INFORMATIONAL ASPECTS OF AUDIOVISUAL IDENTITY MATCHING

by

Lauren Wood Mavica

A Dissertation Submitted to the Faculty of

The Charles E. Schmidt College of Science

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

Florida Atlantic University

Boca Raton, FL

August 2016

INFORMATIONAL ASPECTS OF AUDIOVISUAL IDENTITY MATCHING

by

Lauren Wood Mavica

This dissertation was prepared under the direction of the candidate's dissertation advisor, Dr. Elan Barenholtz, Department of Psychology, and has been approved by the members of her supervisory committee. It was submitted to the faculty of the Charles E. Schmidt College of Science and was accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

SUPERVISORY COMMITTEE:

Elan Barenholtz, Ph.D.
Dissertation Advisor

Howard S. Hock, Ph.D.

Alan W. Kersten, Ph.D.

David L. Wolgin, Ph.D.
Chair, Department of Psychology

David J. Lewkowicz, Ph.D.

Sang Hong, Ph.D.

Janet Blanks, Ph.D.
Interim Dean, Charles E. Schmidt
College of Science

Deborah L. Floyd, Ed.D.
Dean, Graduate College

07/20/2016
Date

ACKNOWLEDGEMENTS

This research project would not have been possible without the support of many people. Foremost, I would like to express my sincere gratitude to my advisor Dr. Elan Barenholtz for the continuous support of my research, for his patience, motivation, and immense knowledge. I could not have imagined having a better mentor and could not say thank you enough for his guidance and support.

Besides my advisor, I would like to thank the rest of my committee: Dr. Alan Kerstin, Dr. David Lewkowicz, Dr. Sang Hong and especially Dr. Howard Hock, for their encouragement, insightful comments, and thought provoking questions.

ABSTRACT

Author: Lauren Wood Mavica

Title: Informational Aspects of Audiovisual Identity Matching

Institution: Florida Atlantic University

Dissertation Advisor: Dr. Elan Barenholtz

Degree: Doctor of Philosophy

Year: 2016

In this study, we investigated what informational aspects of faces could account for the ability to match an individual's face to their voice, using only static images. In each of the first six experiments, we simultaneously presented one voice recording along with two manipulated images of faces (e.g. top half of the face, bottom half of the face, etc.), a target face and distractor face. The participant's task was to choose which of the images they thought belonged to the same individual as the voice recording. The voices remained un-manipulated. In Experiment 7 we used eye tracking in order to determine which informational aspects of the model's faces people are fixating while performing the matching task, as compared to where they fixate when there are no immediate task demands. We presented a voice recording followed by two static images, a target and distractor face. The participant's task was to choose which of the images they thought belonged to the same individual as the voice recording, while we tracked their total fixation duration. In the no-task, passive viewing condition, we presented a male's voice recording followed sequentially by two static images of female models, or vice versa,

counterbalanced across participants. Participant's results revealed significantly better than chance performance in the matching task when the images presented were the bottom half of the face, the top half of the face, the images inverted upside down, when presented with a low pass filtered image of the face, and when the inner face was completely blurred out. In Experiment 7 we found that when completing the matching task, the time spent looking at the outer area of the face increased, as compared to when the images and voice recordings were passively viewed. When the images were passively viewed, the time spend looking at the inner area of the face increased. We concluded that the inner facial features (i.e. eyes, nose, and mouth) are not necessary informational aspects of the face which allow for the matching ability. The ability likely relies on global features such as the face shape and size.

FIGURES

Informational Aspects of Audiovisual Identity Matching

People can be identified by the presentation of either their face or their voice. The presentation of isolated faces and voices of unfamiliar people carries redundant information that may be used to infer specific characteristics of those individuals. For example, people can correctly infer properties such as height, weight, and age, based on static images of the faces of unfamiliar people, as well as based on recordings of their voices (Borkenau & Liebler, 1992; Lass & Davis, 1976). It has also been shown that people have the ability to discern which face and voice were recorded from the same unfamiliar person, first using moving stimuli (i.e. videos) (Kamachi, Hill, Lander & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004) and then using static images (Krauss, Freyberg & Morsella, 2002; Mavica & Barenholtz, 2013; Smith, Dunn, Baguley & Stacey, 2016a; Smith, Dunn, Baguley & Stacey, 2016b).

Presently, we know the ability to match faces and voices from static images exists but it remains uncertain as to what information people are using in order to perform this task. Based on redundant information, extracted from faces and voices, it seems people could be using information gained from a model's face in order to match the face to the corresponding voice. For example, you hear a voice that has the qualities of a tall person, in order to make the match you look for a face that also has the same qualities of tallness. The current study reviews the existing literature and investigates what informational aspects of faces could account for the ability to match an individual's face to their voice, using only static images.

Audiovisual Identity Matching

As mentioned above, people can be identified by their face or voice separately and common information, such as height and weight, can also be correctly inferred from both (Lass & Davis, 1976). Early studies investigating multisensory speech perception (Kamachi, Hill, Lander & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004) concluded that an inference could not be made as to which face and voice belong to the same unfamiliar person, when presented with static images of the faces. In particular, these studies suggested that people could only perform this task at above chance levels when the stimuli consisted of moving faces (i.e. videos), where temporal consistencies and nonverbal cues, such as those influencing prosody and intonation, across the two modalities are matched (Kamachi, Hill, Lander & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004).

In one study, Lachs and Pisoni (2004), participants were presented with a voice recording of a person speaking a single word. They were then presented two silent videos sequentially, one was the correct model saying the target word, and the other silent video was a different model saying the target word, used as a distractor. Participants had to infer which of the two videos were recorded from the same model as the voice recording. The order of presentation was also reversed; first a single silent video was presented, followed by two voice recordings to choose from. In both conditions participants were able to make the correct match at better than chance levels. In a following experiment, to test if the participants were using implicit knowledge of the mappings between vocal characteristics and facial features, they presented static images of the speakers, rather than silent videos, and used the same testing paradigm. However, participants were

unable to perform the matching task at above chance levels when the visual stimulus consisted of static images of the model's faces. Relying on this information, the authors suggested that that people could only perform this task at above chance levels when the stimuli consisted of moving faces, where the temporal consistencies across the two modalities could be used to make the match.

In a similar study, Kamachi, Hill, Lander and Vatikiotis-Bateson (2003), participants were presented with a silent video of an unfamiliar person speaking a full sentence. Next, they were presented with two voice recordings, each was a different sentence than was presented in the video. They also reversed the order and first presented a voice recording followed sequentially by two silent videos. By changing sentences between the videos and voice recordings they were able to show that identity-specific information about the models can be accessed across different utterances. The participants needed to choose which of the videos and voice recordings were recorded from the same model. The participants were able to correctly choose which stimuli matched only when the stimuli were presented as forward moving videos. In follow up conditions, when the auditory stimulus was presented backwards, or the visual stimulus was presented as static images, performance was reduced to chance. Similar to Lachs and Pisoni (2004), the authors suggested that the matching task was completed by the ability to map individual variations of nonverbal cues regulated by facial motion during speech and, therefore, the match could not be made with static images.

However, Krauss, Freyberg and Morsella (2002) showed that people were able to match voice recordings to static images of unfamiliar model's. They were interested in studying the phenomenon that occurs when you are surprised by a person's appearance

when you first meet them after having only talked to them on the phone or hearing them on the radio. In each trial, participants were presented with two voice recordings of full sentences, recorded from the same person, followed by two full-length, full body photographs, one of the target model and one of a distractor model. The task was to choose which photograph corresponded to the two voice recordings. They reported evidence of a cross-modal matching ability; in fact, the correct match was made 76.5% of the time. In a following experiment participants estimated the age, height, and weight of the models from either the presentation of their voice recordings or from the model's full body photographs. The results revealed that people are equally as accurate making estimates of the model's physical characteristics from their voices as they are from their photographs. Interested in how close, on average, the ratings were to the model's actual measurements, the absolute difference between the participant's ratings and the model's actual values were computed. Results here revealed that participant's ratings of a model's age, height, and weight, from a voice are nearly as close as their ratings from a full body photograph. The authors suggested that people could have performed the matching task by estimating the physical characteristics (e.g. age, height, and weight) from the voice of the model and matching based on the photograph that most closely corresponded to their estimates. However, they did not test whether accurate measurements or the degree of similarity in the ratings of the voice recordings and photographs predicted performance in the matching task.

While some previous studies (Kamachi, Hill, Lander & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004) suggested that people cannot match an unfamiliar voice to a static image of the face, these studies have several limitations. First, in natural settings, faces

and voices are normally perceived simultaneously. However, the aforementioned studies used a delayed match to sample testing paradigm, where the first face or voice presented must be held in an active state in working memory while the participants were presented, in sequence, with two alternatives to choose from. It should be noted that even with moving stimuli, the ability to accurately determine which faces and voices belonged to the same person was fairly weak; across both studies performance averaged about 60-65% correct (Kamachi, Hill, Lander & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004). Thus, if any ability was actually present in the static conditions, it might have been obscured by the added difficulty in the task because it required employing working memory. Second, in one of the studies (Lachs & Pisoni, 2004), the auditory stimuli presented were isolated words such as "cat", rather than full sentences. It is possible that the reason the participants were not able to make the correct match may have been due to insufficient auditory information.

