinent similarities, differences, and functional connections between voice recognition, face recognition, and identity-specific semantic memory. These studies discuss cognitive modeling of the person recognition system (e.g. Campanella & Belin, 2007; Damjanovic, 2011; Stevenage et al., 2012) and the interactive effects of the face and voice modalities (e.g., Latinus et al., 2010; Schweinberger et al., 2010; Stevenage et al., 2014), as well as drawing pertinent distinctions between the two modalities (e.g. Hanley et al, 1998, Barsics, 2014; Biederman et al, 2018).

A subset of this literature seeks to specifically model the voice pattern, using paradigms of voice discrimination of newly learned (i.e. “learned-to-familiar”) voices. These studies, well-reviewed by Maguinnness et al. (2018), provide behavioral and functional imaging evidence of the theory that voice patterns are stored as deviations from a prototypical voice, first proposed by Papcun et al. (1989). Recent empirical support for this theory comes from Latinus et al. (2013), who presented subjects sets of 64 generated voices (32 male, 32 female) that varied along three dimensions (fundamental frequency, formant dispersion, and harmonics-to-noise ratio) and found greater BOLD activity in the temporal voice area (TVA) elicited by the voices that deviated most from the average values, suggesting greater sensitivity to deviant features. Subjects’ ratings of

distinctiveness correlated with degree of deviance from the average. In a similar study of discriminating voices, Baumann and Belin (2010) found that voices closer in a two-dimensional sound space comprised of f0 and f1 were perceived as more subjectively similar. They conclude that these two dimensions alone comprise a reasonably sufficient representation of an acoustic discriminability space. Another theory, not exclusive to the prototype model, is that voice patterns are stored as an average of all heard instances of a voice. To test this, Fontaine et al (2017) hypothesized that subjects would be better at recognizing voice averages, i.e. combined vocal morphs of varying numbers of vowel sounds, than singular vowels. They discovered that averages of vowels, ranging from one to five vowel morphs, showed no recognition improvement in newly learned voices However, they were better recognized in a famous voice recognition task, with roughly linear recognition improvements from ~60% to ~70% as the number of vowels within the morphed increased (and thus the morphs were more average sound representations).

This highlights one of the key motivations for the present study; voice discrimination and voice recognition are dissociable abilities, as first argued by Van Lancker & Krieman (1987) and supported by multiple studies in the ensuing decades (see Stevenage, 2018 for review). Case studies of phonagnosia in particular (Xu et al, 2015; Roswandowitz et al, 2014), starkly highlight this double dissociation. While the voice discrimination literature is rich with parametric analyses, it focuses heavily on newly learned (in the lab), rigorously controlled voices, learned often one word at time. We posit that studies that examine

"learned-to-familiar" voices are not studying voice *recognition* at its best; they are studying voice *discrimination* at its best. Looking at the early stages of voice pattern formation, and testing perception of single words or vowel sounds, the learned-to-familiar literature provides an incomplete, albeit invaluable, understanding of the capabilities of natural familiar voice recognition.

The familiar voice recognition literature studies a broader scope of naturalistic voice recognition often by using personally familiar voices, most common in the Forensic Voice literature (e.g. Ladefoged & Ladefoged, 1980; Rose & Duncan, 1995; Yarmey et al., 2001). These studies provide empirical evidence for effects of clip length and listening conditions (.e.g. whispering versus normal volume, verbal content), as well as a broad distinction between the recognizability of high versus low familiarity targets.

These personally familiar voice studies, given the difficulties in procuring personally familiar voice clips for large numbers of subjects, understandably suffer both from low sample sizes and a minimal range of degrees of familiarity. The alternative is to use sets of celebrity voices. Early studies in this field (e.g. Van Lancker et al, 1985; Meudell et al., 1980) established empirical understanding of accuracy under differing lengths, set sizes, and conditions (such as forward vs. backward speech). Schweinberger at al. (1997) ran the most extensive famous voices study prior to the present date, looking at open sets of voices and using a precise step-wise method of playing voice samples until subjects could report familiarity with the voice. They demonstrated recognition improvements occurring with each .25 sec of clip length, as well as the effects of different retrieval cues

(added voice stimulus, target occupation, initials of name) to aid name-specific recognition. Using a similar paradigm of assessing a "familiarity signal" using open sets of celebrity targets, Bethmann et al. (2012) concluded that the anterior temporal lobes (ATL) and superior temporal sulcus (STS) play a fundamental role in familiar voice recognition by finding greater fMRI BOLD response in these regions. The differential activation of familiar versus unfamiliar voice stimuli was proportional to the subjects' degree of familiarity with the familiar target.

These open set paradigms are essential for studying a "familiarity signal," the feeling of knowing *that* a voice is familiar but not necessarily being able to identify it. However, while they successfully probe the nature of such a signal, accounting for both degree of familiarity and length of exposure, they neglect to account for the acoustic voice features so rigorously measured in the unfamiliar and learned-to-familiar discrimination studies. In assessing the impressive scope of voice recognition literature, we were surprised to find that no studies have systematically compared familiar target voices and unfamiliar foil voices, in a way that bridges the discrimination-recognition gap.

1.1. Present Study: Detecting the Familiar Voice Signal

The present study assessed the accuracy of distinguishing celebrity voices, at various levels of familiarity, from the voices of foils that differed from the celebrity in several key parameters – fundamental frequency, subharmonic-to-harmonic ratio, and speaking rate. These three voice features were chosen as they have been implicated in successful voice discrimination and remain relatively

constant over the variable vocal signal (Bauman & Belin, 2010; van Dommelen, 1990; Skuk & Schweinberger, 2014; Krieman et al., 2017).

One unique aspect of voice recognition when compared to the parallel person individuation route of face recognition is that while familiar faces can be readily recognized from an unrestricted set of thousands of faces, e.g., “any celebrity” (e.g., Hacker et al., 2018), the accuracy of recognition of voices declines precipitously as the number of possible voices increases beyond a handful (Legge et al., 1984; Biederman, et al., 2018; Xu et al., 2015; Shilowich & Biederman, 2016). Furthermore, past studies utilizing familiar celebrity voices have mostly used unconstrained sets, with varying success rates as low as 10% (Meudell et al., 1980; Van Lancker et al., 1985; Schweinberger et al., 1997). By using a single celebrity target voice—that is under conditions of minimal uncertainty--we can observe familiar voice recognition within a range of accuracy sufficiently above floor and below ceiling so that the effects of individual distinguishing features between target and foil can be assessed. Using a simple match-not match paradigm, we assessed the role of: a) the three auditory parameters, b) clip length, and c) rated familiarity of the target, with accuracy in recognizing a celebrity voice against a foil that was matched for sex, age, and accent.

2. Materials and Methods

Subjects. 195 USC students were recruited through the USC Psychology Department subject pool and received credit in their psychology courses for their participation. The subjects participated over the internet, on the experimental

research website Testable.org. Participants filled out a questionnaire on their educational background, age, sex, handedness, history of brain trauma, self-assessments of hearing ability, voice recognition ability, and years of exposure to American culture. An item asked the participants to imagine Barack Obama's speaking voice and rate the vividness of the auditory image on a five-point scale. This was included in light of the prior findings (Xu et al, 2015; Shilowich & Biederman, 2016) revealing a relationship between an individual's voice recognition ability and the vividness of their voice imagery. Barack Obama was chosen as the exemplar as his was the most familiar voice to all participants in prior studies at USC and the familiarity of his voice was similarly rated highly (an average familiarity of 4.75 out of 5) by the participants in the present study.

2.1 Celebrity Voice Familiarity and Distinctiveness Pretest

In the section preceding the voice recognition task, participants were provided a list of 100 celebrity headshots with their names printed below the headshot (Figure 2.1), to provide ratings of the Familiarity and Distinctiveness of each celebrity's speaking voice to that participant. For each trait, ratings were made on a five-point sliding scale, and on both scales '1' indicated they have not heard the person speak. The remaining scale values (2 to 5) for Familiarity went from "I've heard the target speak very infrequently" to "very frequently". For the Distinctiveness scale, participants were asked to rate each celebrity's voice on a scale from "very common voice" (2) to "very unique voice (5)."

Figure 2.1: Example of Familiarity and Distinctiveness Rating Survey. Subjects rated all 100 celebrity targets before beginning the experiment.

2.2 Voice Recognition Task

The USC Voice Recognition Task (USC-VRT) can be accessed at https://www.testable.org/t/373f87d29. It consists of 100 trials, administered as two blocks of fifty trials each. After an instruction block and three practice trials with accuracy feedback, the testing blocks began, with no accuracy feedback. Each trial consisted of a two second fixation cross followed by a 2.5 second presentation of a celebrity headshot with name printed below it and the text “Is this celebrity the speaker?” printed above it. The celebrities were mostly entertainers but politicians and newscasters were also included. After the 2.5 second presentation, the headshot and texts remained on the screen while a sound clip played.

Figure 2.2: Sample Trial. After seeing the headshot and name for 2.5 seconds, a sound clip of either 1, 2, or 4 seconds played; the speaker in the clip was either the pictured celebrity or a similar sounding non-famous foil. Subjects chose either 'M' for "Match" or 'N' for "Not Match."

The clip was either one, two, or four seconds long, and the voice either matched the prompted celebrity target or was the voice of a non-famous person. Both parameters, clip length and celebrity versus foil speaker, were selected randomly by the Testable (www.testable.org) function randomPick. The voice clips were retrieved by the author from television and radio interviews, with no semantic clues to the identity or profession of the speaker. Foil voice clips were obtained from the same media forms from non-celebrity speakers (e.g. local guests, journalists, etc.). The gender and race of the speakers were always matched between target and foil; accent and age were closely matched to the author's best subjective judgment. Upon hearing the voice, participants selected either 'M' if the voice matched the celebrity target or 'N' if the voice did not match the target. Participants were instructed to respond as quickly as possible, even if the clip

hadn't finished playing yet. There was no indication prior to each trial as to how long the clip was going to be.

2.3 Sound Parameter Analysis

The voice stimuli clips were analyzed using Voice Sauce (Shue et al., 2011) in MATLAB (Mathworks, 2012) to compare the voice stimuli parametrically. The top fifty most familiar (determined by mean familiarity rating) celebrity targets and their voice foils were chosen for the voice parameter analysis. These fifty celebrities were chosen in a post hoc analysis; the lesser-known targets were not highly familiar (rating >3) to any subject, and the focus of the post hoc analysis was on subjects' performance on highly familiar trials. The mean familiarity rating for the unused trials was 1.75 (SD = .46); the high familiarity trials we used had a mean rating of 3.60 (SD = .48). The trials chosen for analysis had each case (three clip lengths of each identity, celeb vs foil speaker) represented by the full range of familiarity. Each of the voice sample's full four second clips was used to extract mean f0 values using the Snack algorithm (Sjölander, 2004) and subharmonic-to-harmonic ratio using the SHR algorithm (Sun, 2002). Furthermore, a speaking speed analysis was performed by averaging the number of syllables spoken over the three different clip lengths. The syllable count was performed subjectively by the author. The correlation over two counts separated by seven months was r = .97.

There were no discernible differences between the distributions of parametric values for the celebrity voices compared to the non-famous foil voices.

Table 2.1 shows the descriptive statistics of each parameter, grouped by sex and celebrity status.

Table 2.1:

Descriptive statistics of voice parameter distributions by sex and celebrity status

		F0		SHR		Syllables/sec	
		Celebs	Foils	Celebs	Foils	Celebs	Foils
Male	Mean	115.26	120.35	0.61	0.62	4.70	5.15
	SD	22.86	24.65	0.06	0.06	1.04	0.85
	Skew	0.84	1.41	-0.49	-0.64	0.44	0.49
Female	Mean	187.87	190.82	0.60	0.60	5.40	4.52
	SD	22.61	21.86	0.05	0.09	0.56	0.72
	Skew	0.47	0.36	-0.22	-0.33	0.54	-0.39

2.4 Data analysis inclusion criteria

195 individuals initiated the experiment, eight of whom did not complete it and were dropped from the analysis. Seven additional subjects were dropped for their insufficient immersion in American culture (<5 years). This was done because the subsequent analyses depended highly on familiarity with the targets and these seven subjects were familiar with less than 20% of the targets. Finally, four subjects were removed for questionable testing behavior; the accuracy of their performance was at chance and their reaction times were as fast as possible. After the removal of those 19 participants, 176 subjects remained. Congenital phonagnosic subject AN (Xu et al, 2015) also partook in the experiment; her results were not included in the aggregate main analyses but will be discussed.

2.4.1 Subject characteristics of included subjects

The 176 subjects included in the analysis had a mean age of 20.3 (SD = 3.0). 133 of the participants were female and 156 were right-handed. None reported any brain damage or neurological insult that would affect their hearing or voice perception.

3. Results

3.1. Subject Characteristic Effects

The overall accuracy on the USC-VRT was 67.4% (SD = 8.4%), with chance being 50%. Neither sex nor handedness caused significant voice recognition differences. The mean male score was 68.1% (SD = 8.2%) and females averaged 67.3% (SD = 8.5%); $t(174) = .535$, *ns*. Right handed subjects averaged 67.1% (SD = 8.5%), left handed subjects averaged 70.1% (SD = 7.0%), $t(174) = -1.50$, *ns.* There was a modest positive correlation between age and recognition accuracy, $r(175) = .29$, $p < .001$, likely in part explained by a smaller, correlation between age and familiarity with the celebrity targets, $r(175) = .16$, $p < .03$. These two relationships can be seen in Figure 3.1.

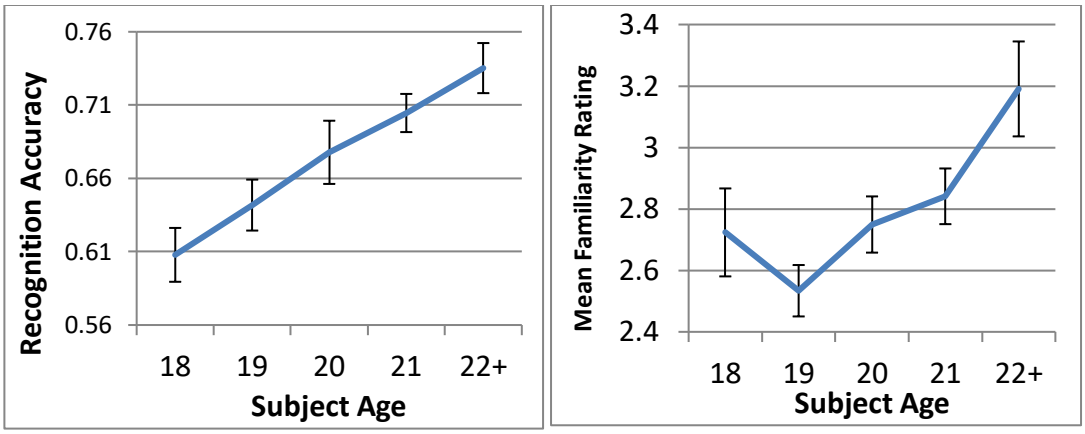


Figure 3.1: Recognition accuracy (left) and mean familiarity ratings (right) as a function of subject age. Error bars are 1 SE for each group.