Another important component to the Kamachi, Hill, Lander & Vatikiotis-Bateson (2003) study, suggesting that people do not have this ability, is the fact that their models were all of the same ethnicity, age and gender; they used Japanese females to film and record their stimuli. The use of the Japanese models could have hindered them in finding the matching ability for several reasons. First, Japanese females have been found to artificially raise the pitch of their voices (Horvat, 2000; Loveday, 1986), which could affect or mask the voice qualities which are relied on in the matching task. Second, as a whole in the Japanese culture, people have been found to spend less time fixating on faces during social interaction, as compared to other ethnic groups (Argyle & Cook, 1976). Therefore, Japanese participants could have different gaze behaviors or even

reduced expertize in the subtle features of the face (Horvat, 2000). In sum, the lack of statistical significance in Kamachi, Hill, Lander & Vatikiotis-Bateson (2003) could be related to the vocal characteristics of the Japanese models, or to cultural norms in the looking behavior of the participants.

As mentioned, Krauss, Freyberg and Morsella (2002) did find evidence of the matching ability using static images. Although, the full body photographs used in their study provided additional information, beyond what an isolated face would, and therefore, could not be directly compared to other studies that reported the matching task could not be performed at better than chance levels when the images were static (Kamachi, Hill, Lander & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004).

In order to directly compare the disparate findings Mavica and Barenholtz (2013) investigated this further. They presented participants with varying amounts of facial and vocal information, used two different testing paradigms, and gathered ratings data on several physical and personality dimensions. In Experiment 1, they simultaneously presented one voice recording along with a target face and a distractor face, using static images. In Condition 1 the face images were frontal headshots, including the neck and shoulders of the models. In order to ensure this ability was present across utterances they presented three different voice recordings, each a different statement. Participants had to choose which of the two faces they thought belonged to the same individual as the voice recordings. Because the auditory and visual stimuli were presented simultaneously, participants did not have to maintain the faces or voices in working memory in order to perform the task, which may have hampered performance in previous studies. The results indicated that people are able to choose, based on a static image of a face, what face and

voice belong to the same person. Next, in Condition 2, they removed bodily information (i.e. neck and shoulders) from the photographs to see if participants were using body shape and size in order to make inferences about how a person's voice should sound; they used the same audio stimuli as in the previous condition. Here, they again found the ability to do the match was still above chance levels. In Condition 3, in order to determine whether previous findings may have been due to insufficient auditory information available, they presented isolated words (i.e. "clouds") rather than full statements and still found significantly better than chance performance. Performance in Condition 1 and Condition 2 were both significantly better than in Condition 3; although, the first two conditions did not significantly differ from one another. Participants also rated the faces and voices along six physical dimensions (e.g. age, height, weight, attractiveness, socioeconomic status, and masculinity/femininity), and five personality dimensions (e.g. openness, conscientiousness, extraversion, agreeableness, and calmness). For each model, the degree of similarity between the ratings of their face and voice was turned into a difference score. Each model's difference score was compared to the participant's performance in the matching task. None of the dimensions were found to be significantly correlated with performance. Overall, the results from the ratings data did not support the view that the matching ability is due to matching faces and voices along these specific dimensions, as none of the factors significantly correlated with performance. In Experiment 2, in order to replicate previous studies that utilized a sequential match to sample task and found no ability to do the match when presenting static images (Kamachi, Hill, Lander & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004), Mavica and Barenholtz (2013) conducted a sequential match to sample task, rather than

simultaneous presentation. Here, they presented a voice recording followed sequentially by two static images, using the same images and voice recordings as in Experiment 1. Again, they found that accuracy was above chance and the ability persisted.

The results from Mavica and Barenholtz (2013) were confirmed in two later studies (Smith, Dunn, Baguley & Stacey, 2016a; Smith, Dunn, Baguley & Stacey, 2016b). In the first study (Smith, Dunn, Baguley & Stacey, 2016a), participants rated static images and silent videos of model's faces and voice recordings, separately, on dimensions of fitness and quality (e.g. masculinity/ femininity, age, health, height, and weight). Half of the participants rated silent videos of talking faces, and the other half rated static images of the faces; all participants rated the same full sentence voice recordings. It was found that the ratings for each face and voice, recorded from the same model, fall into a similar range, regardless of whether the ratings were from silent videos or static images. They did not compare performance in a matching task to the ratings data, like Mavica and Barenholtz (2013), but were just interested in redundant, or concordant information that could be separately assessed from the faces and voices. In the face and voice matching phase of the experiment, they presented silent videos and static images along with full sentence voice recordings and used a same-different testing paradigm, in order to rule out the possibility of a possible response bias. In each trial, participants were presented with a voice recording followed by either a silent video, or a static image of a face (depending on the condition), and responded if they thought the face and voice were matching (i.e. recorded from the same person) or not matching. The results confirmed earlier studies, and found that when using this testing paradigm both moving (Kamachi, Hill, Lander & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004) and

static images of faces (Krauss, Freyberg & Morsella, 2002; Mavica & Barenholtz, 2013) could be matched with the corresponding voice at above chance levels.

In their next study, Smith, Dunn, Baguley and Stacey (2016b) investigated previous contradictory results obtained (Kamachi, Hill, Lander & Vatikiotis-Bateson, 2003; Krauss, Freyberg & Morsella, 2002; Lachs & Pisoni, 2004; Mavica & Barenholtz, 2013) and in three experiments tested whether procedural differences could account for the inconsistencies. In their first experiment they presented a sequential two-alternative, forced choice matching task using both static images and silent videos of faces, along with full sentences. In the first condition, the face stimuli (e.g. static images or silent videos) was presented followed sequentially by two voice recordings; in the second condition, a voice recording was presented followed sequentially by two face stimuli (e.g. static images or silent videos). Participants had to choose which of the two alternatives they thought was the correct match. Here they found performance was better than chance when the face stimuli were silent videos but not when they were static images. Next, they used a simultaneous presentation testing paradigm. They presented two face and voice combinations one after another, using static images in one condition and silent videos in another condition. Participants had to choose which of the two alternatives they thought was the correct combination. Here, again, they found that performance was better than chance when the face stimuli were videos but not when they were static images. Finally, in their last experiment they first presented a voice recording, then simultaneously presented two static images of faces and found better than chance performance when using static images. Overall, Smith, Dunn, Baguley and Stacey (2016b) concluded that

procedural differences could account for the previously found inconsistencies in face and voice matching abilities.

Eye Tracking

Societal norms of appropriate eye gaze during face-to-face interactions are well established, and the eyes of a person attract the most interest (Argyle & Cook, 1976). Faces attract the interest of newborn infants (Atkinson & Braddick, 1989), a face preference is observed as soon as 9 minutes after birth (Goren, Sarty & Wu, 1975). In fact, from a very young age, 7-11 weeks, infants look more toward the eyes than to other regions of the face for static, dynamic and talking faces (Haith, Bergman, & Moore, 1977). The eyes are thought of as the "window into the soul" and attract interest because this is where one can access social cues (Langton, Watt & Bruce, 2000). However, the eyes are not the only source of information. The mouth of a speaker may carry emotional information, based on its expression but, it also contains important information related to the temporal and acoustic characteristics of speech (Lansing & McConkie, 2003). Recent research shows that when social and attentional cues are of specific relevance, the eyes are generally favored (Birmingham, Bischof & Kingstone, 2008; Emery, 2000). When, however, the task specifically requires speech processing, the mouth becomes a significant source of fixations as well, particularly during the actual speech utterance (Lansing & McConkie, 1999; Lansing & McConkie, 2003; Vo, Smith, Mital, & Henderson, 2012).

Relatedly, it has been shown that as noise increases in a speech signal the amount of time spent gazing at the mouth of a speaker increases from about 37% with no noise to 56% with noise that is so great that the speech is almost unintelligible (Vatikiotis-

Bateson, Eigsti, Yano & Munhall, 1998). Since the tendency to gaze more toward the mouth is particularly pronounced when the noise is degraded (Lansing & McConkie, 2003; Vatikiotis-Bateson, Eigsti, Yano & Munhall, 1998), this likely reflects the fact that redundancy from the visual articulations can enhance comprehension under noisy conditions (Sumby & Pollack, 1954). For example, Buchan, Pare, and Munhall (2007) tracked eye gaze while participants performed either emotion judgment or speech recognition tasks. When emotion judgments were performed gaze was directed toward the eyes, whereas gaze was directed more toward the mouth in the speech recognition task. In both conditions, however, when noise was added to the auditory signal fixations at the mouth increased. Overall, this indicates that the eyes of a person are not the only feature of interest, gaze is sensitive to the distribution of information in the face and is influenced by the task at hand. And, fixations at the mouth seem to take place on an 'as needed basis', based on the task and on the informational demands of the situation.

Similarly, Lansing and McConkie (1999) examined gaze behavior in silent speech reading and observed differences in gaze behavior when the speech recognition task emphasized recognition of individual words versus determining the intonation of a sentence. Gaze was more often at the mouth when identifying words and more often at the eyes when carrying out the intonation task. Interestingly, in a later experiment Lansing and McConkie (2003) added a one second still image of the first and last frame of the video to the beginning and the end of each trial; this provided a still image control condition. They found that during still image periods the participant's gaze was biased toward the model's eyes. Once the faces started to move the gaze then shifted to the

mouth. Therefore, the deployment of attention to certain parts of a face do depend on the current task and do take place on an 'as needed basis'.

The deployment of attention during speech perception in adults has been studied in the context of intelligibility in noise and in social behavior. Infants, as mentioned, from a very young age look more toward the eyes than to other regions of the face (Haith, Bergman, & Moore, 1977). Lewkowicz and Hansen-Tift (2012) demonstrated that when infants view speaking faces they shift their gaze from a speaker's eyes to the mouth when they are between 4-8 months of age. This shift is thought to occur in order to facilitate language learning, based on the audiovisual redundancy in the speech signal. An infant's main task during this period of infancy (i.e. 4-8 months) is to learn language, therefore, an increase in mouth fixations could reveal this as a language encoding strategy activated in order to enable the infants to encode unfamiliar speech. Evidence of gazing more toward the mouth is a sign that infants are actively engaged in language encoding. A second shift back to the eyes is seen at 12 months of age in response to the infant's native but not non-native language. When 12 month olds viewed videos of their native language they looked more at the eyes however, when they viewed videos of a non-native language they continued to look more toward the mouth. This was suggested to be in response of a growing expertise for their native language and unfamiliarity with the non-native language. The 12 month olds newfound language expertise frees them to shift their attention back to the eyes in order to gain social cues when looking and listening to their native, but not non-native, language.

Presumably, by 12 months of age, once infants begin to master their native speech, they no longer need to rely as much on the audiovisual redundancy of the speech

signal in order to encode it. This demonstrates that we find in infants, as well as adults, the allocation of attention to different features of the face, depending on the task at hand. Before language learning, infants gaze more toward the eyes of a speaker in order to gain important social cues. Then, older infants who are in the midst of learning language, allocate their attention to the mouth where more information for the task of language learning is contained. Finally, we can see gaze shifting back to the eyes when it is no longer needed at the mouth and again is needed to gain social cues from the eyes. A critical question, then, is whether there is a similar tendency in adults to fixate the mouth when attempting to encode an unfamiliar language. Data from adults gathered by Lewkowicz and Hansen-Tift (2012) suggested that this may not be the case. They found that monolingual English-speaking adults did not fixate on the mouth of a talker speaking in an unfamiliar language (Spanish) more than they fixated on the eyes. Instead, like in their response to a familiar language (English), they fixated primarily on the speaker's eyes. This finding suggests that the observed fixation of the mouth in younger infants may be tied to a specific developmental epoch (e.g. learning to produce speech sounds), rather than representing a general speech encoding processing mechanism. However, the adults in this study were not asked to perform any experimental task other than to watch the videos. This raises the possibility that these adult participants did not engage in speech encoding. Hence, when there was no task to engage in task specific changes in the adult's gaze patterns were not recorded.

Based on the previous findings, Barenholtz, Mavica and Lewkowicz (2016) asked whether they could find a similar behavioral trend, a mouth preference, present in adults, as it is in infants, when the adults actively encoded speech (i.e. completed a task) in an

unfamiliar language. They presented short videos to monolingual and bilingual adults in their native and non-native languages and tracked their eye gaze in task and no-task conditions. They were interested in seeing if there were differences in gaze patterns when encoding their native versus non-native language, if there were differences in bilinguals' gaze patterns to their first and second languages, of which they are equally familiar with, and if the difference in gaze patterns between their native and non-native languages depend on encoding (i.e. completing a task). In the first condition they presented two videos of the same model speaking two different sentences, sequentially, followed by a voice recording of one of the two previous sentences; the participants were asked to report if the voice recording was presented first or second in the previous sequence of two. Each participant was presented with this task in their native and non-native languages. The results reveled that adults encoding speech in an unfamiliar language exhibited gaze patterns that are similar, in some respects, to infants who are first learning their native language, or to adults perceiving speech with noise or uncertainty in the signal—that is, there was enhanced allocation of attention to the mouth when presented with the unfamiliar, non-native language. Next, using the same task, when bilingual speakers were presented with two languages that they were equally familiar with, Spanish and English, there was no significant difference in the time spent looking at the mouth in either language. Finally, they presented the same videos and voice recordings as in the previous conditions. Although here, they asked the participants to passively view the videos with no task and, thus, did not impose any specific processing requirements on them. In this no-task condition, similar to Lewkowicz and Hansen-Tift (2012), there was no difference found in the time spent looking at the mouth versus the eyes when

presented with the familiar versus the unfamiliar languages. Adults are attracted to the eyes and have been shown to fixate the eyes when there are no immediate task demands (Lansing & McConkie, 2003; Lewkowicz and Hansen-Tift, 2012), and it could be the case that this is what was displayed in this final condition while people passively viewed the videos without completing a task. Here, again, we find evidence of task specific gaze patterns present in adults. People seek out information in the location of the speaker's mouth not only under conditions where the auditory stimulus is degraded but also when it is specified in an unfamiliar language.

Overall, we can record a change in the time spent looking at areas of the face that contain information that can be used to complete the task at hand. And, when there are no task demands, the gaze shifts back to the eyes, where the task of gaining social cues can resume. However, during speech perception, the face carries a number of important cues and the choice of which feature to fixate depends on a combination of cognitive and developmental factors (Lewkowicz & Hansen-Tift, 2012), the quality of the underlying stimulus (Vatikiotis-Bateson, Eigsti, Yano & Munhall, 1998), and on the familiarity of the language being spoken (Barenholtz, Mavica & Lewkowicz, 2016). As first demonstrated by Yarbus (1967), cognitive factors interact with the environment to produce goal-directed eye fixations. He recorded the gaze of people while they viewed the same object given different sets of instructions, such as, "give the ages of the people", and, "remember the clothes worn by the people". He found that eye fixations are directed where the most information is provided for the immediate task, and as the task changes so do the gaze patterns. Taken together, these results suggest that fixations are directed where useful information is provided.

The Current Study

Based on previous research (Krauss, Freyberg & Morsella, 2002; Lass & Davis, 1976; Mavica & Barenholtz, 2013; Smith, Dunn, Baguley & Stacey, 2016a; Smith, Dunn, Baguley & Stacey, 2016b), it is likely that people have some implicit knowledge of the mappings between subtle facial features and vocal characteristics which gives them the ability to match a static image of a face and voice to the person they were recorded from. The ability possibly relies on some form of learning, taking place through the course of everyday social interaction, of a set of general mappings between specific facial features and vocal characteristics (Mavica & Barenholtz, 2013). To this point, this has been speculated but never revealed in the data. Based on evidence of a change in the allocation of gaze to areas of the face that contain information that can be used to complete specific tasks (Lansing & McConkie, 1999; Lansing & McConkie, 2003; Lewkowicz & Hansen-Tift, 2012; Barenholtz, Mavica & Lewkowicz, 2016; Yarbus, 1967), we are interested in investigating the informational aspects of the face that the matching ability could possibly rely on and in recording the difference between the gaze patterns when the task is to match a static face to the corresponding voice as opposed to when there are no specific task demands.

In seven experiments the aim was to isolate which facial features or informational aspects of faces account for the ability to complete a static face to voice matching task. In the first six experiments we systematically manipulated the photographs, each in a different manner, in order to remove certain sources of potential information. If, by removing or manipulating a feature in the images, the capacity to do the matching task is

diminished or eliminated, we could begin to put together a depiction of the most important informational aspects involved in the matching ability.

In each of the first six experiments, we simultaneously presented one voice recording along with two manipulated images of faces, a target face (which consisted of a manipulated image of the same individual from whom the voice was recorded), and a distractor face (which consisted of a manipulated image of a face that belonged to one of the other voices used in the experiment). The participant's task was to choose which of the two images displayed they thought belonged to the same individual as the voice recording. The voices remained un-manipulated. One image of each model was manipulated in six different manners. In Experiment 1 we presented participants front view headshots, including the neck and shoulders of the models, cropped from the bridge of the nose down. In Experiment 2 we presented participants with images of the models cropped from the bridge of the nose to the top of the head. In Experiment 3a we presented participants with images cropped from the bridge of the nose down with the lip area cropped out. In Experiment 3b we presented participants with the cropped out lips from Experiment 3a. In Experiment 4 we presented the headshots inverted, rotated 180 degrees. In Experiment 5 we presented low pass filtered images of the headshots. Finally, in Experiment 6 we presented participants with images of the models with a blur covering all information about the inner face (i.e. eyes, nose, and mouth). In all of the experiments we presented three full auditory statements.

In addition to reducing the visual information available in the images, in Experiment 7 we used eye tracking in order to determine which informational aspects of the model's faces participants spent more time looking at while performing the matching

task, as compared to when the participants had no immediate task demands. Where people look on a face seems to be dependent on which parts of the face provide the information necessary to complete the current task (Barenholtz, Mavica & Lewkowicz, 2016; Yarbus, 1967). Based on this, in Experiment 7 we compared total fixation duration while participants were completing a voice to static face matching task, as compared to when they passively listened to the voice recordings and viewed static images of the faces. First, in the matching condition, we presented naïve participants with a voice recording followed sequentially by two static images of the faces (un-manipulated). One was the target face, while the other was a distractor face. The participant's task was to choose which of the two images displayed they thought belonged to the same individual as the voice recording. Next, in the no-task, passive viewing condition we presented another group of naïve participants with a voice recording in one gender followed sequentially by two static images of models of the opposite gender. Participants passively listened to the voice recordings and viewed the videos with no task demands and, thus, we did not impose any specific processing requirements on them. Gender incongruent pairings were presented in the passive viewing condition in order to prevent the participants from automatically trying to do the matching task.

Experiment 1: Bottom Half of Face

Attention to the face, particularly the mouth, during speech perception is important. The mouth region contains important information related to the temporal and acoustic characteristics of speech and gaze is often directed toward the mouth when people are viewing speaking faces (Argyle & Cook, 1976; Lansing & McConkie, 1999; Lansing & McConkie, 2003). It is possible that participants have some implicit knowledge of the mappings between subtle facial and vocal characteristics which gives them the ability to match a voice and face to the person they were recorded from. The ability could rely on a form of implicit learning, taking place through the course of everyday social interaction (Mavica & Barenholtz, 2013). If this were the case, then the lower half of the face could be important because people often gaze toward the mouth while they are listening to the voice, which is when they would create the mappings between the subtle face and voice characteristics. Therefore, in Experiment 1 we isolated the lower half of the face, cropping a frontal headshot from the tip of the nose down to the top of the shoulders.