There was an overall correlation with age and accuracy of .29 and age and familiarity of .16; as familiarity and accuracy were correlated (Fig. 3.1, right panel). The improvement in recognition performance with age is in large part due to higher familiarity of the older subjects with the target voices. This is not surprising as the celebrities were initially generated by phonagnosic AN (at the time a USC sophomore) six years prior to the running of the current study. N-Sizes of each bin in Fig. 3.1: 18 years = 14; 19 years = 50; 20 years = 54; 21 years = 37; 22+ years = 21. One-way ANOVAs between the age groupings were significant both for accuracy, $F(4,171) = 9.62$, $p < .001$, $\eta_p^2 = .184$, and familiarity ratings, $F(4,171) = 4.37$, $p < .002$, $\eta_p^2 = .093$.

The mean score for all familiar trials rated greater than 1 on a 5-point scale was 72.6% (SD = 8.4%). Seven subjects reported below average hearing, but performed at the mean on familiar trials (>1 rating), 69.8% (SD = 7.5%), in comparison with normal hearing subjects, $t(174) < 1.00$. Eleven subjects reported below average voice recognition abilities; their scores for familiar targets were

lower on average, 69.0% (SD = 7.2%), than those who reported average or better voice recognition abilities (72.8%, SD = 8.4%), but not significantly so, $t(174) = 1.48$, *n.s.* There was a modest correlation, $r(175) = .24$, $p < .002$, between voice imagery vividness ratings and voice recognition performance. Twenty-two subjects rated Obama's speaking voice as highly familiar but rated their vividness of imagining his voice as a three or less on the five-point scale. Consistent with the correlation between vividness of imagery and voice recognition accuracy found in Xu et al. (2015) and Shilowich & Biederman (2016), these subjects performed significantly worse on the recognition test (M = 68.9%, SD = 9.0%) than the subjects who gave a high vividness ratings (>3) to the auditory image of Obama's voice (M = 73.1%, SD = 8.2%) - $t(174)= 2.27$, $p < .025$, $d = .49$.

3.2 Main recognition results

Familiarity - As documented in prior studies (Xu et al., 2015; Shilowich & Biederman, 2016) and as shown in Fig. 3.2, the higher the rated familiarity of a voice, the more accurately it was judged, $F(4,668) = 90.6$, $p < .001$, $\eta_p^2 = .352$. There was a high positive correlation between familiarity ratings and accuracy, $r = .51$, $p <.001$; the correlation with distinctiveness accuracy was lower but still highly significant, $r = .49$, $p < .001$. The correlation between Familiarity and Distinctiveness was very high, $r = .86$, $p <.001$.

Clip Length – The longer the clip length, the higher the accuracy as shown in Figure 3.2; $F(2,350) = 18.37$, $p < .001$, $\eta_p^2 = .095$. For clips lengths of 1, 2, and 4 seconds, matching accuracy was 64.3% (SD = 10.1%); 68.1% (SD = 10.9%),

and 69.3% (SD = 10.8%). The interaction between Length and Familiarity was not significant, $F(8,864) < 1.00$.

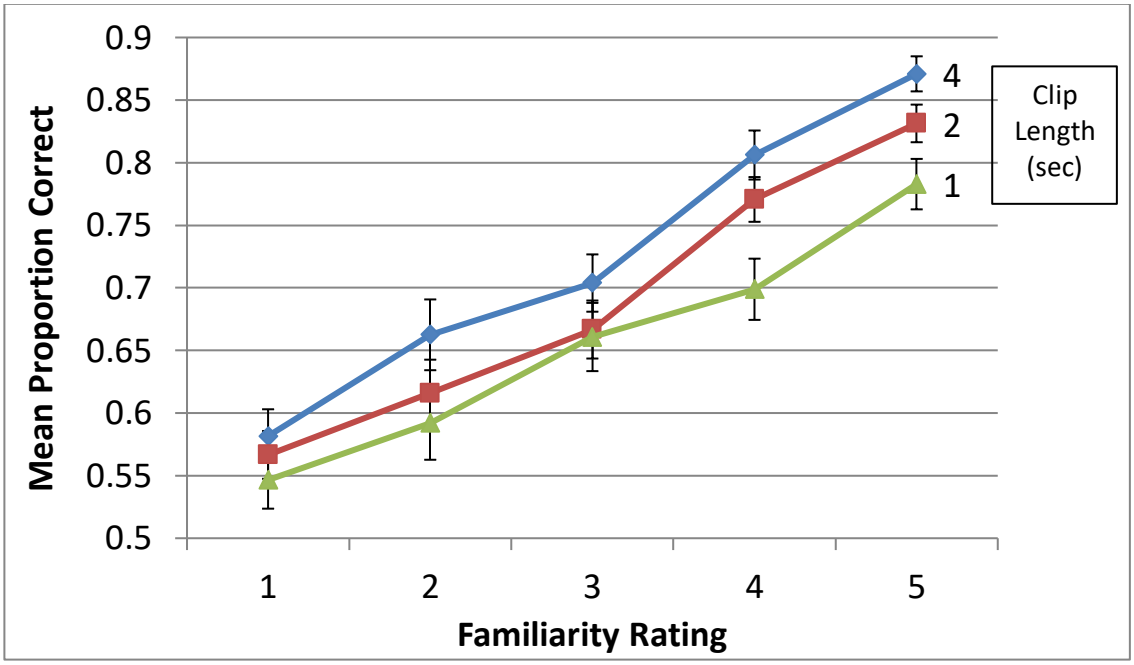


Figure 3.2: Recognition accuracy by clip length and familiarity. Accuracy increased both with increasing clip length and familiarity. Error bars indicate 1 SE.

Matching – There was a pronounced effect on recognition accuracy of whether the voice sample was the celebrity target or a non-famous voice with mean accuracy of 78.9% (SD = 9.7%) for positive match trials compared to a mean accuracy of 66.5% (SD = 14.3%) for negative matches. A repeated measures ANOVA showed that this difference was highly significant $F(1,175) = 15.93$, $p < .001$, $\eta_p^2 = .083$.

There was a pronounced interaction between Familiarity and Match Case (Match trials vs Foil trials), in which Match trial performance improved monotonically with increasing Familiarity, but Foil trial performance only improved at the highest two levels of familiarity. This interaction between Familiarity and

Match case was significant at $F(4,620) = 26.67$, $p < .001$, $\eta_p^2 = .147$. There was no interaction between Clip Length and Match Case – $F(2, 350) = .024$, *.n.s.;* however, there was a three-way interaction between Clip Length, Familiarity Rating, and Match Case of the trial, shown in Figure 3.3. In Match trials, there was in improvement between one and two seconds but not two a four. In Foil trials, there was no differential performance between the clip lengths at Familiarity of 3 and below; at levels 4 and 5 there is a linear additive relationship between Familiarity and Clip Length. The three-way interaction of Length, Familiarity, and Match was significant at $F(8,1400) = 6.60$, $p < .001$, $\eta_p^2 = .036$.

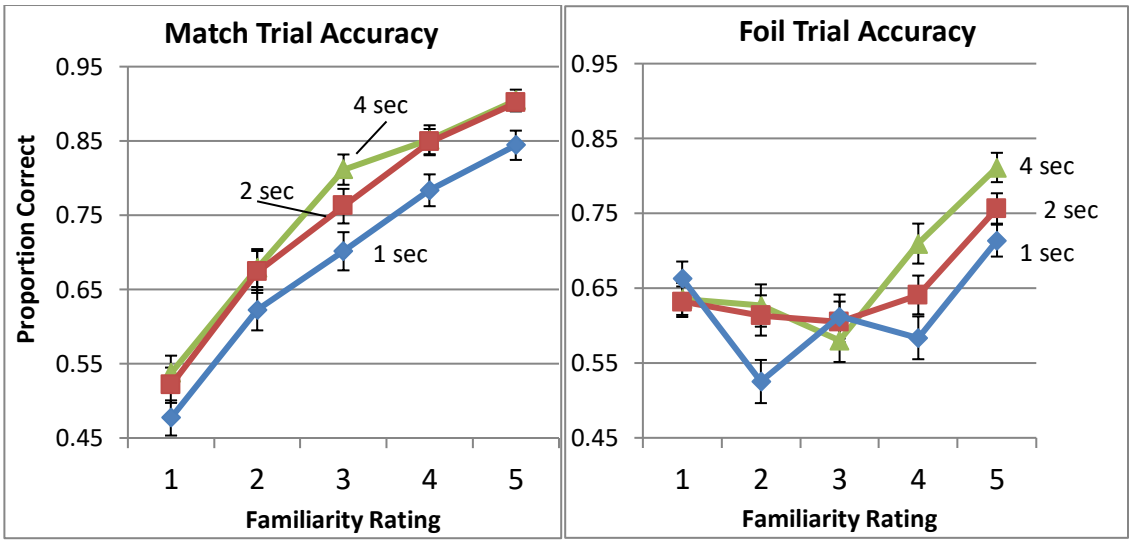


Figure 3.3: Recognition accuracy by clip length, familiarity, and matching condition. Performance in matching cases is linear by familiarity, with an improvement in performance between 1 and 2 seconds. Below familiarity ratings of 4, familiarity and clip length do not affect accuracy in foil trials (right panel). Error bars indicate 1 SE.

Reaction Time – As shown in Fig. 3.4, the shorter the clip length, the shorter the mean correct RTs. In a 3x5 repeated measures ANOVA of Clip Length and Familiarity, the main effect of Clip Length was $F(2, 350) = 232$, $p < .001$, $\eta_p^2 =$

.570. The same ANOVA yielded a significant effect of familiarity, $F(4, 700) = 3.40$, $p < .009$, $\eta_p^2 = .019$. The interaction between clip length and familiarity ratings was also significant, $F(8,1400) = 2.19$, $p < .05$, $\eta_p^2 = .012$. A plausible interpretation of why longer clip lengths were associated with both higher accuracy and longer RTs is that subjects used the additional durations of the clips productively; the longer clip durations could provide additional distinctive voice characteristics that could inform their decision.

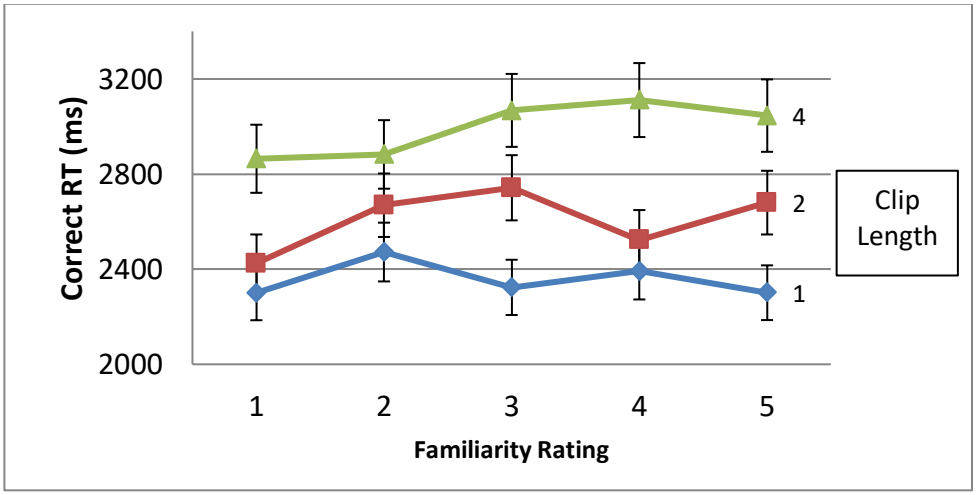


Figure 3.4: Mean Correct RT across clip lengths and familiarity ratings. Error bars are 1 SE

Match trials were significantly faster on average than foil trials; mean correct RT for match trials was 2,568 ms (SD = 544 ms); for foil trials, M = 2,642 ms (SD = 576 ms). This difference is better understood in terms of the highly significant interaction between familiarity and target identity, $F(4,700) = 21.34$, $p < .001$, $\eta_p^2 = .109$. Shown in Figure 3.5 below, match trials were much faster in all familiar (>1 rating) trials; whereas foil trials were significantly faster only on the trials with the lowest familiarity rating (of 1). This reaction time difference at 1

Familiarity levels was likely the result of a criterion shift; at 1 Familiarity, subjects exhibit a conservative bias. Knowing they cannot recognize the voice, they respond "Not Match," quickly. In all familiar trials, Match trials are quicker likely because subjects answer as quickly as they accumulate enough positive evidence for a Match; Foil trials require more exhaustive processing to eliminate uncertainty. The interaction between Clip Length and Match Case was not significant – $F(2,350) < 1.00$, *n.s.*

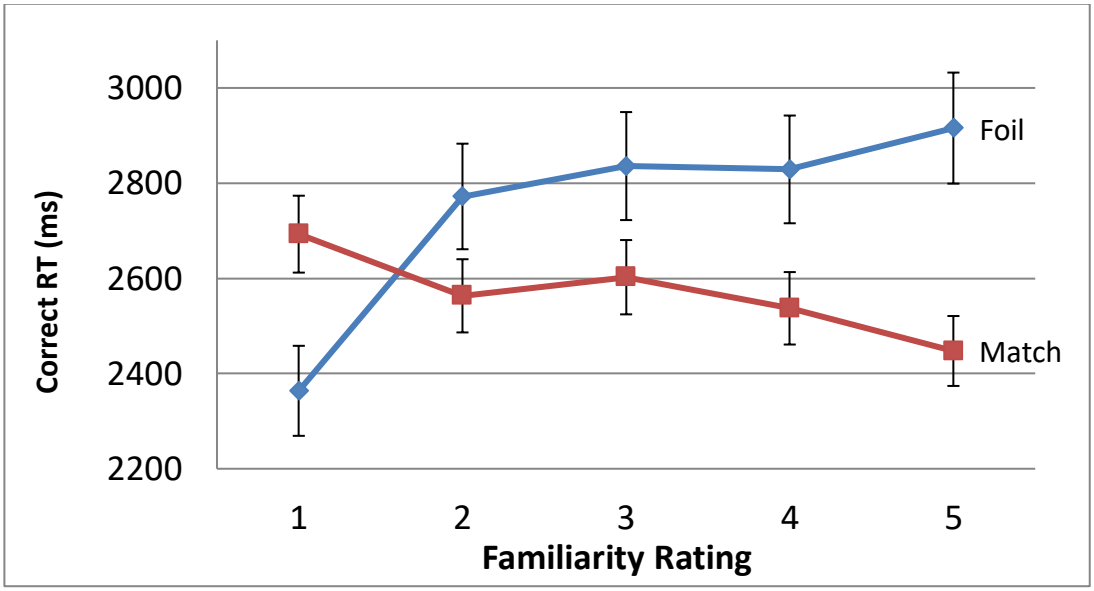


Figure 3.5: The interaction between target identity and familiarity on correct RT. Subjects are quicker to successfully recognize a voice (Match) than to correctly reject a foil voice (Foil). At the highest familiarity rating of 5, the mean difference between these processes is 468 ms. Error bars are 1 SE.