Method

Participants

Participants were 22 English-speaking undergraduate students, enrolled in an introductory psychology course at Florida Atlantic University (FAU), who received partial credit for participation. All participants reported having normal or corrected-to-

normal vision and hearing, and gave informed consent according to the guidelines set forth by (FAU).

Stimuli

A Mac desktop computer with a 17-inch monitor was used to present the visual and audio stimuli to the participants using Matlab. The visual stimuli presented to participants were manipulated frontal headshots. The face images consisted of the top half of the models faces cropped from the bridge of the nose to the top of the head (See Figure 1A), and the auditory stimuli presented were voice recordings of three full spoken statements: (a) There are clouds in the sky, (b) The boy took his sister to the park, and (c) One, two, three, four, five. All of the images and the voice recordings were drawn from Caucasian undergraduate students (mean age: 21) from FAU. There were 32 males and 32 females, each supplying one headshot and three full statements. We were specifically interested in the ability to use vocal characteristics rather than dynamic properties, such as the rate of speech or intonation. Therefore, in order to control for speed and varying articulation patterns, the models listened, through headphones, to pre-recorded statements, played on a loop. The models were instructed to speak along with the statements they heard through the headphones.

Procedure

In each trial, participants were presented with images of two manipulated faces of the same gender, while simultaneously listening to a voice recording of one of the models pictured in the images; participants had to choose which of the two faces they thought belonged to the same individual as the voice recording by responding on a standard keyboard (See Figure 2). Block one always presented the statement "There are clouds in

the sky;" block two presented the statement "The boy took his sister to the park;" and block three presented the statement "One, two, three, four, five." Each face was shown twice per block, once with the true voice and once as a distractor face; each voice was presented only once per block. Participants were tested on all of the faces and voices in each of the three separate blocks. The blocks were further divided into a male group and a female group. All the female models were displayed first followed by the males, or vice versa, counterbalanced across participants. The faces and voices were randomly displayed across trials and across participants. The face pairs were also randomly displayed and only by chance would two faces be paired with each other more than once per experiment.

Results

Across all six experiments, there were no significant differences in performance across the three experimental blocks (i.e. for the three different statements); therefore, in subsequent analyses the data from the three blocks were collapsed into a single measure of performance. Individual t-tests found better than chance (50%) performance in Experiment 1 ($M = 0.5309$, $SE = 0.0094$, $t(21) = 3.29$, $p = .003$) (See Figure 3). Additional t-tests comparing performance of male and female participants found no significant differences (females: $M = .5473$, $SD = .0319$, males: $M = .5145$, $SD = .0496$, $p = .05$), and comparing performance on the male and female models revealed no significant differences (females: $M = .5370$, $SD = .4990$, males: $M = .5399$, $SD = .5008$, $p = .05$). Across all six experiments, there were no significant differences in performance between male and female participants or between the performance on the male of female

models; therefore, in subsequent analyses the data were collapsed into a single measure of performance.

Discussion

The results from Experiment 1 indicate that people are able to choose, based on a static image of the bottom half of the face, what voice was recorded from each model. This suggests that the bottom half of the face provides information that allows for the ability to make the correct match. This information does not have to be exclusive to the bottom half of the face, however. There are properties of the face that one may be able to concordantly extract from both the top and bottom half of the face. If the ability relies on information only contained in the bottom half of the face, then presenting the top half of the face should diminish the matching ability. Moreover, when viewing a speaking face, the mouth is not the only point of interest. The eyes are an important source of information and again, are a place people gaze at while they are listening to speech sounds. Based on this, in Experiment 2, using the same testing paradigm and same three full, complete statements as in Experiment 1, we presented the top half of the face to the participants.

Experiment 2: Top Half of Face

In general, adults often fixate on the eyes (Argyle & Cook, 1976), which are thought to attract interest because they carry important social and emotional cues (Langton, Watt & Bruce, 2000). Eye contact is used to regulate turn taking in conversation, and express emotional state and intimacy (Kleinke, 1986). While we did find in Experiment 1 that the matching task can be accomplished with only the bottom half of the face, in Experiment 2 we presented the top half of the face in order to narrow down what informational aspects the matching ability relies on. If the ability relies on information that is global, such as shape and size of the face, then the matching ability should persist when presented with only the top half of the face. On the other hand, if the matching ability relies on specific features that are located only on the bottom half of the face the ability should diminish when presented with only the top half of the face.

Method

Participants

Participants were 33 English-speaking undergraduate students, enrolled in an introductory psychology course at FAU, who received partial credit for participation. None of the students participated in any of the other experiments, and all were naïve to the purpose of the experiment. All participants reported having normal or corrected-to-normal vision and hearing, and gave informed consent according to the guidelines set forth by FAU.

Stimuli

The face stimuli for Experiment 2 consisted of the top half of the models' faces. The faces were cropped from the tip of the nose up to the top of the head, hair included (See Figure 1B).

Procedure

The procedure for Experiment 2 was the same as for Experiment 1.

Results and Discussion

In Experiment 2, individual t-tests found better than chance (50%) performance ($M = 0.5196$, $SE = 0.0065$, $t(32) = 2.70$, $p = .011$) (See Figure 3). Experiments 1 and 2 indicate that people are able to choose, based on a static image of the bottom half of the face, and a static image of the top half of the face, what voice was recorded from each model. This suggests that the ability either relies on information that can be extracted concordantly from both halves of the face, such as the global size and shape of the face, or it could be the case that different informational aspects of the face work separately to contribute to the matching ability. When one feature is not available the ability will rely more heavily on other features. Following Experiment 1, where we presented the bottom half of the face, in Experiment 3a we presented images of the bottom half of the face with the lips cropped out.

Experiment 3a: Bottom Half of Face – Lips Cropped Out

Soon after birth, infants have been found to discriminate vowel sounds (Trehub, 1973). Kuhl and Meltzoff (1996) reported that 4.5-month-olds imitated the vowel that matched the vowel they both saw and heard. And, later, Legerstee (1990) presented audio and visual displays of vowels to 3-month-olds; infants who were presented with the matching audiovisual displays imitated the vowel sounds more than infants who were exposed to mismatched displays. We can see from an early age that the mouth, the lip area in particular, conveys important information about speech, and is fixated often during speech encoding (Lewkowicz & Hansen-Tift, 2012). And, it is beneficial for infants and adults to fixate on the lip area while listening to speech sounds that are degraded (Vatikiotis-Bateson, Eigsti, Yano & Munhall, 1998) or unfamiliar (Barenholtz, Mavica & Lewkowicz, 2016). In Experiment 3a. we presented the bottom half of the face with the lip area cropped out, while still presenting the same three full, complete statements as in Experiment 1.

Method

Participants

Participants were 25 English-speaking undergraduate students, enrolled in an introductory psychology course at FAU, who received partial credit for participation. None of the students participated in any of the other experiments, and all were naïve to the purpose of the experiment. All participants reported having normal or corrected-to-

normal vision and hearing, and gave informed consent according to the guidelines set forth by FAU.

Stimuli

The stimuli for Experiment 3a were the same as for Experiment 1, except the facial stimuli consisted of the bottom half of the models' faces with the lips of the model cropped out (See Figure 1C).

Procedure

The procedure for Experiment 3a was the same as for Experiment 1.

Results and Discussion

Individual t-tests did not reach statistical significance ($M = 0.5064$, $SE = 0.0075$, $t(24) = .85$, $p = .402$) (See Results and Discussion for Experiment 3b.)

Experiment 3b: Cropped Out Lips Only

Here, we presented a group of participants with the cropped out lips from the stimuli presented in Experiment 3a. as a control.

Method

Participants

Participants were 34 English-speaking undergraduate students, enrolled in an introductory psychology course at FAU, who received partial credit for participation. None of the students participated in any of the other experiments, and all were naïve to the purpose of the experiment. All participants reported having normal or corrected-to-normal vision and hearing, and gave informed consent according to the guidelines set forth by FAU.

Stimuli

The stimuli for Experiment 3b were the cropped out lips from Experiment 3a (See Figure 1D).

Procedure

The procedure for Experiment 3b was the same as for Experiment 1.

Results and Discussion

Individual t-test did not reach statistical significance ($M = 0.4896$, $SE = 0.0054$, $t(33) = -3.57$, $p = .091$) (See Figure 3). Participants were unable to correctly make the match in Experiment 3a when they were presented with bottom half of the face with the lip area cropped nor were they able to correctly make the match in 3b when they were

presented with the cropped out lip area from the previous experiment. Following Experiment 1, in which participants correctly matched images of the bottom half of the faces to the corresponding voice recordings at better than chance levels, in Experiment 3a and 3b we continued to consider not only the bottom half of the face, but the importance of the lip area in the matching ability. In Experiment 1 we found the participants were able to complete the matching task when presented with only the bottom half of the face, when we then removed the lips in Experiment 3a the ability completely diminishes. This could suggest evidence that the mouth is a particularly important informational aspect of the face involved in the matching ability. Although, when presented with just the mouth the ability diminishes, and there is not a significant difference between Experiments 3a and 3b. It could also be the case that the removal of the lips in this manner, caused participants to actually focus more on the mouth and ignore other features of the face that provide information to complete the task. For example, the participants could have spent their time visualizing what the lips could possibly look like and therefore, spent more times fixating in the lip area rather than inspecting the face globally. We could also speculate that with the removal of the lips the faces were perceived differently. Participants could have been taken by the oddity of the images and been unable to process the face in an organized manner.

Experiment 4: Inverted Whole Faces

Faces are analyzed in a configural manner, meaning that when you attend to a face you process not only the features (e.g. eyes, nose, and mouth), but also the spatial relationships amongst the features (Leder, & Carbon, 2006). The face inversion effect (Yin, 1969), in which an inverted face is much harder to recognize than an upright face, is due to a decrease in the ability to process the spatial relationships of the face (Rossion & Gauthier, 2002). When viewing inverted images of faces we must rely on featural processing. If presenting inverted images in the matching task interferes with performance, we can infer that relational, not featural, information is involved in the matching ability. In order to investigate whether the matching ability is related to relational information, or featural information about faces, we presented the face images inverted, rotated 180 degrees, and again used the same procedure and voice recordings as outlined above.

Method

Participants

Participants were 33 English-speaking undergraduate students, enrolled in an introductory psychology course at FAU, who received partial credit for participation. None of the students participated in any of the other experiments, and all were naïve to the purpose of the experiment. All participants reported having normal or corrected-to-normal vision and hearing, and gave informed consent according to the guidelines set forth by FAU.

Stimuli

The stimuli for Experiment 4 were whole images of the headshots except the images were presented inverted, rotated 180 degrees (See Figure 1E).

Procedure

The procedure for Experiment 4 was the same as for Experiment 1.

Results and Discussion

For Experiment 4, individual t-tests found better than chance (50%) performance ($M = 0.5212$, $SE = 0.0074$, $t(32) = 2.87$, $p = .007$) (See Figure 3). We presented inverted whole faces in which the facial features were still prominent but the relational information between the features was degraded. The results revealed better than chance performance; the ability to correctly make the match persisted. We can now assume that the matching ability does not rely on relational information about the inner face features, such as the distance between the eyes, nose, and mouth.

Experiment 5: Low Pass Filter

Low pass filtered images do not allow accurate location of closely bordering edges. The measurement of small-scale distances, such as distance between the eyes or the height of the eyebrows, are dependent on high frequencies, while large-scale distances, such as face shape, are dependent on low frequencies (Costen, Parker, & Craw, 1996). In low pass filtered images low frequencies are eliminated and, therefore, so too is small-scale information. Experiment 5 investigates very similar concepts to Experiment 4, there we decreased the ability to process spatial relationships by inverting the face images. Here, we disturb the processing of spatial relationships by using a low pass filter. If our interpretation is correct and extracting the spatial relationships of faces by inversion does not diminish the matching ability, then when we should see the same result when we present the low pass filtered images. On the other hand, if the matching ability is diminished, we can then suggest that small-scale information may be essential to the matching task.

Method

Participants

Participants were 35 English-speaking undergraduate students, enrolled in an introductory psychology course at FAU, who received partial credit for participation. None of the students participated in any of the other experiments, and all were naïve to the purpose of the experiment. All participants reported having normal or corrected-to-

normal vision and hearing, and gave informed consent according to the guidelines set forth by FAU.

Stimuli

The stimuli for Experiment 5 were whole images of the headshots that were degraded with a low pass filter that, using Matlab, smoothed the images and removed high frequency noise from them (See Figure 1F).

Procedure

The procedure for Experiment 5 was the same as for Experiment 1.

Results and Discussion

Individual t-tests found better than chance (50%) performance ($M = 0.5269$, $SE = 0.0072$, $t(34) = 2.33$, $p = .03$) (See Figure 3). The ability persisted and results suggest that the matching ability may rely more on global features, such as face size and shape, rather than on local features, such as distance between the eyes. When presented participants with the low pass filtered images performance was found to be better than chance. Removing the small-scale information did not hamper the ability to complete the task. Here again, as in Experiment 4, we found that the local features of the face are not necessary to the matching ability.

Experiment 6: Removal of all Facial Features

Findings from Experiments 1 and 2 suggest that information concordantly available in both the top and the bottom halves of faces can account for the matching ability. Furthermore, findings from Experiments 4 and 5 suggest that the local features of the face are not necessary to correctly complete matching task. Combined, these results point to the global face shape and size somehow correlating with vocal characteristics. Additionally, Venter, et al. (2001) and his team, working with machine learning, have developed the ability to predict a participant's gender, age, and height from voice recordings and found a way to isolate a set of genes that determine appearance. First, they recorded a few minutes of speech from each of 1,000 participants. They took precise body measurements with an MRI, including throat and jaw size, and matched and correlated the information with a machine that could learn and make predictions. They also collected DNA samples from thousands of participants which were then matched and correlated with 3D models of the participants' faces. Amongst other findings, a correlation between voice and face shape was found (McFarland, 2015). Following from this, in Experiment 6, in order to investigate whether the informational aspects involved in the matching ability are related to the global face shape and size, we presented images with a blur that concealed all the inner facial features such as the eyes, nose, and mouth while still preserving the global facial characteristics.

Method

Participants

Participants were 30 English-speaking undergraduate students, enrolled in an introductory psychology course at FAU, who received partial credit for participation. None of the students participated in any of the other experiments, and all were naïve to the purpose of the experiment. All participants reported having normal or corrected-to-normal vision and hearing, and gave informed consent according to the guidelines set forth by FAU.

Stimuli

The stimuli for Experiment 6 were whole images of the headshots except that, using Photoshop, the facial stimuli had the inner face of the models blurred out of recognition, including the eyes, nose, and mouth of the models (See Figure 1G).

Procedure

The procedure for Experiment 6 was the same as for Experiment 1.

Results and Discussion

Individual t-tests found better than chance (50%) performance in Experiment 6 ($M = 0.5247$, $SE = 0.0087$, $t(29) = 2.83$, $p = .008$) (See Figure 3). The results revealed that the matching task can still be completed with better than chance performance, even with a complete lack of visual information about the eyes, nose, and mouth. This indicates that information gained from the inner facial features are not necessary at all to complete the matching task and that the global features may be what the matching ability relies on.

Discussion

Overall, the results taken from Experiments 1-6 indicate that people are able to choose, based on a static image of a face, what voice was recorded from each individual when relying on varying degrees of visual information. An ANOVA comparing performance in Experiments 1-6 found a significant difference in performance among the experiments [$F(6, 205) = 5.18$, $p = .001$]. Gabriel's post hoc analysis revealed that performance in Experiment 3a (bottom half- cropped out lips) and 3b (the cropped out lips) did not significantly differ from each other. However, all of the remaining experiments (i.e. Experiments 1, 2, 4, 5, 6) have performance that is significantly better than the performance in Experiment 3a and 3b, and have performance that does not significantly differ from each other. In other words, of all of the experiments with performance better than chance, none of them had performance that was statistically different from each other. Therefore, none of the manipulations, besides 3a and 3b, seemed to affect the matching ability more or less than others.

The images and voice recordings were identical to Mavica & Barenholtz (2013), therefore, we could compare the performance on the individual models in the respective experiments. Across some of the experiments, performance of the models was positively correlated with the performance from Mavica and Barenholtz (2013). Also found were correlations between some of the experiments within the current study, [Experiment 4 and Mavica & Barenholtz (2013): $r(46) = .433$, $p < .003$; Experiment 5 and Mavica & Barenholtz (2013): $r(46) = .391$, $p < .007$; Experiment 4 and Experiment 5: $r(47) = .611$, $p < .001$; Experiment 1 and Experiment 4: $r(47) = .403$, $p < .005$; Experiment 1 and Experiment 5: $r(46) = .457$, $p < .001$] (See Figure 4). Thus, there was a degree of consistency with regard to how people performed on the matching task with each

individual model's images across not only experiments within the current study but also with Mavica and Barenholtz's (2013) study. It seems to follow that participant's performance on each of the models' in Experiments 4 and 5 would most closely correlate with models' performance in Mavica and Barenholtz (2013). These two conditions displayed full frontal headshots and the images contained information that was most similar to the former study.

While overall performance in the matching task was significantly better than chance, except for 3a and 3b, it was still quite poor, with the best performance across all of the experiments only reaching about 53% correct. It should be noted, however, that this level of performance is not far below performance reported in previous studies using static stimuli, which ranged between 55-60% correct. This overall performance measure reflects variability in performance across the models, with some model's consistently yielding much better performance than the rest and a smaller number yielding considerably worse performance. The results revealed that the majority of models (~60%) produced better than chance performance. Bonferroni adjusted alpha levels of .0008 (.05/64) per test were used to correct for multiple comparisons. A total of 7 models yielded greater than chance performance, while 5 models yielded worse than chance performance. Next, we considered whether any of the models yielded performance that was significantly better or worse than the average performance across all of the models in the significant experiments (52%). The 5 top models yielded performance that was significantly better than this average while the 3 bottom models yielded performance below this average (See Figure 5). In addition, there is a high degree of consistency, both good and bad, across the experiments in this study, as well as compared to Mavica and

Barenholtz's (2013) experiments. This is evident from the high degree of correlation in performance for the different models across the experiments. There are a proportion of models who were consistently matched correctly and another proportion who were consistently matched incorrectly. This suggests that participants across all the studies (Mavica & Barenholtz (2013) included) maintained shared expectations about which faces and voices matched; what varied was the degree to which the different models conformed to those expectations. In the case of the top-performing models, their facial and vocal properties were consistent with participant's expectations, in the case of the bottom–performing models, they were inconsistent. In both cases, the relationship between participant's expectations and actual face-voice pairings were systematic and do not reflect random guessing.