Phonagnosic Subject AN – AN's accuracy on the task was low, scoring an average of 65.0% on all trials compared to 77.9% (SD = 14.2%) for the high familiarity (4 and 5 ratings) of nonphonagnosic subjects. We used these trials as a basis for comparison because AN's mean familiarity was very high (4.77) and she rated no celebrity target lower than a 4, which is to be expected given that she,

herself, generated the list of celebrities in Xu et al. (2015). The highest mean familiarity of the control subjects was 4.55 with a mean of 2.75. Further analysis revealed that she had a strong bias to respond that the voice sample was not a match to the target. She was 85% correct on not-match trials but only 42% correct on match trials. This is more than four SD's below the mean of comparable trials by control subjects who had an 85.5% (SD = 10.2%) accuracy level on high familiarity, positive match trials. Figure 3.6 shows AN's performance compared to the high familiarity trials of conservative subjects (n = 42), who, like AN, have a tendency to answer not-match more frequently, and liberal subjects, who are biased towards match responses. We see here that differences in criteria do not reflect differences in sensitivity.

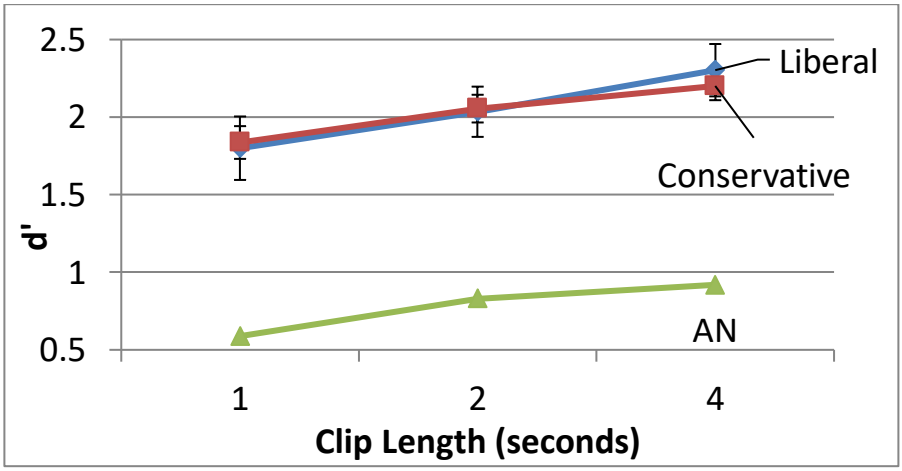


Figure 3.6: Phonagnosic subject AN compared to the high familiarity target trials of liberal subjects (who have negative criterion values, a bias towards answering "match") and conservative subjects (who have positive criterion values, a bias towards answering "not match," like AN).

3.3 Signal Detection Analysis

We analyzed the data as a signal detection task (discriminating the familiar pattern signal from the unfamiliar pattern noise). Sensitivity (measured by *d'*)

increased monotonically as rated familiarity increased as shown in Fig. 3.7; in a 3x5 repeated measures ANOVA between the familiarity categories and clip length, the main effect of familiarity on *d*' was significant at $F(4,700) = 49.6$, $p < .001$, $\eta_p^2 = .221$ Voice sample length did not have a significant effect, $F(2,350) = 1.11$, *n.s*. The interaction between familiarity and clip length on *d'*, shown in Figure 3.7, was significant at $F(8, 1400) = 2.73$, $p < .005$, $\eta_p^2 = .015$.

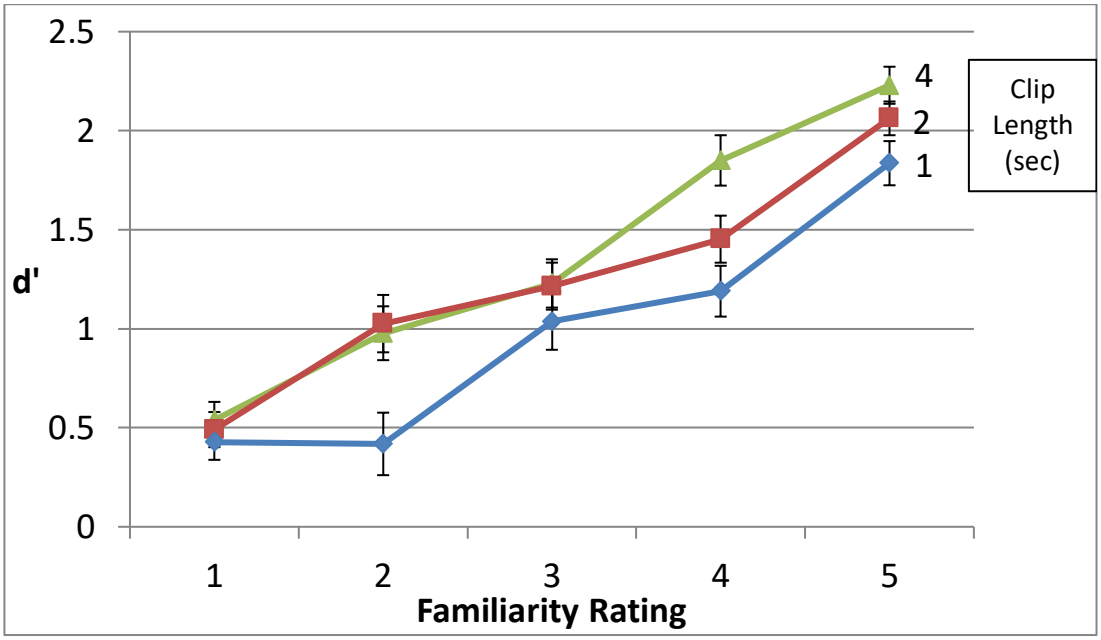


Fig 3.7: *d'*s for matching a celebrity name (the target) against a sample voice as a function of the familiarity rating of the target's voice and segment length. Sensitivity increased both with familiarity of the target voice and the length of voice sample. Error bars indicate 1 SE.

3.4 Speech Parameter Analysis

To assess the effects of the three measured speech parameters – fundamental frequency, subharmonic-to-harmonic ratio, and speaking rate –we compared recognition performance as a function of the absolute difference in a parameter's value between the foil's voice and the celebrity target's voice, as well

between each target voice and the mean parameter values for that target's sex. These analyses were performed on the top 50 most familiar celebrities, defined by average familiarity rating; the analyses focused on the differences between low (2 and 3 ratings) and high (4 and 5 ratings) familiarity, and the least familiar targets had no subjects rate them as highly familiar.

3.4.1 Target to Foil Parametric Differences by Clip Length and Familiarity

We assessed the effects of each parameter with a 3-way repeated measured ANOVA; each parameter difference was binned into three groups – low, medium, and high difference between target and foil. Values for these categories are found in Table 3.1 below. Familiarity was binned into high familiarity (ratings of 4 or 5) and low familiarity (ratings of 2 or 3). The three clip lengths of 1, 2, and 4 seconds comprised the length category. All trials rated greater than a Familiarity level above 1 were included in the analysis

Table 3.1:
Parametric Bin Values, [Target – Foil].

	BIN		
Parameter	**Low**	**Medium**	**High**
F0 [Δhertz] Mean *(SD)*	< 16 9.9 (*4.3*)	16 - 32 23.0 (*5.2*)	> 32 40.6 (*6.3*)
SHR [Δratio] Mean *(SD)*	< .045 .023 (*.01*)	.045 - .095 .071 (*.02*)	> .095 .121 (*.03*)
Speaking Rate [Δsyllables/sec] Mean *(SD)*	< .85 .35 (*.23*)	.85 – 1.5 1.1 (*.20*)	> 1.5 1.8 (*.31*)

Each trial was binned according to low, medium, and high differences between target and foil. The value ranges are listed on top, each in the units corresponding to each parameter (listed next to the parameter as [Δ unit]; the bin means and standard deviations are below the ranges.

Figure 3.8 shows that increases in fundamental frequency difference between target and foil improved recognition, with a main effect of $F(2,350) = 90.31$, $p < .001$, $\eta_p^2 = .340$. Both levels of familiarity showed improvement over the differences in f0, but the high familiar group showed greater sensitivity to the medium f0 category. The interaction between familiarity and f0 was significant, $F(2,350) = 4.89$, $p < .008$, $\eta_p^2 = .027$. There was also a significant interaction between clip length and f0 difference, $F(4,700) = 4.56$, $p < .001$, $\eta_p^2 = .025$ in that recognition only improved from two to four seconds, benefiting from the extra time. At the highest f0 difference level, all three lengths are comparable, indicating that one second suffices for recognition with a sizable f0 difference. The three way interaction between Length, Familiarity, and Fundamental Frequency was not significant, $F(4,700) = 1.20$, *n.s.*

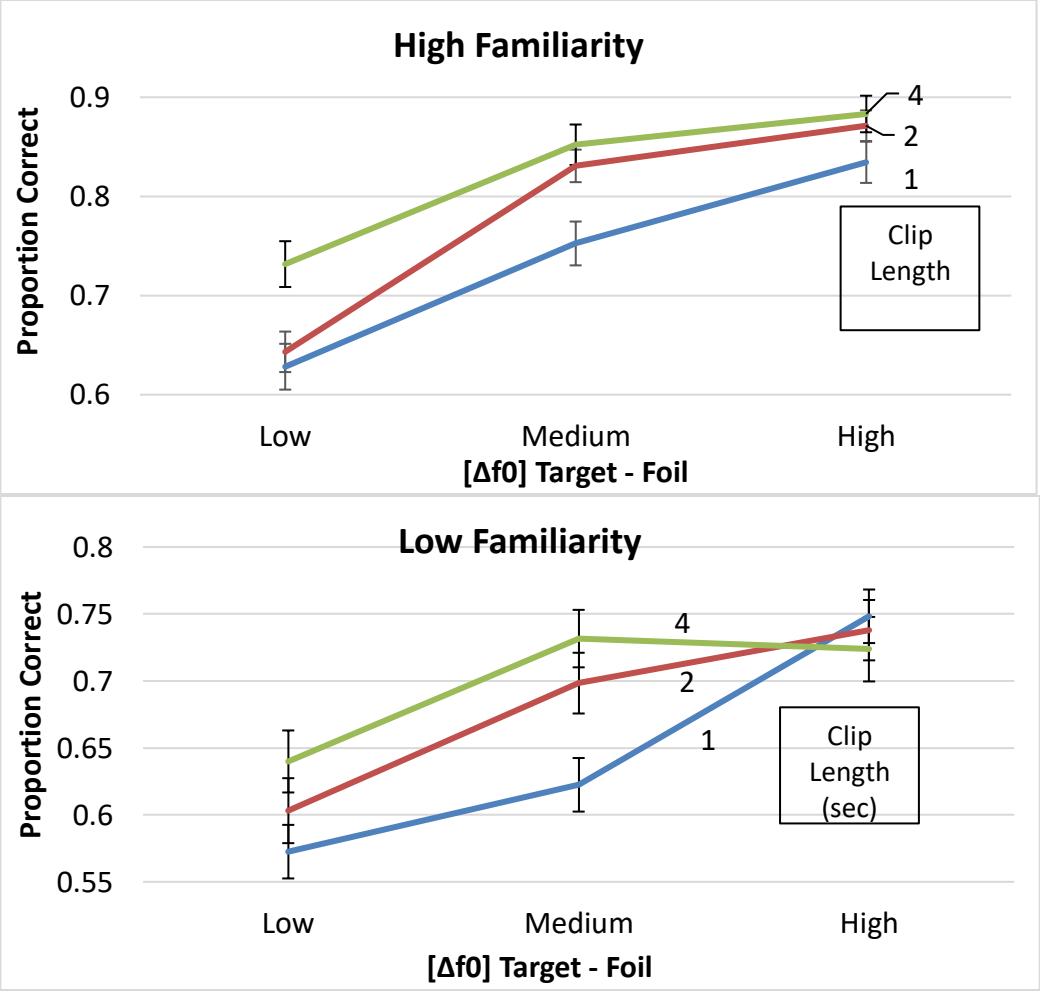
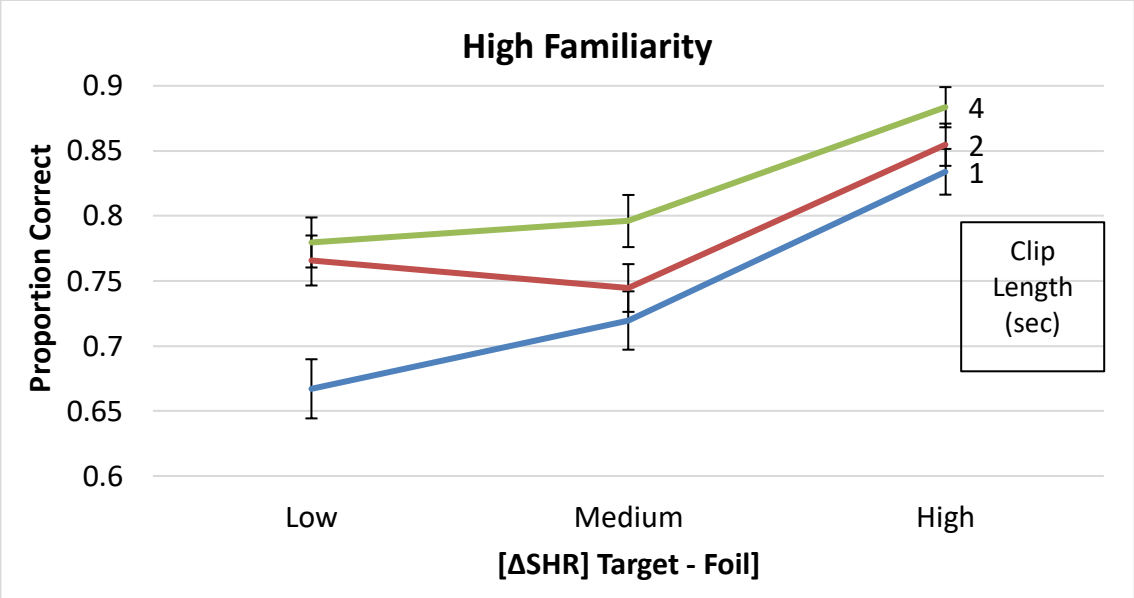


Figure 3.8: Fundamental frequency difference between target and foil voice predicts recognition accuracy. High familiarity trials are in the top panel; low familiarity trials are below. The [Δf0] values for each bin are: Low, >16 Hz; Medium, 16-32 Hz; High, >32 Hz. Error bars are 1 SE.

The same three way repeated measures ANOVA was performed with subharmonic-to-harmonic ratio. As seen in Figure 3.9 below, there was a significant main effect of SHR on recognition performance, with higher levels of SHR difference resulting in higher recognition accuracy, $F(2,350) = 44.08$, $p < .001$, $\eta_p^2 = .201$.