Experiment 7: Compare gaze patterns in the matching task versus passive viewing

Previous results suggest that eye fixations are directed where useful information is provided (Barenholtz, Mavica & Lewkowicz, 2016; Buchan, Pare, & Munhall, 2007; Lewkowicz & Hansen-Tift, 2012; Yarbus, 1967). Based on this, we asked whether we can use eye tracking in order to see where people are fixating and, therefore, where people are getting their information from while they are completing the static face to voice matching task. The global shape and size of a face may contain information that people extract while trying to match a face to the corresponding voice (Venter, et al., 2001). We found evidence of this in Experiments 4, 5, and 6; when spatial information was disturbed or smoothed in the images and even when there was complete removal of information about the inner facial features the matching ability persisted. Therefore, we predicted that we would find significantly more fixations directed toward the outer areas of the face, such as the hairline or jawline, than at the inner face, such as the eyes, nose, and mouth, when people are trying to complete the matching task, as compared to a no-task, passive viewing condition in which the participants had no specific task demands. On the other hand, it has been found that when tasks require speech processing the mouth becomes a significant source of fixations (Lansing & McConkie, 1999; Lansing & McConkie, 2003; Vo, Smith, Mital, & Henderson, 2012) and it has also been found that when viewing speaking faces, without task demands, the eyes are a considerable focus of

fixation (Barenholtz, Mavica & Lewkowicz, 2016; Lewkowicz & Hansen-Tift, 2012). Therefore, in the matching task it is a possibility that we could find significantly more fixations directed toward the mouth than the eyes because the task concerns speech signals however, the task is not a speech processing task and therefore, it is also a possibility that we will see more fixations directed toward the eyes.

Using an eye tracking system, in order to compare the gaze behavior of participants while they engaged in the matching task versus the gaze behavior while participants had no immediate task demands, in Experiment 7 we replicated Mavica and Barenholtz's (2013) Experiment 2 sequential matching task, along with a similar no-task, passive viewing condition. In the matching condition, in each trial, we presented participants with a voice recording followed sequentially by two static images of models, one of the target model and one of a distractor model. The participants' task was to choose which model matched the voice recording. In the no-task, passive viewing condition, in each trial, we presented a voice recording of a male model followed sequentially by two static images of female models, or vice versa, counterbalanced across participants.

In order to determine total fixation duration, we defined four principal areas of interest (AOI's). The eye AOI was defined by a rectangle drawn just above the eyebrows and through the bridge of the nose with the vertical sides drawn to the model's hairline on each side of the face. The mouth AOI was defined by a rectangle drawn between the upper lip and bottom of the nose and the lower part of the chin with the vertical sides drawn just outside the corners of the mouth on each side. The inner face AOI was defined by a rectangle from the top of the eyebrows to the bottom of eyes and from the

outermost corner eye to eye; it was connected to a square covering the nose and a rectangle covering the mouth. This AOI contained the major facial features, such as the eyes, nose, and mouth of the face. The outer face AOI was first defined by an oval drawn from the top of the head (i.e. hair included) to the bottom of the chin and from side to side including the ears and then we subtracted the total fixation duration of the inner face AOI from the whole image. This gives us only the fixation duration of outer area of the face such as the jawline and hairline, and the overall global shape and size of the face (See Figure 6).

Method

Participants

Participants were 60 English-speaking undergraduate students (28 in the matching condition and 32 in the passive viewing condition), enrolled in a psychology course at FAU, who received extra credit for participation. None of the students participated in Experiment 1-6 and all were naïve to the purpose of the experiment. All participants reported having normal or corrected-to-normal vision and hearing, and gave informed consent according to the guidelines set forth by FAU.

Stimuli

Participants' eye movements were recorded with an eye tracking system (T120; Tobii Technology, Stockholm, Sweden) and analyzed with the Tobii Studio 3.0.6 software. Gaze was monitored using near infrared and both bright and dark pupil-centered corneal reflection. Stimuli were presented on a 17-inch flat panel monitor with a screen resolution of 1280 X 1024 pixels. The eye tracker allows head movement within a 30 x 22 x 30 cm volume when seated 50 to 80 cm in front of the screen; this provides

fairly natural conditions for assessing gaze. The sampling rate is 120 Hz; for each eye 120-gaze data points are collected per second. All participants were tested in a quiet room that was illuminated by the stimulus display and were seated ~60 cm from the screen. A standardized five-point calibration was performed prior to tracking.

Visual stimuli were the same stimuli presented in Mavica and Barenholtz (2013) (e.g. frontal headshots); the stimuli were not manipulated. The audio stimuli were the same stimuli as presented in Block 1 of Experiments 1-6 (i.e. "There are clouds in the sky").

Procedure

In the matching condition, a sample trial was presented to the participants by the experimenter with detailed verbal and written instructions on how to complete the task. Participants were asked if they understood the task and upon agreement the experimenter left the room. There was no time limit and eye movements were tracked throughout the session. In each trial, the participants first heard a voice recording and then were presented with two static face images displayed sequentially one after another. The faces were displayed for 3 seconds with an inter-stimulus interval of 1 second. Participants chose which face, the first or second, they thought matched the voice recording (See Figure 7). Each face was shown twice, once with their true voice and once as a distractor face; each voice was presented only once. Participants were tested on 32 male and 32 female faces and voices. The faces and voices were pseudo-randomly displayed across trials.

In the no-task, passive viewing condition, participants were simply told to watch and listen to the stimuli that would be presented on the computer screen and to maintain

fixation at the screen for the entire duration of the experiment. Each trial in block 1, the participants were presented with a voice recording of a male model followed sequentially by two static face images of female models (un-manipulated). In block 2 participants were first presented with a voice recording of a female model and then with two static images of male models, or vice versa, counterbalanced across participants. The gender incongruent stimuli were presented in order to prevent the participants from automatically trying to do a matching task.

Results

The study used a between-subjects design with two conditions. We first calculated the proportion of total fixation duration that each participant spent looking at the whole image divided by the amount of time spent looking at each AOI (eyes, mouth, inner face, and outer face) for each of the 64 models presented in each of the two conditions. As mentioned, we were interested in accessing if there was a difference in the amount of time spent fixating the inner face versus the outer face and if there was a difference in the amount of time spent fixating the mouth versus the eyes while engaging in the matching task, as compared to the no-task, passive viewing condition where participants had no specific task demands. For each participant we calculated a mouth-vs.-eyes difference score and an inner face-vs.-outer face difference score. Analysis of the mouth-vs.-eyes difference scores revealed no significant differences between the matching and passive viewing condition. However, analysis of the inner face-vs.-outer face difference scores, specifically, a between subjects t-test, revealed that the difference scores were greater in the matching condition, $M = .57$, $SD = .27$ than in the passive viewing condition $M = .40$, $SD = .31$, $t(131) = -2.186$, $p < .031$ (See Figure 8). Thus, participants looked more at the

outer face in the matching condition than in the passive viewing condition. We were also interested in seeing if there was a difference in the amount of time fixating the whole face AOI in the matching condition as compared to the passive viewing condition. A between subjects t-test revealed that the time spent fixating the whole face was greater in the matching condition, $M = 79.25$, $SD = 10.77$ than in the passive viewing condition $M = 66.36$, $SD = 15.75$, $t(131) = -5.567$, $p < .001$. Thus, participants spent more time fixating at the whole face when they were completing the matching task. Using individual t-tests, the results for the matching task revealed significantly worse than chance (50%) performance ($M = 0.4519$, $SE = 0.0120$, $t(27) = -4.01$, $p = .001$).

Discussion

Overall, we found that when people were engaged in the matching task, using static images, the time spent looking at the outer area of the face, such as the jaw line, increased as compared to when people passively viewed the images. The results also revealed that the time spent overall looking at the whole face in the matching condition was significantly greater than the time spent in the passive viewing condition. These results, taken together, reveal that participants engaged in the matching task spent more time inspecting the outer edges of the faces and more time inspecting the faces in general. These results are not surprising however; it makes sense that the participants would spend more time fixating a face when trying to make a decision about it as opposed to when they were passively viewing it. And, based on the results from Experiments 1-6 we know that the inner face features are not necessary in the matching ability. Therefore, increased fixations to the outer face point to the participants taking global shape and size of the face into consideration when making their decision.

Results from the sequential matching task revealed worse than chance performance. This is not surprising, however, because of the heavy demands put on working memory. Each trial contained three stimuli (one voice recording, and two static images), each of which were displayed for about three seconds, each separated by one second. This heavy demand on working memory likely hampered the matching ability (Barenholtz, Mavica, & Lewkowicz, 2016). In the current experiment it was not imperative that the participants be able to perform better than chance in the matching task. We were simply interested in recording the total fixation duration while the participants actively completed the matching task as compared to when they were doing something else (e.g. passive viewing). We chose to use the match to sample testing paradigm and to display each image at such length in order to maximize the information that we could analyze from the eye tracker and minimize fixations not involved in the matching task. If we presented the faces for a longer duration, we were concerned that the participants would make their decision on whose face the voice was recorded from within three seconds and then continue to fixate areas of the face that are not involved in the decision making process. In that case, we would not be able to decipher which fixations were task dependent. In sum, we needed each image to be displayed for an ample amount of time, knowing that it would likely hamper the matching performance but not too long as to record fixations that were not task dependent.

General Discussion

Overall, we found that participants were able to match an unfamiliar voice to a static image of the face of the person from whom the voice was recorded at significantly better than chance levels in most of the experiments. Performance was better than chance and equal (i.e. not significantly different from each other) when shown the bottom half of the face (Experiment 1), the top half of the face (Experiment 2), when the images were inverted whole faces (Experiment 4), when the images were low pass filtered (Experiment 5), and with removal of all facial features (Experiment 6) but, performance was significantly worse when shown the bottom half of the face with the lip area cropped out (Experiment 3a) and when presented with the cropped out lip area only (Experiment 3b).

The results of Experiments 1 and 2 initially suggested that features of the face that can be concordantly extracted from the bottom or top half of the face may be responsible for the matching ability. With evidence that there is a correlation between face shape and vocal characteristics (Venter, et al., 2001) we investigated the importance of the inner features of the face in the matching task. In Experiment 4 we presented the images inverted 180 degrees, (which decreases the ability to process the spatial relationships of the face) and in Experiment 5 we presented low pass images (which will not allow for accurate location of closely bordering edges). In both cases, we found better than chance

performance. Next, we blurred the interior facial features, such as the eyes, nose, and mouth, while preserving the global face size and shape. Here, with no information about inner features at all, the ability persisted. We concluded that the facial features, such as the eyes, nose, and mouth, are not necessary components that allow for the matching ability. This is not a suggestion that they are not considered when completing the task, this is an indication that they are not necessary in order to complete the task.

The results of Experiment 7 support the indication, obtained from the results of Experiments 1-6, that the global features of the faces, such as face shape and size, may be important information to seek while engaging in the matching task. We found fixations toward the outer edges of the face during the matching task increased as compared to the no-task, passive viewing condition. In the passive viewing condition, we found increased fixations in the inner face AOI, which contains the eyes, nose, and mouth. Because the stimuli were gender mismatched we assumed participants would not engage in a specific task and would, therefore, gaze more toward the eyes. However, the results did not show this, possibly because the stimuli were presented in a gender incongruent order. This gender incongruence could have caused fixations at features of the face that provide information about the gender, or even the masculinity and femininity of the faces. This would cause the gaze to shift away from the eyes and toward other areas of the face, such as the hairline, the cheek bones and/or the jaw line. Therefore, the lack of increased allocation of attention to the eyes AOIs in the no-task, passive viewing condition could have been muddied by the gender incongruent presentation order.

In conclusion, based on all seven experiments, it seems that the inner facial features of the faces (e.g. the eyes, nose, and mouth) do not play as important of a role in

the face and voice matching task as do the outer facial features, such as the jawline and hairline. Global features of the faces shape and size seem more to contribute to the ability to do the matching task. Without access to information about the inner features the task could still be completed at better than chance levels. Therefore, the ability likely relies on global features that can be extracted without information about the inner facial features.

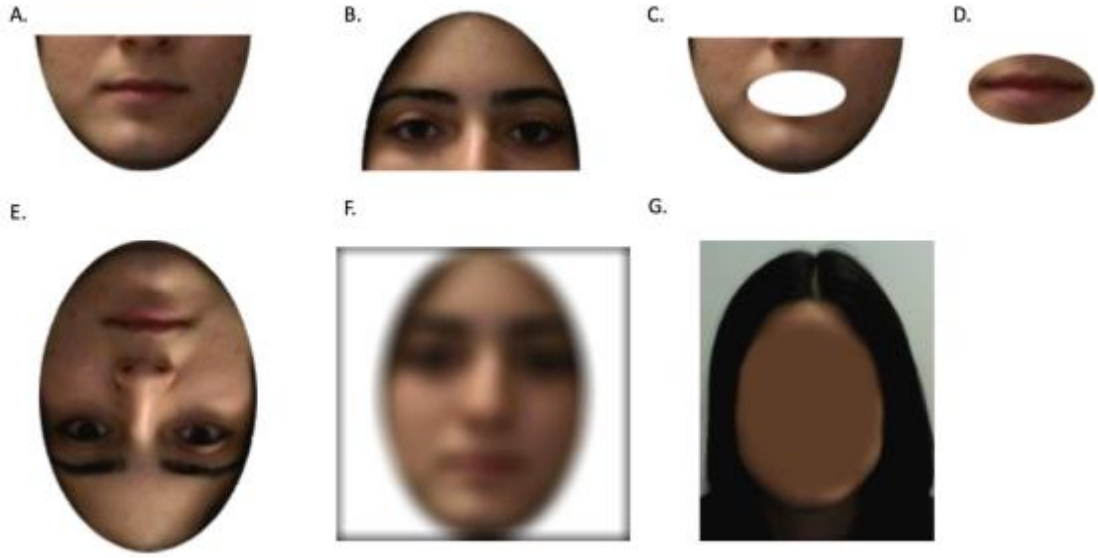

Figure 1. Example stimuli for Experiment 1 (A), Experiment 2 (B), Experiment 3a (C), Experiment 3b (D), Experiment 4 (E), Experiment 5 (F), and Experiment 6 (G).

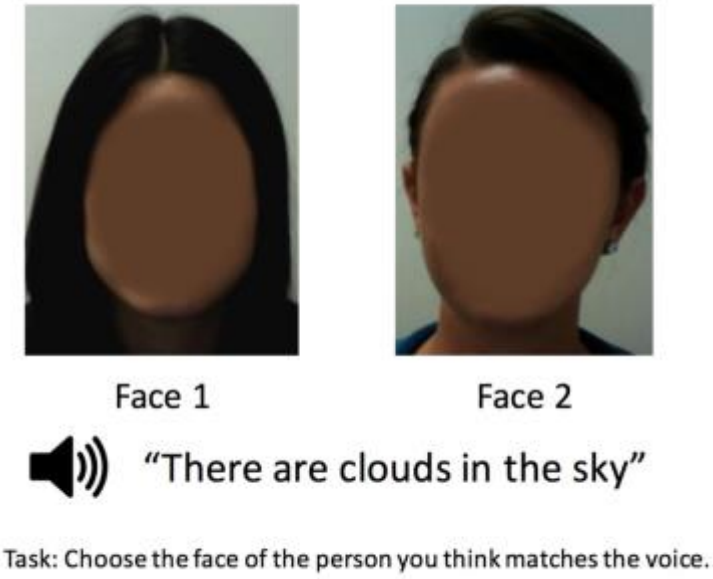

Figure 2. Example of a single trial in the matching task (Experiment 6 stimuli pictured). In each trial participants were simultaneously presented with two images of faces and a recording of a voice. Their task was to choose which face belonged to the same person as the voice.

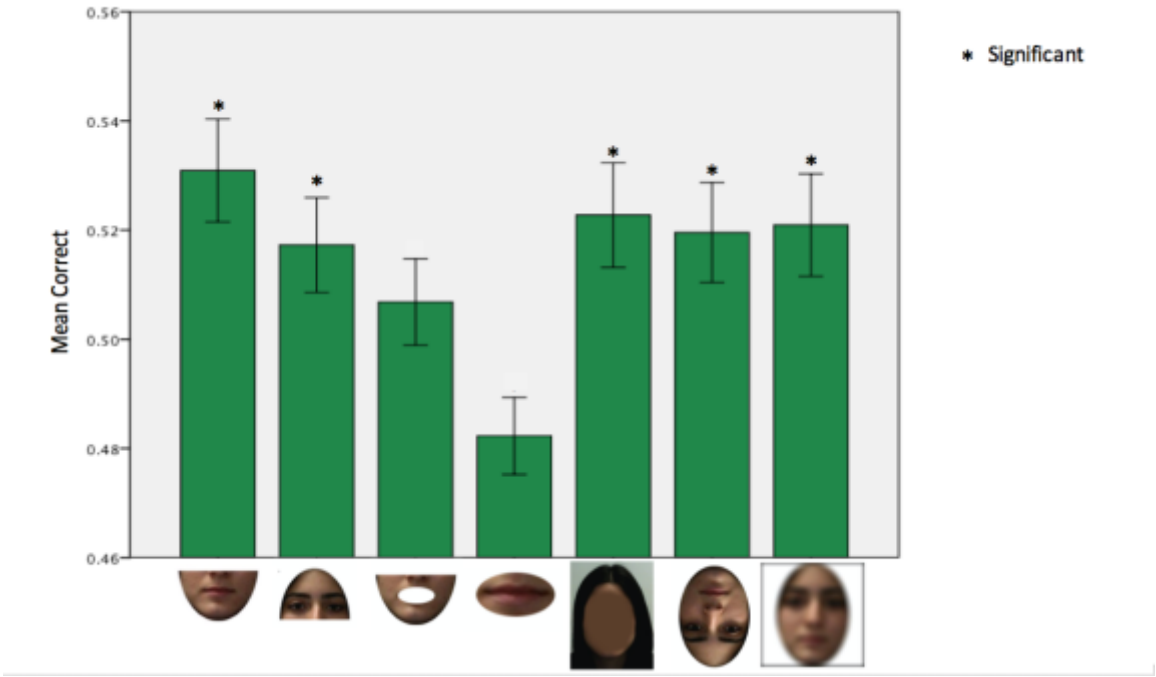

Figure 3. Average performance in each of the Experiments. Error bars indicate +/- 1 standard error of the mean.

Correlations

		EXP1	EXP2	EXP4	EXP5	EXP6	EXP2013
EXP1	Pearson Correlation	1	.209	.018	.403**	.457**	.207
	Sig. (2-tailed)		.097	.906	.005	.001	.110
	N	64	64	46	47	46	61
EXP2	Pearson Correlation	.209	1	.221	.130	.135	.098
	Sig. (2-tailed)	.097		.135	.380	.365	.451
	N	64	65	47	48	47	61
EXP4	Pearson Correlation	.018	.221	1	.028	.087	.269
	Sig. (2-tailed)	.906	.135		.850	.564	.074
	N	46	47	47	47	46	45
EXP5	Pearson Correlation	.403**	.130	.028	1	.611**	.433**
	Sig. (2-tailed)	.005	.380	.850		.000	.003
	N	47	48	47	48	47	46
EXP6	Pearson Correlation	.457**	.135	.087	.611**	1	.391**
	Sig. (2-tailed)	.001	.365	.564	.000		.007
	N	46	47	46	47	47	46
EXP2013	Pearson Correlation	.207	.098	.269	.433**	.391**	1
	Sig. (2-tailed)	.110	.451	.074	.003	.007	
	N	61	61	45	46	46	61

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 4. A correlations table of the participant's performance on each of the models. EXP2013 refers to Mavica and Barenholtz (2013).