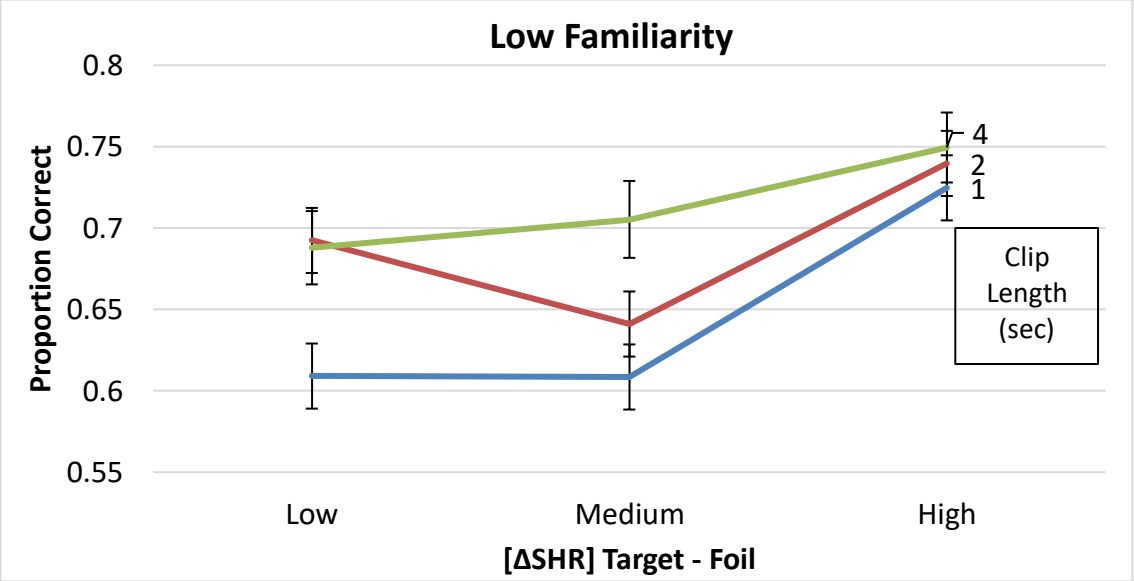
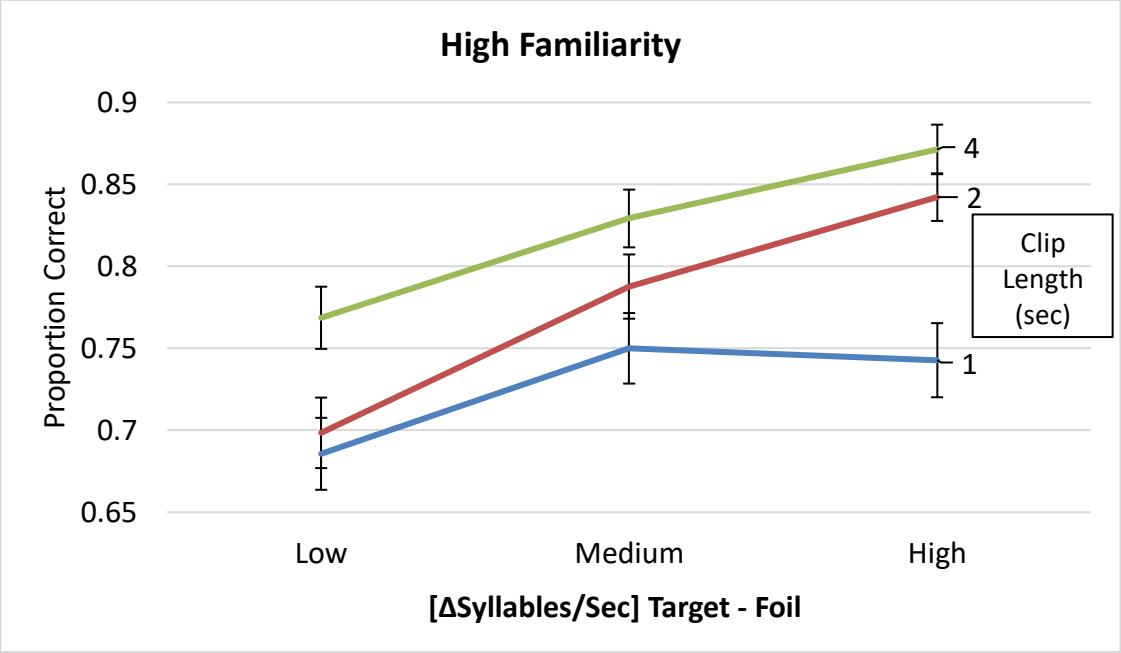


Figure 3.9: Subharmonic-to-harmonic ratio difference between target and foil voice predicts recognition accuracy. High familiarity trials are on top, low familiarity trials are on bottom. The [ΔSHR] values for each bin are: Low, > .045; Medium, .045 - .095; High, >.095. Error bars are 1 SE.

Familiarity did not interact significantly with SHR, $F(2, 350) = 1.89$, *n.s.*, with both groups showing greatest sensitivity between the middle and high SHR difference categories. The interaction between Clip Length and SHR was significant, $F(4,700) = 2.72$, $p < .029$, $\eta_p^2 = .015$, with differences in SHR improving

most at the shortest clip length. The overall three-way interaction was not significant, $F(4,700) = .246$, *n.s.*, but at the one second clip length the lower familiarity group only benefits from the highest SHR difference, whereas the higher familiarity trials show improvement at the medium level. As with fundamental frequency, the highest difference in SHR, trials in which the difference between target and foil SHR was > .095, sufficed for peak performance even at one second stimulus length.

Speaking rate difference between foil and target, measured in syllables per second, improved recognition accuracy, with a main effect of $F(2,350) = 23.32$, $p < .001$, $\eta_p^2 = .118$. The effect of familiarity on sensitivity to speaking rate fell short of significance, $F(2,350) = 2.47$, $p < .086$. Clip Length had a significant interaction with speaking rate, $F(4,700) = 2.56$, $\eta_p^2 = .014$, with subjects' one second clip length trials less sensitive to the higher differences in rate, likely due to the limited range of syllables at that length. The three way interaction, shown in Figure 3.10, between Length, Familiarity, and Speaking Rate, was not significant – $F(4,700) = <1.00$, *n.s.*



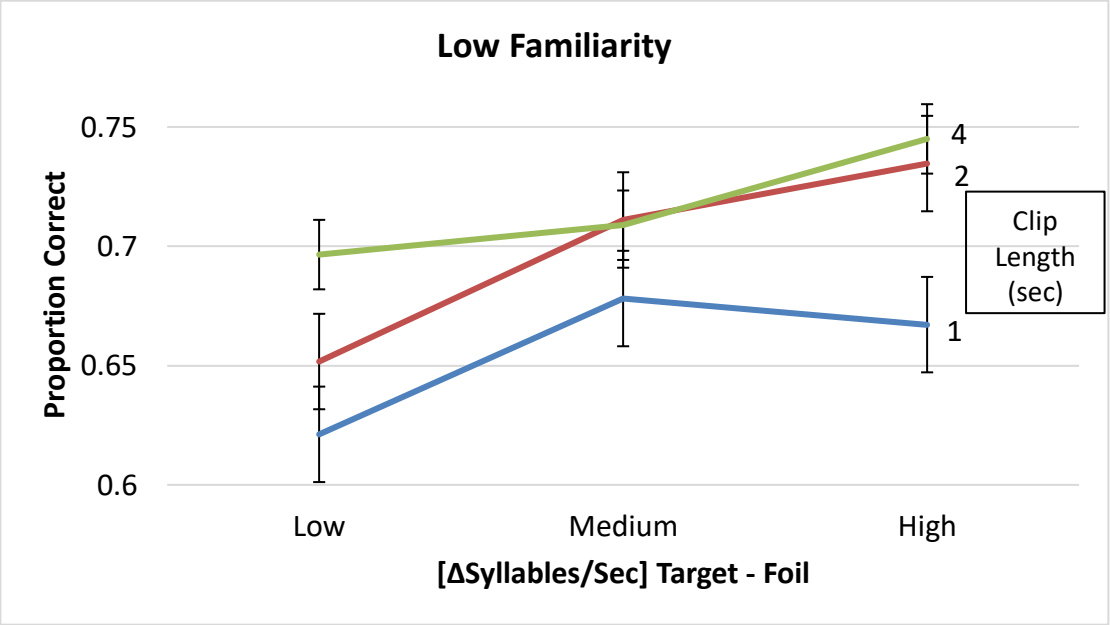


Figure 3.10: The effect of Target minus Foil Speaking Rate Differences and Clip Length on recognition accuracy. The [ΔSyllables/sec] values for each bin are: Low, > .85; Medium, .85 – 1.5; High, > 1.5. Error bars are 1 SE.

In addition to the absolute difference in speaking rate, there was improved performance specifically when the target was the faster speaker. To assess these effects, we binned speaking rates into High (>1.2 syllables/sec difference) and

Low (<1.2 syllables/sec difference) and ran a 2x2x2 repeated measures ANOVA with Speaking Rate Difference (High and Low), Familiarity (High ratings, 4-5, vs Low, 2-3) and Faster Speaker (Target vs Foil). The results are seen in Figure 3.1 below. Recognition accuracy was higher when the target was the faster speaker, with a significant main effect of $F(1,175) = 28.02$, $p < .001$, $\eta_p^2 = .139$. The two-way and three-way interactions were not significant.

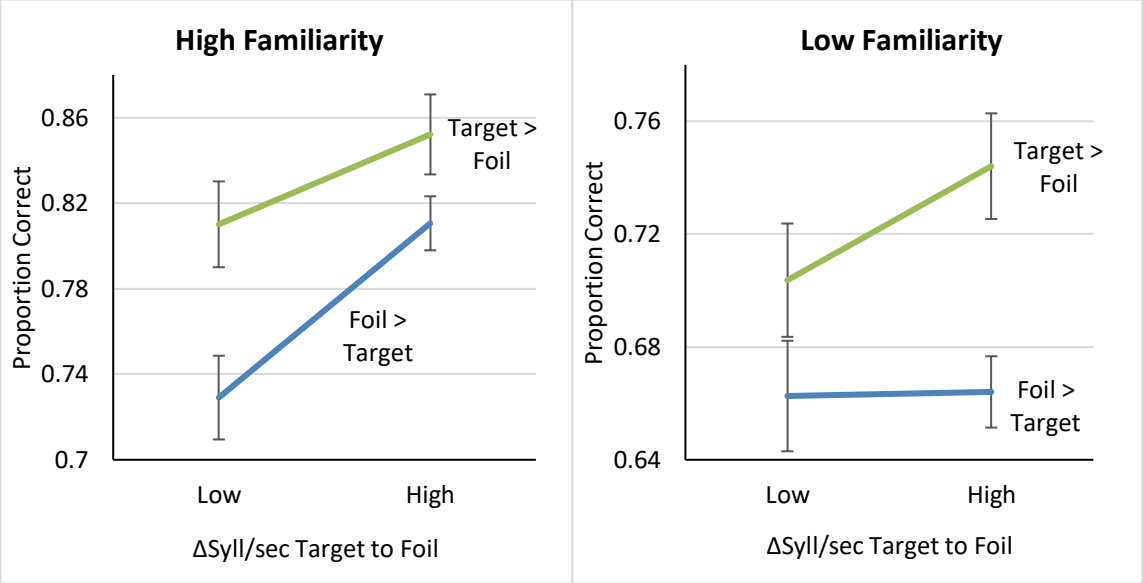


Figure 3.11: Effect of which speaker is faster. In all cases, accuracy is higher when the target is the faster speaker. At low familiarity, there is no sensitivity to speaking rate difference when the foil is the faster speaker. The Low difference group contains trials in which the difference was less than 1.2 syllables/sec; the High difference contains trials in which the difference was greater than 1.2 syllables/sec. Error bars are 1 SE.

3.4.2 Target to Foil Parameter Interactions

To assess the combined effects of the speech parameters, we used all clip lengths of highly familiar trials (>3) and ran 3x3 repeated measures ANOVAs on each parameter pair.

Figure 3.12 shows the combined effects of Target-to-Foil differences in both subharmonic-to-harmonic ratio and fundamental frequency. In general, the larger the Target-to-Foil differences on either variable, the greater the accuracy, with still a higher level of accuracy achieved with larger differences on both variables. The interaction was significant, $F(4,700) = 11.50$, $p < .001$, $\eta_p^2 = .062$, with much of it attributable to ceiling effects and variable sensitivity to the parameter differences, such as the lack of an effect between Low and Medium differences in speaking rate. This benefit of having multiple differences on the speech parameters was witnessed for all parameter pairs.

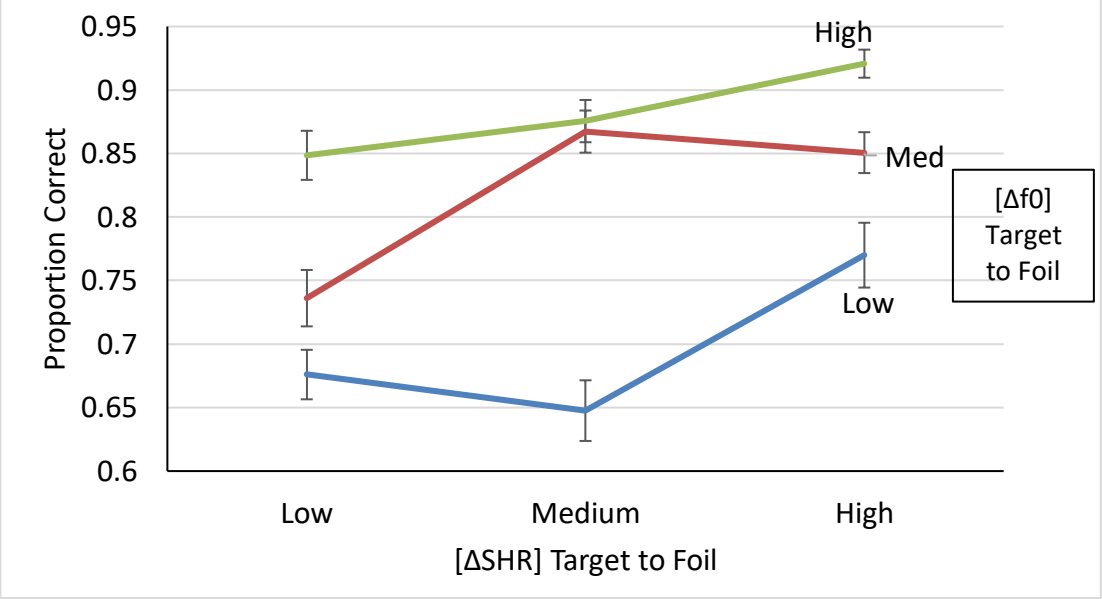


Figure 3.12: Interaction between f0 and SHR differences between target and foil voices. Error bars are 1 SE.

The interaction between speaking rate and SHR was similar to that for SHR and f0 differences, as shown in Figure 3.13 below. The repeated measures ANOVA for the interaction was significant, $F(4,700) = 13.80$, $p < .001$, $\eta_p^2 = .135$.

Similar to the interaction with f0, there were ceiling effects and differential sensitivity to various ranges of the variables.

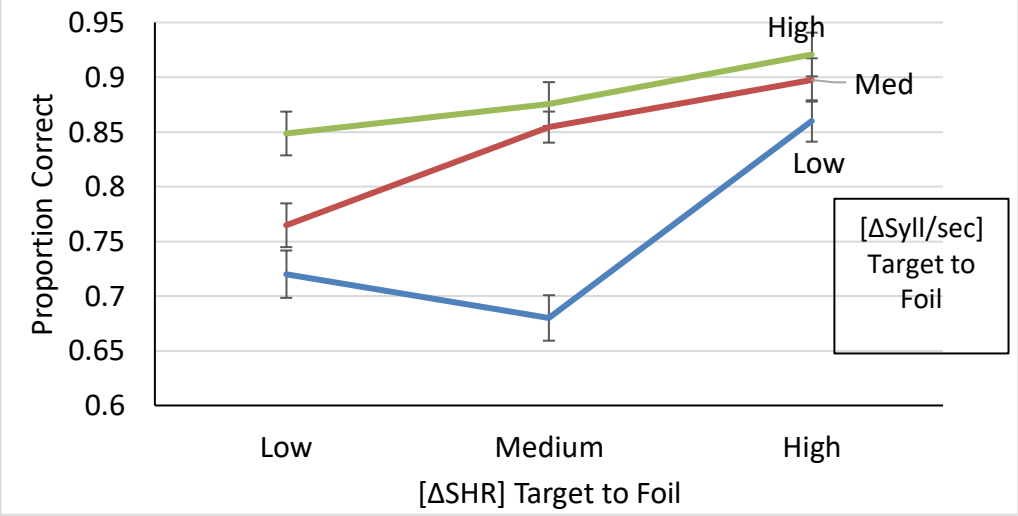


Figure 3.13: The effect of differences in SHR and speaking rate on accuracy. Error bars are 1 SE.

The interaction for the final parameter pair, fundamental frequency and speaking rate, was not significant $F(4,700) = 1.66$, *n.s.* At the highest levels of speaking rate difference, there was a monotonic improvement in recognition over the different [Δf0] categories.

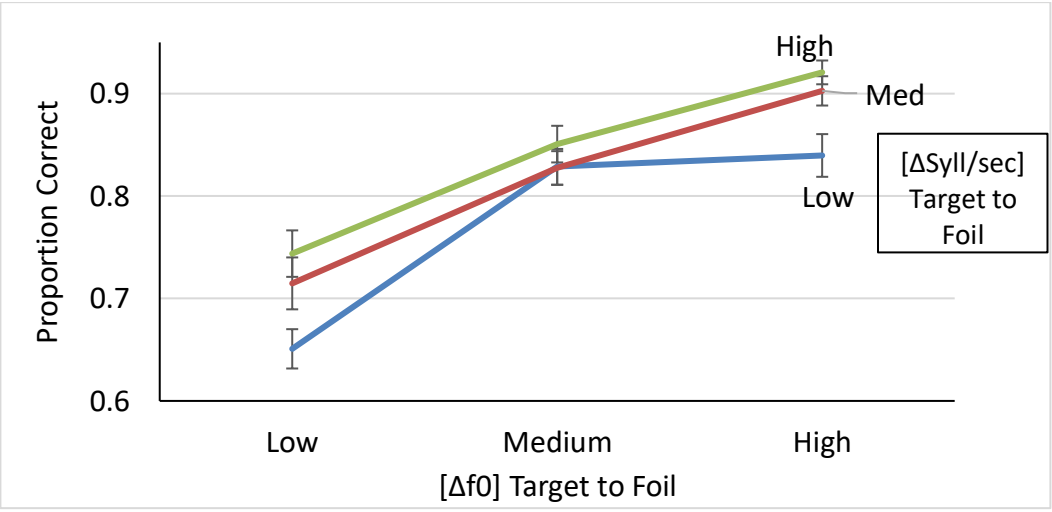


Figure 3.14: The effect of fundamental frequency and speaking rate. Error bars are 1 SE.

3.4.3 A Rigorous Definition of Voice Distinctiveness

We examined each parameter by comparing the target voice to the mean voice (calculated from the average values of our stimulus set), as objective measures of distinctiveness. Separate means were calculated for each sex; male voices (n = 72) had a mean f0 of 117.8 Hz (SD = 23.7 Hz) and female voices (n = 28) had a mean f0 of 186.4 Hz (SD = 25.0 Hz), both in the normal range of f0 (Skuk & Schweinberger, 2014). Male voices had an average SHR of 0.611 (SD = 0.056); females had an average SHR of 0.596 (SD = 0.071), also both in the normal range (Sun, 2002). Male voices averaged 4.7 syllables/second (SD = 1.0 syllable/sec) and females averaged 5.3 syllables/sec (SD = 0.6 syllables/sec). We compared each target's distance-to-foil to that that target's distance-to-mean on each parameter. The bin values for the distances to the mean are seen in Table 3.2 below.

Table 3.2

Parametric Bin Values, [Target Voice's Value – Target's Sex Mean Value]

	BIN		
Parameter	**Low**	**Medium**	**High**
F0 [Δ hertz]	< 8.4	8.4 - 20	> 20
Mean (*SD*)	4.6 (*2.6*)	13.0 (*3.5*)	28 (*5.6*)
SHR [Δ ratio]	< .025	.025 - .054	> .054
Mean (*SD*)	.012 (*.01*)	.043 (*.01*)	.081 (*.02*)
Speaking Rate [Δ syllables/sec]	< .59	.59 – 1.15	> 1.15
Mean (*SD*)	.34 (*.19*)	.86 (*.19*)	1.44 (*.33*)

Each trial was binned according to low, medium, and high differences between the target and mean for that target's sex. The value ranges are listed on top, each in the units corresponding to each parameter (listed next to the parameter as [Δ unit]; the bin means and standard deviations are below the ranges.

Comparing the effects of target-foil and target-mean values for each parameter, only fundamental frequency yielded a significant interaction. As shown in Figure 3.15, the interaction was a consequence that at lower levels of f0 differences between target and foil, subjects could better recognize those voices with larger f0 distances from the mean f0. At the higher levels, ceiling effects limited the magnitude of this benefit of f0 distinctiveness. The 3x3 repeated measures ANOVA between the foil and the target f0 distances was significant, $F(4,700) = 5.46$, $p < .001$, $\eta_p^2 = .030$.

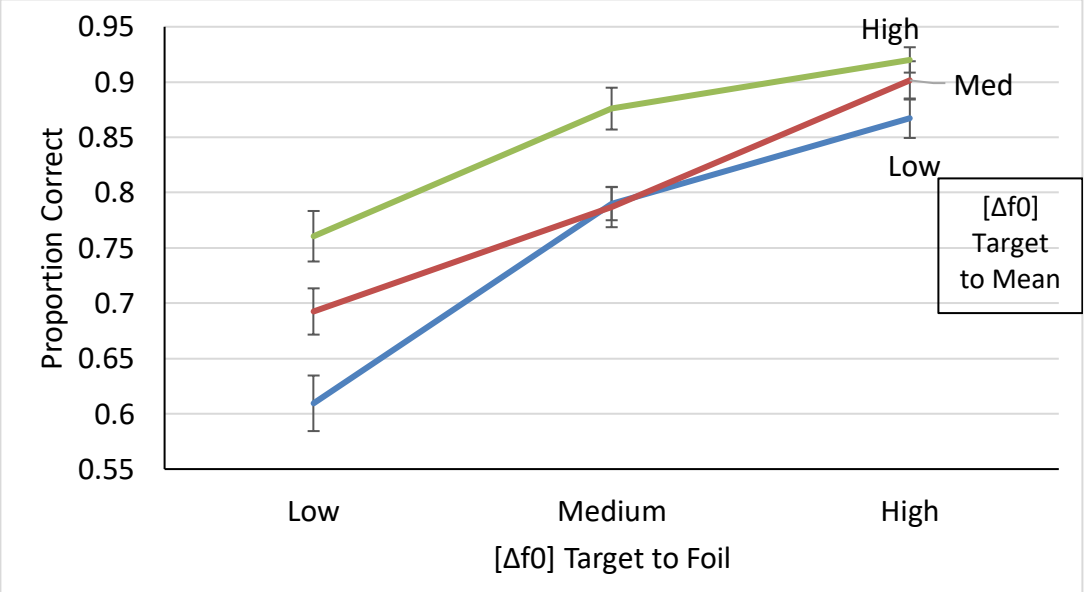


Figure 3.15: Interaction between fundamental frequency distance to foil and distance to mean (distinctiveness). At low difference to foil, more distinct target voices are better recognized. The [Δf0] values for each Target to Foil bin are: Low, >16 Hz; Medium, 16-32 Hz; High, >32 Hz. The [Δf0] values for each Target to Mean bin are: Low, >8.4 Hz; Medium, 8.4-20 Hz; High, >20 Hz. Error bars are 1 SE.

The results for the three-way interactions between each parameter's distance to mean, Familiarity, and Match Case. were somewhat surprising. We had expected that increasing distinctiveness of the parameters would affect both

match and foil trials, but match recognition was only sensitive to syllable rate (as shown in Fig. 3.18 below).

Looking at each of the distinctiveness parameters in turn, Fundamental Frequency had a significant main effect of improving recognition as the target's f0 became further from the mean, $F(2,350) = 5.95$, $p < .003$, $\eta_p^2 = .034$. As suggested above, distinctiveness did not affect match recognition at either familiarity level. There was a significant interaction between Familiarity and f0 distance to mean, $F(2,350) = 4.91$, $p < .008$, $\eta_p^2 = .028$, with high Familiarity increasing sensitivity to distinctiveness on the foil trials. The interaction between Match Case and f0 distance to mean was not significant, $F(2,350) = 1.10$, *n.s*. However, the three way interaction, shown in Figure 3.16 was significant at $F(2,350) = 6.12$, $p < .002$, $\eta_p^2 = .035$. The increased foil performance at the low distinctiveness level for the low Familiarity trials is likely due to the criterion shift seen in the prior section; on low familiarity trials, subjects were more conservative (likely to say "not match"). This is enhanced if the expected voice had a non-distinctive pitch.

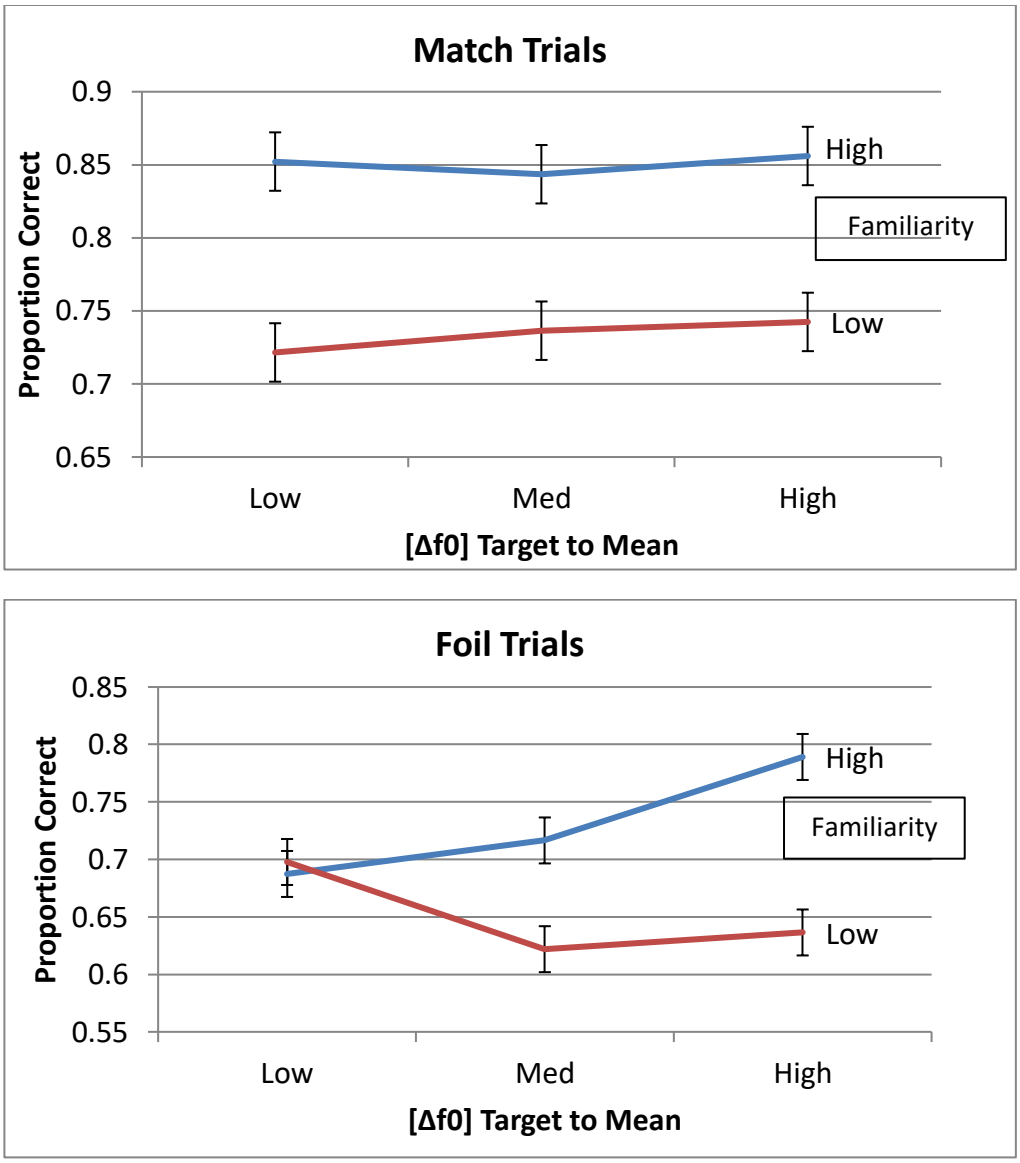


Figure 3.16: Interaction between Fundamental Frequency distinctiveness (distance to mean voice), Familiarity, and Match Case. Match recognition (top) was not affected by distinctiveness of the target voice; correct Foil rejection (bottom) improved with increasing f0 distance when subjects were highly familiar with the target. Error bars are 1 SE.

The relationship with subharmonic-to-harmonic ratio distance to Familiarity and Match Case, shown in Figure 3.17, was similar to those with fundamental frequency. There was a significant interaction between SHR distance and Match Case, $F(2,350) = 14.76$, $p < .001$, $\eta_p^2 = .082$, wherein match recognition was

unaffected by differences in SHR distinctiveness but foil performance improved moderately (from medium to high) though the effect was a weak one; $F(2,350) = 3.15$, $p < .044$, $\eta_p^2 = .019$. The interaction with familiarity was not significant, $F(2,350) = 2.06$, *n.s*. Highly familiar trials showed greater sensitivity to SHR distinctiveness in the foil trials, but the overall three-way interaction fell short of significance – $F(2,350) = 2.45$, $p < .087$.

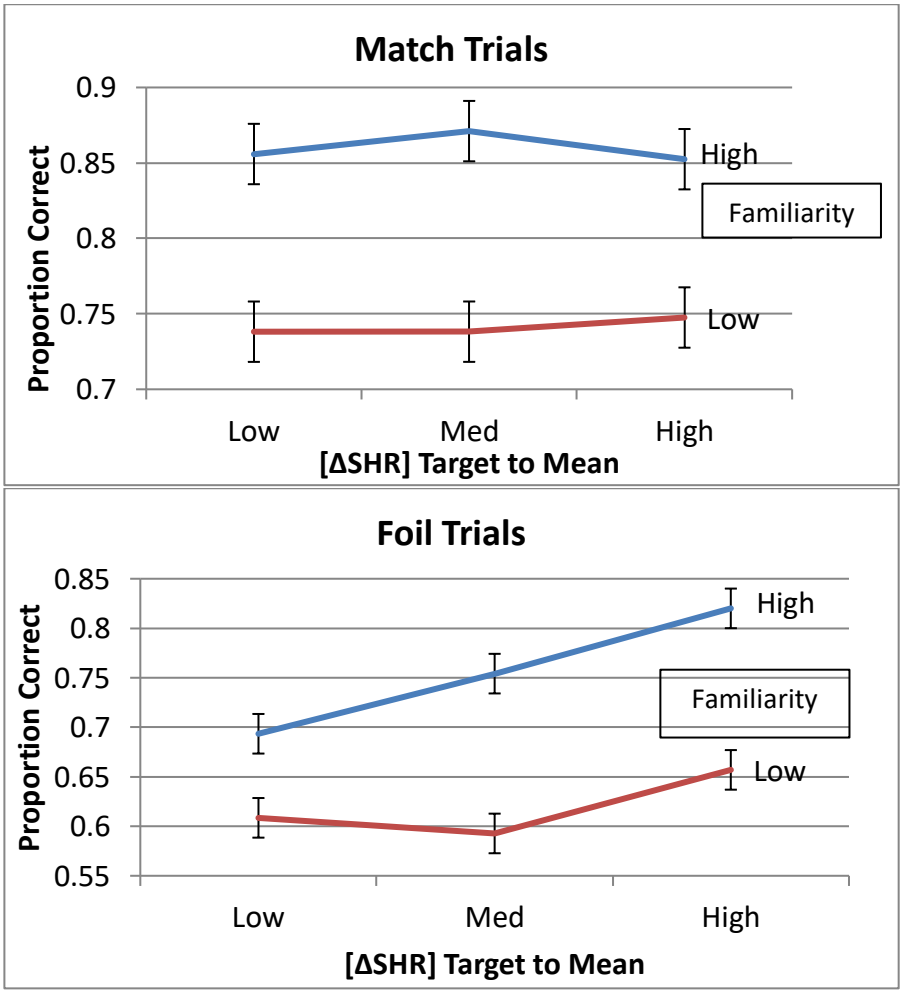


Figure 3.17: Interaction between SHR distinctiveness (distance to mean voice), Familiarity, and Match Case. Match recognition (top) was not affected by distinctiveness of the target voice; correct Foil rejection (bottom) improved with increasing SHR distance when subjects were highly familiar with the target. Error bars are 1 SE.

The third distinctiveness parameter that was assessed was speaking rate, which had a main effect in the three-way ANOVA of $F(2,350) = 28.27$, $p < .001$, $\eta_p^2 = .154$. Low Familiarity Speaking rate, had only a minimal effect on Low Familiarity trials, High Familiarity did (Figure 3.18). There were significant interactions with Speaking Rate and Match Case, $F(2,350) = 6.31$, $p < .002$, $\eta_p^2 = .039$, as well as Speaking Rate and Familiarity, $F(2,350) = 6.18$, $p < .002$, $\eta_p^2 = .038$. Both match recognition and foil rejection showed improvement in High Familiarity trials at the most distinct levels of Syllables/sec. The three-way interaction was not significant, $F(2,350) = .422$, *n.s.*

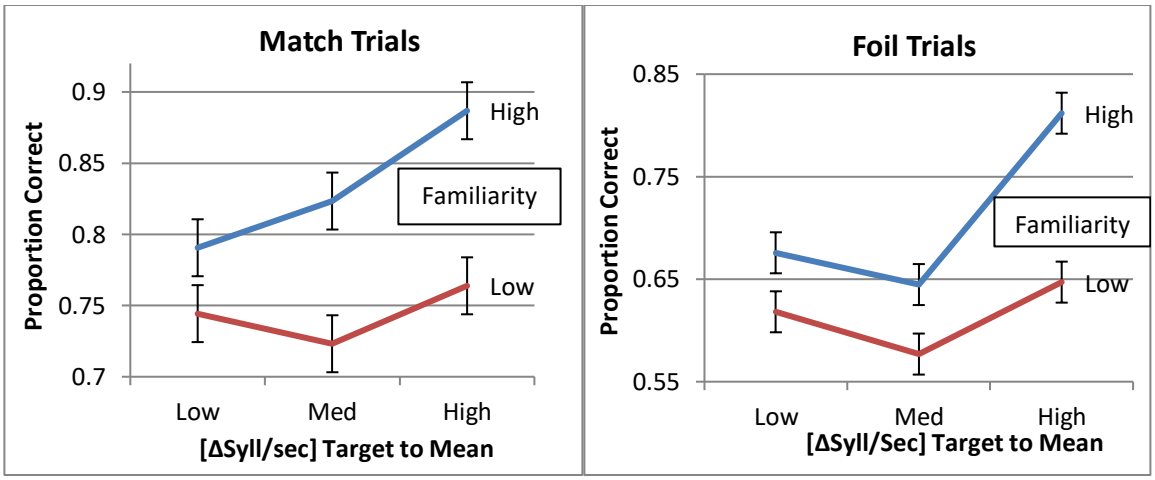


Figure 3.18: Effect of Distinctiveness of Speaking Rate (in syllables per sec), Familiarity, and Match Case. Highly distinctive speaking rates improved both Match recognition and Foil rejection in High Familiarity trials. Error bars are 1 SE

3.4.4. Two types of distinctiveness, three dimensions of voice features

Since we found pronounced effects of the target-to-foil difference and the target-to-mean difference, with interactions between the parameters, we next looked at how the parameters work in concert. For each type of distinctiveness (-to-foil and -to-mean), we binned the parametric differences into “Small” and

"Large" by dividing the set of voices at the median value for each parameter. Figure 3.19 shows the improvements in recognition accuracy by the number of parameters that vary by a "Large" (greater than median) amount. With both forms of distinctiveness, there was marked improvement in recognition when the voices differed greatly on more than one dimension, with highest recognition accuracy on trials in which all three parameter differences exceeded the median.

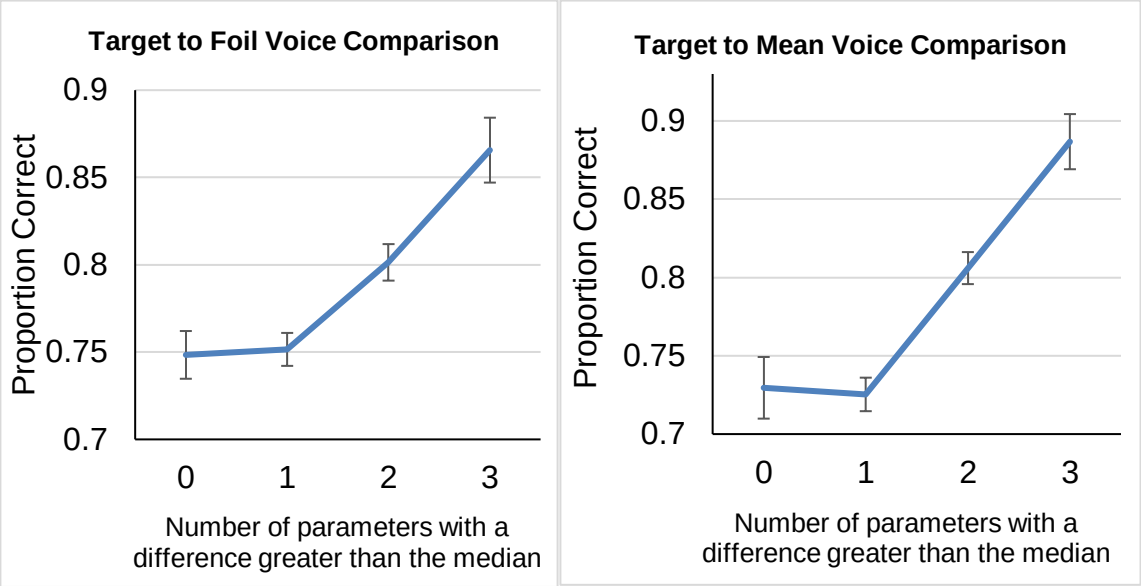


Figure 3.19: Recognition accuracy by number of parameters with large differences. All three voice parameters were binned into small and large differences (less than and greater than the median value for that parameter) within both measures of distinctiveness, target-to-foil (left) and target-to-mean (right). Recognition accuracy improved when the two voice values differed by a large amount on at least two parameters, and the highest accuracy occurred when all three parameters differed greatly. Error bars are 1 SE.

3.5 Linear Regression Analyses of Familiar Voice Recognition

We performed a linear regression analysis incorporating ten parameters to predict the accuracy of a given trial. The ten variables included in the analysis were: 1) familiarity rating 2) clip length 3) distinctiveness rating 4) f0 difference between

celebrity and foil voices 5) SHR difference between celebrity and foil voices 6) syllables per second difference between celebrity and foil voices 7) celebrity voice f0 absolute deviation from mean f0 8) celebrity voice SHR absolute deviation from mean SHR 9) celebrity voice speaking rate absolute deviation from mean 10) celebrity voice speaking rate raw deviation from the mean. As seen in the prior section, the differences between the three measured voice parameters for the celebrity versus foil are predictive of recognition accuracy. The last four variables in the regression were included as measures of objective measures of celebrity voice distinctiveness; each voice parameter was compared to the mean of the voice samples from this set. In the regression analyses, the celebrity-to-mean f0 and SHR differences were the absolute values of the differences, the distance to that voice's sex mean value. The absolute distance in speaking rate was only significant as a predictor of response time; however, the raw difference (celebrity speaking rate minus mean) proved highly predictive of accuracy. The raw differences for f0 and SHR did not have significant effects in any of the analyses.

The first regression analysis was a step-wise linear regression of the aforementioned ten variables, including all levels of familiarity. The results of this analysis can be seen in Table 3.3. These step-wise regression results show each variable in order of its contribution to R-squared. Variables were dropped from the model if the significance of their contribution to a change in F was $p > .100$; this criterion left the absolute SHR and speaking rate differences between target and the mean out of the regression analysis in Table 3.3 The total R-Square was .367 (adjusted R-square = .360), $F(8, 725) = 52.49$, $p < .001$, $\eta_p^2 = .367$.

Table 3.3:

Regression Model of Familiar Voice Recognition Parameters

Variable	Unstandardized Beta	Std. Error	Standardized Beta	t-value	Significance
(Constant)	.072	.078		.926	.355
Familiarity Rating (1-5)	.047	.004	.372	12.16	<.001***
Target to Foil [∆f0]	.003	.000	.259	6.59	<.001***
Clip Length	.022	.004	.146	4.93	<.001***
Distinctiveness Rating (1-5)	.077	.019	.126	4.03	<.001***
Target to Foil [∆SHR]	.472	.114	.136	4.12	<.001***
Target to Foil [∆Syllables/sec]	.032	.008	.122	3.78	<.001***
Target to mean ∆Syllables/sec	.019	.005	.110	3.48	<.001***
Target to Mean [∆f0]	-.001	.000	-.081	-2.25	<.025**

Clip length, familiarity rating, distinctiveness, and five different voice parameters (f0, SHR, and syllables/sec differences between target and foil; f0 and syllables/sec differences between target and mean) predict accuracy of a given trial at R = .606, R-Square = .367, adjusted R-Square = .360

The subsequent analysis looked at only high-familiarity trials (familiarity >3); the results are seen in Table 3.4 below. Familiarity was dropped as a measured variable (at the high levels the contribution was not significant at $p < .05$); all other variables' positions in order of variance-explained were preserved, except for target to mean f0 difference, which was dropped for lack of statistical significance. The fit of the model was better in this high familiarity analysis – R = .668, R-Square = .447, adjusted R-Square = .423, $F(6, 143) = 19.23$, $p < .001$, $\eta_p^2 = .457$.

Table 3.4:

Regression Model of Highly Familiar Voice Recognition Parameters.

Variable	Unstandardized Beta	Std. Error	Standardized Beta	t-value	Significance
(Constant)	.142	.129		1.097	.275
Target to Foil [∆f0]	.002	.001	.301	3.78	<.001***
Clip Length	.027	.007	.237	3.80	<.001***
Distinctiveness Rating (1-5)	.104	.031	.230	3.42	.001***
Target to Foil [∆SHR]	.612	.189	.227	3.23	.002**
Target to Foil [∆Syllables/sec]	.035	.013	.181	2.68	.008**
Target to mean ∆Syllables/sec	.021	.008	.167	2.48	.014**

Clip length, distinctiveness, and four different voice parameters -- f0, SHR, and syllables/sec differences between target and foil; syllables/sec difference between target and mean -- predict accuracy of a given trial at R=.668, R-Square = .447, adjusted R-Square = .423.

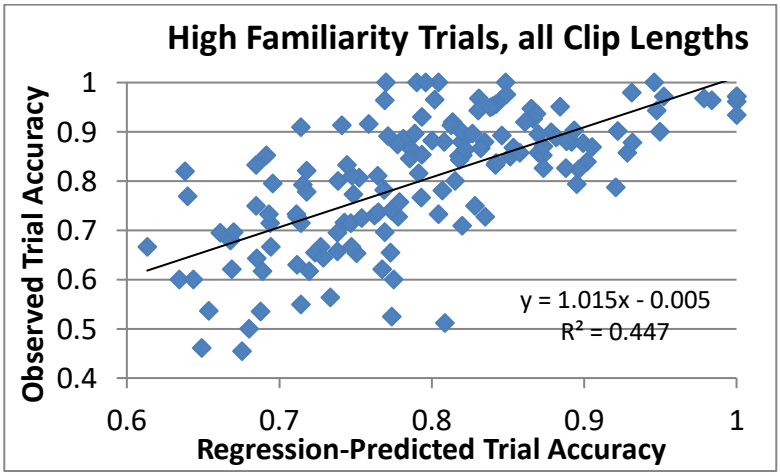


Figure 3.20: Model Fit of Linear Regression of Highly Familiar Voice Recognition. A linear regression model accounting for six variables – clip length, distinctiveness ratings, target-to-foil absolute difference in f0, SHR, and syllables per second, and target-to-mean raw difference in syllables per second – accounts for 44.7% of the variance of subjects' accuracies recognizing highly familiar voices (rated 4 or 5 on the 5 point scale).

Figure 3.20 (above) shows a plot of the high familiarity linear regression model results. We plotted each trial's model-predicted accuracy (calculated from the beta weights in Table 3.2) against its observed accuracy to visualize the goodness of fit. Figure 3.21 shows similar plots for trials of each clip length. The total R-Square for all clip lengths was .447; the regression had a better fit at longer clip lengths, increasing from 1 to 2 to 4 second stimulus lengths with respective R-Squares of .376, 458, and .483.

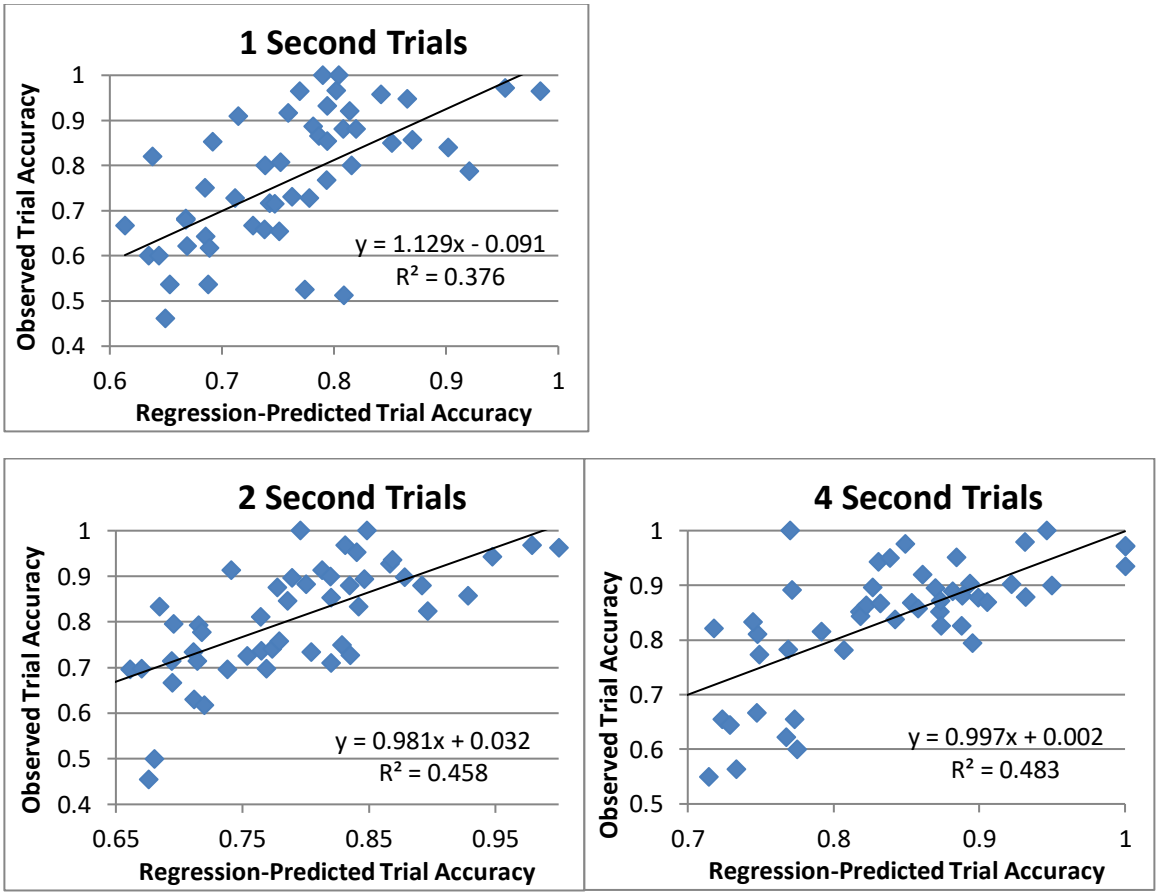


Figure 3.21: Model fit of linear regression model at each clip length. Using the same regression model as in Figure 3.20, we plotted each clip length individually. R-square improved with clip length. R-Square for one second clips was .376, for two second clips, .458, and for four second clips, .483.

The voice parameters themselves are correlated with each other, as seen in Table 3.5 below. This table shows the zero-order correlations between the measured parameters of the regression models, as well as the two dependent variables - accuracy and correct RT - for high familiarity targets.

Table 3.5:

Pearson correlations between all regression parameters, accuracy, and RT.

		Accuracy	Correct RT	Distinct Rating	[Δf0] Target-Mean	[Δf0] Target-Foil	[ΔSHR] Target-Foil	[SHR] Target-Mean	[ΔSyll/sec] Target-Foil	[ΔSyll/sec] Target-Mean	ΔSyll/sec Target-Mean
Accuracy	Corr *(r)*	1	-0.21	0.25	0.26	0.51	0.35	0.07	0.29	0.29	0.21
	Sig.(2-*tail*)		0.009**	0.002***	0.002**	<.001***	<.001***	0.388	<.001***	<.001***	0.012*
Correct RT	Corr *(r)*	-0.21	1	-0.16	-0.16	-0.21	-0.23	<.001	-0.12	-0.09	-0.3
	Sig.(2-*tail*)	0.009***		0.049*	0.046*	0.011*	0.004**	0.974	0.135	0.292	<.001***
Distinct Rating	Corr *(r)*	0.25	-0.16	1	0.04	0.17	-0.01	-0.13	0.08	-0.01	-0.09
	Sig.(2-*tail*)	0.002**	0.049*		0.668	0.038*	0.919	0.125	0.336	0.899	0.262
[Δf0] Target-Mean	Corr *(r)*	0.26	-0.16	0.04	1	0.53	0.23	-0.03	0.31	0.32	0.21
	Sig.(2-*tail*)	0.002**	0.046*	0.668		<.001***	0.005**	0.696	<.001***	<.001***	0.008**
[Δf0] Target-Foil	Corr *(r)*	0.51	-0.21	0.17	0.53	1	0.39	0.05	0.3	0.31	0.23
	Sig.(2-*tail*)	<.001***	0.011*	0.038*	<.001***		<.001***	0.581	<.001***	<.001***	0.006**
[ΔSHR} Target-Foil	Corr *(r)*	0.35	-0.23	-0.01	0.23	0.39	1	0.39	-0.01	0.03	-0.02
	Sig.(2-*tail*)	<.001***	0.004**	0.919	0.005**	<.001***		<.001***	0.922	0.68	0.784
[ΔSHR] Target-Mean	Corr *(r)*	0.07	<.001	-0.13	-0.03	0.05	0.39	1	-0.31	0.01	0.06
	Sig.(2-*tail*)	0.388	0.974	0.125	0.696	0.581	<.001***		<.001***	0.868	0.466
[ΔSyll/sec] Target-Foil	Corr *(r)*	0.29	-0.12	0.08	0.31	0.3	-0.01	-0.31	1	0.35	-0.04
	Sig.(2-*tail*)	<.001***	0.135	0.336	<.001***	<.001***	0.922	<.001***		<.001***	0.595
[ΔSyll/sec] Target-Mean	Corr *(r)*	0.29	-0.09	-0.01	0.32	0.31	0.03	0.01	0.35	1	0.39
	Sig.(2-*tail*)	<.001***	0.292	0.899	<.001***	<.001***	0.68	0.868	<.001***		<.001***
ΔSyll/sec Target-Mean	Corr *(r)*	0.21	-0.3	-0.09	0.21	0.23	-0.02	0.06	-0.04	0.39	1
	Sig.(2-*tail*)	0.012*	<.001***	0.262	0.008**	0.006**	0.784	0.466	0.595	<.001***	

All voice parameters correlate with each other in at least one way (target-to-foil, target-to-mean, or both). Note that the three voice distinctiveness parameters – comparing each target's measured f0, SHR, and syll/sec against the mean value – did not correlate with subjects' reported distinctiveness ratings.

3.6. Subjects' Distinctiveness Ratings

While the overall correlation between the mean distinctiveness rating for the trials and each of their distinctiveness measurements (seen in Table 3.6 above), subjects did show some sensitivity the speech parameters when we performed a repeated measures ANOVAs looking at each subject's mean distinctiveness rating for each of the objective distinctiveness categories. Subjects' subjective distinctiveness ratings increased linearly between the low and high target-to-mean distances for SHR, $F(2,348) = 8.92$, $p < .001$, $\eta_p^2 = .049$. When looking at f0 distance to mean, the low and medium distances (>8.4 Hz and 8.4-20 Hz) showed similar distinctiveness ratings, with a marked increase at the high level (>20 Hz); , $F(2,348) = 13.57$, $p < .001$, $\eta_p^2 = .072$. Subjects showed no significant sensitivity to speaking rate in their distinctiveness ratings, which remained constant over the three speaking rate bins, $F(2,348) = 1.52$, *n.s*. These three relationships are represented in Figure 3.22 below.

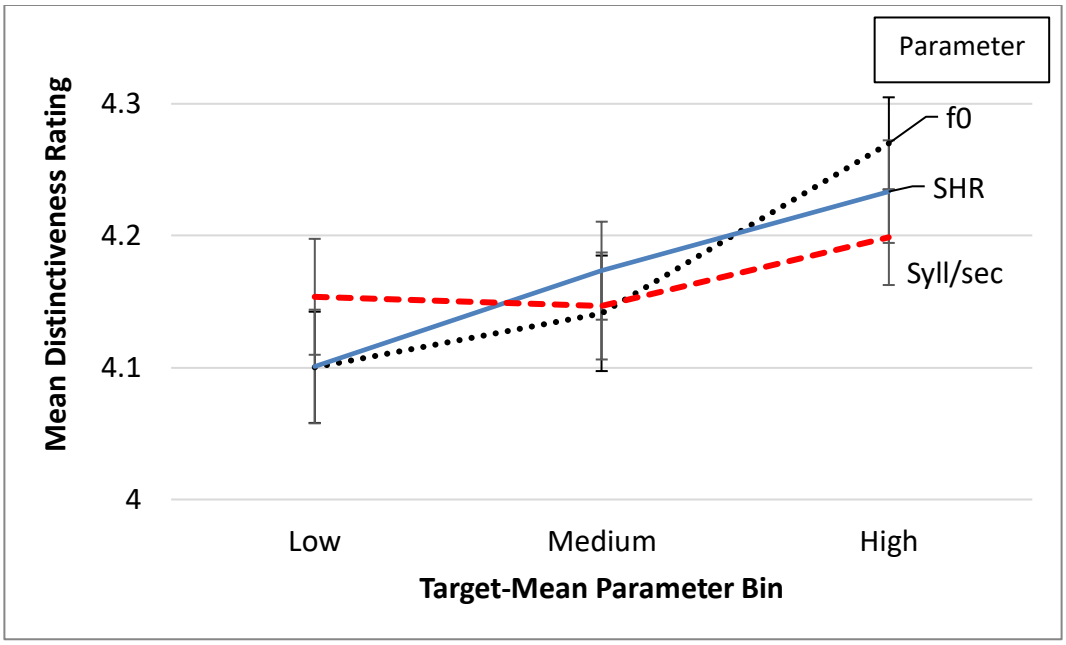


Figure 3.22: Mean distinctiveness ratings of each parameter distinctiveness bin. There was a linear increase in ratings as SHR distance to mean increased; ratings increased only the highest f0 category; syllable rate showed no relationship with subjective ratings. Error bars are 1 SE.

Finally, we looked at the relationship between subjects' familiarity ratings and their distinctiveness ratings; while the correlation between Familiarity and Distinctiveness was very high, $r = .86$, $p < .001$, we found successful differential usage of the scales by looking at performance at each distinctiveness rating for low familiarity and high familiarity trials. We binned Familiarity into Low (2 and 3 ratings) and High (4 and 5 ratings) and Distinctiveness into >3, 4, and 5, and ran a 2x3 repeated measures ANOVA on accuracy. As seen in Figure 3.23, accuracy improved linearly with increasing distinctiveness ratings, but only on high familiarity trials. At low familiarity, subjects' assessments of distinctiveness were likely metacognitively unreliable. The main effects of higher recognition with higher ratings were both significant, with Familiarity at $F(1, 175) = 63.96$, $p < .001$, $\eta_p^2 =$

.268, and Distinctiveness at $F(2,350) = 3.87$, $p < .022$, $\eta_p^2 = .022$. The interaction was significant at $F(2,350) = 3.55$, $p < .030$, $\eta_p^2 = .020$.

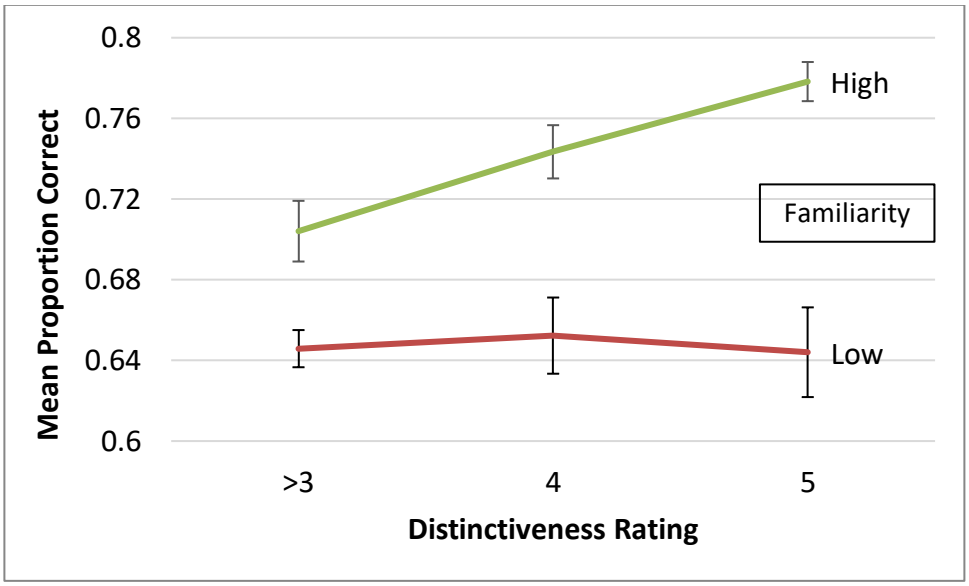


Figure 3.23: Performance by Distinctiveness rating and Familiarity. Subjects' Distinctiveness ratings were predictive of accuracy, but only on highly familiar (>4 rating) trials. Error bars are 1 SE.

4. Discussion

The present study systematically characterizes the nature of familiar voice recognition, using famous celebrity and non-famous foil voice samples. We found main effects of degree of familiarity and subjective rating of voice distinctiveness with each target, length of voice stimulus, and three voice parameters - fundamental frequency, subharmonic-to-harmonic ratio, and syllables spoken per second. Looking at these three parameters as they relate to the distance between a target voice and a similar sounding voice, as well as the distance between a target voice and a sex-based average voice, we found that these three parameters

all significantly contribute to familiar voice recognition, with improved sensitivity with higher familiarity.

It is important to note that the selection of the voice parameters was made independently of the selection of the voice samples themselves, so they likely represent a somewhat unbiased sampling as to how these parameters vary from person to person. These voice parameters were chosen for their constancy over the three clip lengths. Multiple other parameters available within the VoiceSauce analysis program were investigated – multiple differences in harmonic peaks (e.g. H1-H2, H2-H4, etc.), Formants 1-4, Cepstral Peak Prominence, and Root Mean Square Energy. These have all been implicated in characterizing perceived similarity amongst voices (Keating et al., 2015; Kreiman et al., 2017), but unfortunately these parameters varied too greatly between clip lengths to be useful for our analyses. Of course, other parameters distinguish voices as well (e.g. Beckman, 1996; Krieman & Sidtis, 2011), but incorporating exhaustive parametric analysis of auditory features that vary by the phoneme was simply beyond the scope of the present study.

Looking at the main effects of Clip Length and Match Case, we found that positive matches occur within two seconds of voice stimulus; there was no improved performance between two and four seconds for any familiarity level. Conversely, subjects had as many false positives at moderately high familiarity (4) with two seconds of stimulus as they did at all lower levels of familiarity. At the highest familiarity, there was differential performance at all clip lengths up to four seconds. Miss rates were lower than false alarm rates at all lengths and levels of

familiarity; subjects have a liberal bias that increases not just with increasing familiarity by also with decreasing clip lengths. Under cases of higher uncertainty (minimal auditory stimulus), subjects perceive the most familiar voices in the ambiguity. Similar to the evolutionary explanation for pareidolia (e.g. Taubert et al., 2017), in which people tend to find familiar patterns in randomness, when it comes to pattern recognition of any modality, missing a familiar pattern is much more costly than a false alarm. Granted, it is well documented that criterion can be shifted with various reward structures (Ackermann & Landy, 2015; Frithson et al, 2018). However, given our absence any reward structure (we didn't provide feedback of any kind) as well was reports of even higher rates false alarms (up to 50% false alarm rate) of familiar voice recognition under conditions of greater uncertainty (Schweinberger et al., 1997; Lavner et al., 2000; Yarmey et al., 2001), it is likely an innate feature of the voice recognition process.

But, as we have found, false alarm rates vary as a function of the physical properties of the auditory stimuli. It well documented that familiar voice recognition and unfamiliar voice discrimination are separate, dissociable properties (Van Lancker & Krieman, 1989; Stevenage, 2018). Our own case study of AN (Xu et al, 2015) contributed to this knowledge, demonstrating AN's perfectly intact discrimination ability but severely poor familiar voice recognition ability. The literature argues that unfamiliar voice discrimination is a comparison of lower level features (Van Lancker & Krieman, 1989; Krieman & Sidtis, 2011) and that familiar voice recognition is a Gestalt-like pattern more predominantly influenced from top-down processing (Krieman & Sidtis, 2011; Stevenage, 2018; Maguinness et al,

2019). This latter view seems reasonable given the performance on trials; match trials were fast and accurate, with subjects performing as well in *less* than a second than those hearing four full seconds of highly familiar voice stimulus, regardless of any discernable differences between the voices (measured in the distinctiveness parameters). Two or three syllables matching the expected voice pattern was sufficient for recognition.

However, the present study also shows the importance of lower level, quantifiable perceptual comparisons to the converse side of positive recognition, successful discrimination of familiarity itself. This is a unique and perhaps the most valuable contribution of the present study. While many voice studies show sensitivity to various voice parameters in discrimination of unfamiliar voices (Belin et al, 2004; Baumann & Belin, 2010; Garellek et al, 2016), this is the first known study to look at how voice parameters contribute to discrimination of a voice against a highly familiar, long term memory pattern.

As seen in the regression model, subjects are most sensitive to differences in fundamental frequency. From the repeated measures analysis of the different f0 target-to-foil distances, differences of 32 Hz allow for successful foil performance, comparable to positive match performance, with just one second of stimulus. Allowing for two seconds of stimulus, subjects can discriminate differences as low as 16 Hz with the same ceiling performance rate. At the lowest levels of f0 difference, less than 16 Hz, subjects hit the foil performance floor, with false alarm rates of 35%. Furthermore, looking at the interactions with the other two measured

variables, an f0 difference greater than 32 Hz is sufficient for at or nor ceiling performance regardless of the differences in the other two parameters.

However, at f0 differences less than 32 Hz, differences in subharmonic-to-harmonic ratio help subjects to discriminate foils. When the SHR difference between target and foil is low, <.045, performance is at the floor; when it's high, > .095, performance is close to ceiling, regardless of clip length. At low f0, SHR improves recognition at the highest level only, but when f0 differs at the medium (16-32 Hz) level, SHR differences between .045 and .095 bring performance to ceiling.

Speaking rate contributed the least, as per the regression model, but still in a significant way. There was monotonic improvement of recognition with increasing syllable rate at the two and four clips; subjects were not sensitive to the high syllabic differences (> 1.5 syllables/sec) at one seconds of stimulus. So, in the two and four second clips, speaking rate differences contributed additively to differences in the other parameters, but at one second lengths were simply less utilized in the discrimination process. Voices with particularly fast speaking rates (>1 syllable/sec more than average) showed improved recognition compared to the slower rates. This is likely a combination of the unique quality of a fast speaker (e.g. Mila Kunis) and the added benefit of extra syllables, each conveying more information of the other parameters' differences as well.

Regarding the distinctiveness of each voice on each parameter, as measured by the distance to the sex-based mean, it was interesting to find that

distinctiveness of f0 or SHR did not affect positive matching. This perhaps speaks less to a diminished importance of distinctiveness than it does to the high performance of the matching process -- especially in this design, where the expected voice pattern is limited to a single identity. If we opened the possible set to four celebrities as in our prior voice studies (Xu et al., 2015; Shilowich & Biederman, 2016) or had an entirely open set like one used by Schweinberger et al. (1997), distinctiveness may play a much larger role. Also, clip lengths shorter than one second, reduced to one or two syllables, might reveal distinctiveness differences.

While match performance hit ceiling at all levels of distinctiveness, even within 8.4 Hz f0 difference or .025 SHR difference from the mean voice, there was noticeable differential performance on the foil trials. This was most evident at the highest levels, with voices greater than 20 Hz f0 difference, 054 SHR, or 1.15 syllables/sec from the mean showing marked improvement in recognition compared to more average voices. This sensitivity to distinctiveness on these parameters was noticeably absent on less familiar trials (those rated 2 or 3).

This relationship between familiarity and distinctiveness was evident in how subjects subjectively rated the distinctiveness of voices. At low levels of familiarity, there was no correlation between distinctiveness ratings and performance. However, at the higher levels of familiarity (>3), the accuracy of subjects on trials of different distinctiveness ratings showed significant within-subject improvement as the target was rated as being more subjectively distinct. Looking at the relationship between the objective distinctiveness parameters and the

distinctiveness ratings, voice more distinct in f0 and SHR were rated as more subjectively distinct. The difference was subtle, with a mean of 4.25 distinctiveness rating of the most objectively distinct voices versus a mean of 4.1 for the lowest bins, but significant.

This is partly due to the fact that subjective distinctiveness ratings incorporate a multitude of features not captured by our measures, and maybe not capturable by any objective measure. As we saw, distinctiveness and familiarity are highly correlated, $r = .86$. While one explanation is that more distinct voices are more memorable and therefore will be rated as more familiar for a given level of exposure, it is more likely the case that familiarity itself causes an increase in subjective distinctiveness. Simply consider how distinct two identical twins seem when you first meet them versus having known them for years. The objective differences haven't changed, but you've learned a multitude of differentiating cues, some of which may even be unconscious to you. This is perhaps why distinctiveness ratings contributed significantly to the explanation of variance within the regression and the objective measures fell short; the subjective distinctiveness ratings are a metacognitive proxy for many parameters.

5. Conclusion

By studying familiar voice recognition using conversational clips of celebrity voices, paired against similarly sounding nonfamous foils, we could successfully explain 45% of the variance of familiar voice recognition performance using just six variables – 4 voice difference parameters (3 comparisons to the foil voice – f0,

SHR, and syllables/sec – and one distinctiveness measure – speech rate of the famous target), Clip Length, and a subjective rating of distinctiveness. By comparing trials of differing familiarity ratings, we found that sensitivity to these parameters increases with familiarity, helping to bridging the gap between unfamiliar voice discrimination and familiar voice recognition.

References

Ackermann, J. F., & Landy, M. S. (2015). Suboptimal decision criteria are predicted by subjectively weighted probabilities and rewards. *Attention, Perception, & Psychophysics*, *77*(2), 638-658.

Barsics, C. G. (2014). Person recognition ss easier from faces than from voices. *Psychologica Belgica*, *54*(3), 244-254.

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research PRPF*, *74*(1), 110.

Beckman, Mary E. (1996) The Parsing of Prosody. *Language and Cognitive Processes*, 11(1/2), 17-67.

Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in cognitive sciences*, *8*(3), 129-135.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. Nature, 403, 309–312.

Bethmann, A., Scheich, H., & Brechmann, A. (2012). The temporal lobes differentiate between the voices of famous and unknown people: an event-related fMRI study on speaker recognition. *PLoS One*, *7*(10), e47626.

Bishop, J. & Keating, P. (2012) Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex. *The Journal of the Acoustical Society of America*. 132; 1100-1112.

Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology*, *81*(3), 361-380.

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in cognitive sciences*, *11*(12), 535-543.

Damjanovic, L. (2011). The face advantage in recalling episodic information: Implications for modeling human memory. *Consciousness and Cognition*, *20*(2), 309-311.

Fontaine, M., Love, S. A., & Latinus, M. (2017). Familiarity and voice representation: from acoustic-based representation to voice averages. *Frontiers in psychology*, *8*, 1180.

Foulkes, P. & Barron, A. (2000) Telephone speaker recognition amongst members of a close social network. *International Journal of Speech Language and The Law*. 7(2) 180-198.

Frithsen, A., Kantner, J., Lopez, B. A., & Miller, M. B. (2018). Cross-task and cross-manipulation stability in shifting the decision criterion. *Memory*, *26*(5), 653-663.

Garellek, M., Samlan, R., Gerratt, B. R., & Kreiman, J. (2016). Modeling the voice source in terms of spectral slopes. *The Journal of the Acoustical Society of America*, *139*(3), 1404-1410.

Garrido, L., Eisner, F., McGettigan, C., Stewart, L., Sauter, D., Hanley, J. R., ... & Duchaine, B. (2009). Developmental phonagnosia: a selective deficit of vocal identity recognition. *Neuropsychologia*, *47*(1), 123-131.

Hacker, C. M., Meschke, E. X., & Biederman, I. (2019). A face in a (temporal) crowd. *Vision Research*, *157*, 55-60.

Hanley, J. R., Smith, S. T., & Hadfield, J. (1998). I recognise you but I can't place you: An investigation of familiar-only experiences during tests of voice and face recognition. *The Quarterly Journal of Experimental Psychology: Section A*, *51*(1), 179-195.

Keating, P. A., Garellek, M., & Kreiman, J. (2015, August). Acoustic properties of different kinds of creaky voice. In *ICPhS*.

Kreiman, J., Keating, P., & Vesselinova, N. (2017) Acoustic similarities among voices. Part 2: Male speakers. Poster presented at the Acoustical Society of America, New Orleans, LA, December.

Kreiman, J. & Papcun, G. (1991) Comparing discrimination and recognition of unfamiliar voices. *Speech Communication*, 10, 265-275.

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons.

Ladefoged, P. & Ladefoged, J. (1980) The ability of listeners to identify voices. *UCLA Working Papers in Phonetics* 49, 43-51

Latinus, M., & Belin, P. (2011). Anti-voice adaptation suggests prototype-based coding of voice identity. Frontiers in Psychology. 2:175.

Latinus, M., McAleer, P., Bestelmeyer, P.E.G., & Belin, P. (2013) Norm-Based Coding of Voice Identity in Human Auditory Cortex. *Current Biology*. 23,1-6.

Latinus, M., VanRullen, R., & Taylor, M. J. (2010). Top-down and bottom-up modulation in processing bimodal face/voice stimuli. *BMC neuroscience*, *11*(1), 36

Lavner, Y., Rosenhouse, J., & Gath, I. (2001) The Prototype Model in Speaker Identification by Human Listeners. *International Journal of Speech Technology*, 4, 63-74

Legge, G. E., Grosmann, C., & Pieper, C. M. (1984). Learning unfamiliar voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(2), 298.

Maguinness, C., Roswandowitz, C., & von Kriegstein, K. (2018). Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia*, *116*, 179-193.

Meschke, E.X., Hacker, C. M., & Biederman, I. (2018). How Many Faces Can You Recognize? Poster presented at the Annual Meeting of the Vision Sciences Society, St. Petersburg Beach, Fl. May.

Meudell, P. R., Northen, B., Snowden, J. S., & Neary, D. (1980). Long term memory for famous voices in amnesic and normal subjects. *Neuropsychologia*, *18*(2), 133-139.

Papcun, G., Kreiman, J., & Davis, A. (1989). Long-term memory for unfamiliar voices. *The Journal of the Acoustical Society of America*, *85*(2), 913-925.

Plante-Hébert, J., Boucher, V. J., & Jemel, B. (2017). Electrophysiological Correlates of Familiar Voice Recognition. *INTERSPEECH* 3907-3910.

Rose, P. & Duncan, S. (1995). Naïve auditory identification and discrimination of similar voices by familiar listeners. *Forensic Linguistics*. 2(1), 1-17.

Roswandowitz, C., Mathias, S. R., Hintz, F., Kreitewolf, J., Schelinski, S., & von Kriegstein, K. (2014). Two cases of selective developmental voice-recognition impairments. *Current Biology*, *24*(19), 2348-2353.

Schweinberger, S. R., Herholz, A., & Sommer, W. (1997). Recognizing famous voices: Influence of stimulus duration and different types of retrieval cues. *Journal of Speech, Language, and Hearing Research*, *40*(2), 453-463.

Schweinberger, S. R., Herholz, A., & Stief, V. (1997). Auditory long term memory: Repetition priming of voice recognition. *The Quarterly Journal of Experimental Psychology: Section A*, *50*(3), 498-517.

Schweinberger, S R., Kloth, N., and Robertson, D. M.C. Hearing facial identities: Brain correlates of face–voice integration in person identification. *Cortex* 47.9 (2011): 1026-1037.

Shilowich, B. E., & Biederman, I. (2016). An estimate of the prevalence of developmental phonagnosia. *Brain and Language*, *159*, 84-91.

Shue, Y.-L., P. Keating , C. Vicenik, K. Yu (2011) VoiceSauce: A program for voice analysis, Proceedings of the ICPhS XVII, 1846-1849.

Skuk, V. G & Schweinberger, S.R. (2014) Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender. *Journal of speech, language, and hearing research*. 57(1), 285-296.

Sorenson, M.H. (2012) Voice line-ups: Speakers' F0 values influence the reliability of voice recognitions. *International Journal of Speech Language and the Law*. 19(2), 145-158.

Stevenage, S. V. (2018). Drawing a distinction between familiar and unfamiliar voice processing: A review of neuropsychological, clinical and empirical findings. *Neuropsychologia*, *116*, 162-178.

Stevenage, S. V., & Neil, G. J. (2014). Hearing faces and seeing voices: The integration and interaction of face and voice processing. *Psychologica Belgica*, *54*(3), 266-281.

Stevenage, S. V., Hugill, A. R., & Lewis, H. G. (2012). Integrating voice recognition into models of person perception. *Journal of Cognitive Psychology*, *24*(4), 409-419.

Sun, X. (2002). Pitch determination and voice quality analysis using Subharmonic-to-Harmonic Ratio. IEEE International Conference on Acoustics, Speech, and Signal Processing. Orlando, FL, USA.

Taubert, J., Wardle, S. G., Flessert, M., Leopold, D. A., & Ungerleider, L. G. (2017). Face pareidolia in the rhesus monkey. *Current Biology*, *27*(16), 2505-2509.

Van Dommelen, W.A. (1990) Acoustic Parameters in Human Speaker Recognition. *Language and Speech*. 33(30), 259-272.

Van Lancker, D., & Canter, G. J. (1982). Impairment of voice and face recognition in patients with hemispheric damage. Brain and Cognition, 1, 185–198.

Van Lancker, D., & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia*, *25*(5), 829-834.

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters: I. Recognition of backward voices. *Journal of Phonetics, 13*(1), 19-38.

Xu, X., Biederman, I., Shilowich, B. E., Herald, S.B., Amir, O., & Allen, N. E. (2015). Developmental phonagnosia: Neural correlates and a behavioral marker. *Brain & Language*, 149, 106-117.

Yarmey, A. D., Yarmey, A. L., Yarmey, M. J., & Parliament, L. (2001). Commonsense beliefs and the identification of familiar voices. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *15*(3), 283-299.

Zaske, R., Volberg, G., Kovacs G., & Schweinberger, S.R. (2014) Electrophysiological Correlates of Voice Learning and Recognition. *The Journal of Neuroscience*. 34(33), 10821-10831.