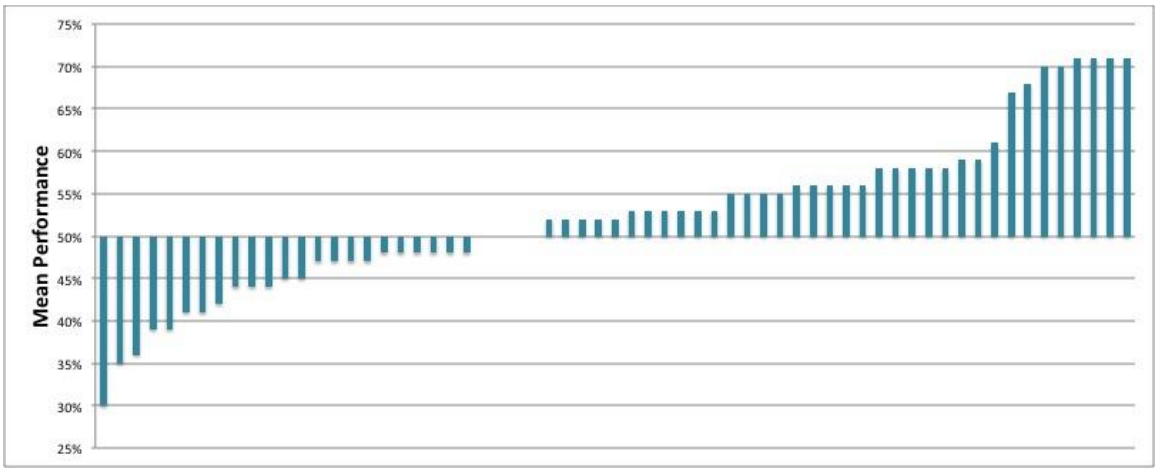

Figure 5. Mean performance in the matching task for each of the models. Each bar represents one of the 64 models, ordered from left to right based on performance.

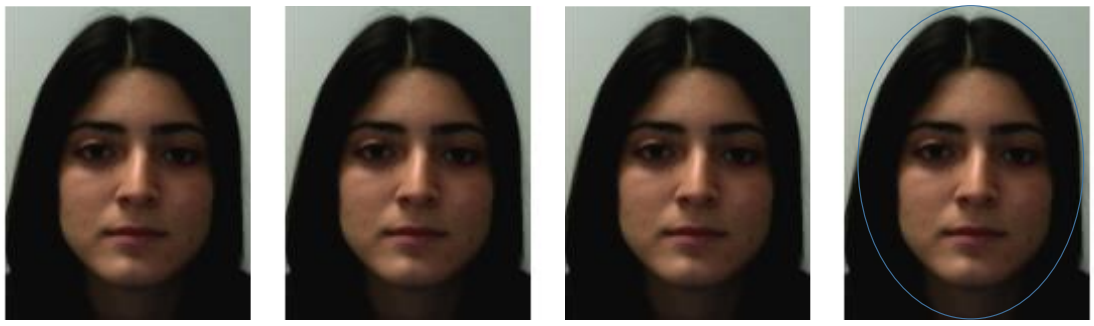

Figure 6. Areas of interest (AOI) are shown as translucent gray shapes on the face images for illustration; they did not appear in the experimental stimuli. From left to right are the eyes, mouth, inner face and whole face. The outer face AOI was calculated by subtracting the data of the inner face AOI from the whole face AOI.

Figure 7. Example stimuli for a single trial in Experiment 7. In the matching condition, in each trial, participants were presented with a voice recording in one gender (e.g. female) followed sequentially by two images of female faces. Participant's task was to choose which face belonged to the same person as the voice recording. In the passive viewing condition, participants were presented with a voice recording in one gender (i.e. male) followed sequentially by two pictures of female faces. There was no task to complete in Condition 2; participants passively viewed the stimuli.

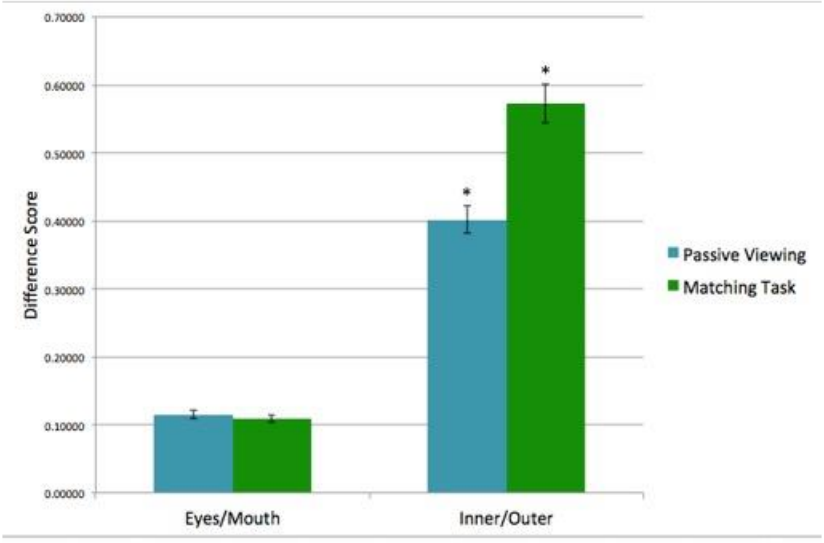

Figure 8. For each participant we calculated a mouth-vs.-eyes difference score and an inner face-vs.-outer face difference score. Analysis of the mouth-vs.-eyes difference scores revealed no significant differences between the matching and passive viewing condition. However, analysis of the inner face-vs.-outer face difference scores revealed that the difference scores were greater in the matching condition than in the passive viewing condition. Error bars indicate + /- 1 standard error of the mean.

References

Atkinson, J., Braddick, O. (1989). Development of basic visual functions. In A. Slater & J.G. Brenner (Eds.), Infant development, pp. 7-41. Hinsdale, NJ: Erlbaum.

Argyle, M., & Cook, M. (1976). Gaze and mutual gaze. Cambridge: Cambridge University Press.

Barenholtz, E., Mavica, L., & Lewkowicz, D. J. (2016). Language familiarity modulates relative attention to the eyes and mouth of a talker. Cognition, 147, 100-105.

Birmingham E, Bischof WF, & Kingstone A (2008) Social attention and real-world scenes: the roles of action, competition and social content. Quarterly journal of experimental psychology (2006) 61(7):986-998.

Borkenau, P., & Liebler, A. (1992). The cross-modal consistency of personality: Inferring stranger's traits from visual or acoustic information. Journal of Research in Personality, 26, 183-204.

Buchan, J. N., Paré, M., & Munhall, K. G. (2007). Spatial statistics of gaze fixations during dynamic face processing. Social Neuroscience, 2(1), 1-13.

Costen, N. P., Parker, D. M., & Craw, I. (1996). Effects of high-pass and low-pass spatial filtering on face identification. Perception & Psychophysics, 58(4), 602-612.

Emery, N.J. (2000) The eyes have it: the neuroethology, function and evolution of social gaze. Neuroscience and biobehavioral reviews 24(6):581-604.

Goren, C., Sarty, M., & Wu, P. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. Pediatrics, 56, 544-549.

Haith, M. M., Bergman, T., & Moore, M. J. (1977). Eye contact and face scanning in early infancy. Science, 198 (4319), 853-855.

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). Putting the face to the voice: Matching identity across modality. Current Biology, 13, 1709-1714.

Kleinke, C. L. (1986). Gaze and eye contact: a research review. Psychological bulletin, 100(1), 78.

Krauss, R.M., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. Journal of Experimental Social Psychology, 38, 618-625.

Kuhl, P.K., Meltzoff, A.N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. Journal of the Acoustical Society of America. 100, 2425-2437.

Lachs, L., & Pisoni, D. B. (2004). Cross modal source identification in speech perception. Ecological Psychology, 16(3), 159-187.

Langton, S. R., Watt, R. J., & Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. Trends in cognitive sciences, 4(2), 50-59.

Lansing, C. R., & McConkie, G. W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. Journal of speech, language and hearing research, 42(3), 526.

Lansing, C. R., & McConkie, G. W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. Attention, Perception, & Psychophysics, 65(4), 536-552.

Lass, N.J., & Davis, M. (1976). An investigation of speaker height and weight identification. Journal of the Acoustical Society of America, 60, 700–704.

Leder, H., & Carbon, C. (2006). Face-specific configural processing of relational information. British Journal of Psychology, 97, 19-29.

Legerstee, M. (1990). Infants use multimodal information to imitate speech sounds. Infant Behavior and Development. 13, 343-354.

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. Proceedings of the National Academy of Sciences, 109(5), 1431-1436.

Loveday, L. (1986). Explorations in Japanese Sociolinguistics. Amsterdam, The Netherlands: John Benjamins.

Mavica, L. W., & Barenholtz, E. (2013). Matching voice and face identity from static images. Journal of Experimental Psychology: Human Perception and Performance, 39(2), 307.

McFarland, M. (2015, December 14). 5 amazing and alarming things that may be done with your DNA. Retrieved June 16, 2016, from https://www.washingtonpost.com/news/innovations/wp/2015/12/14/5-amazing-and-alarming-things-that-may-be-done-with-your-dna/

Rossion, B., & Gauthier, I. (2002). How does the brain process upright and inverted faces? Behavioral and cognitive neuroscience reviews, 1(1), 63-75.

Smith, H. M., Dunn, A. K., Baguley, T., & Stacey, P. C. (2016a). Concordant Cues in Faces and Voices Testing the Backup Signal Hypothesis. Evolutionary Psychology, 14(1), 1474704916630317.

Smith, H. M., Dunn, A. K., Baguley, T., & Stacey, P. C. (2016b). Matching novel face and
voice identity using static and dynamic facial images. Attention, Perception, & Psychophysics, 1-12.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. The journal of the acoustical society of America, 26(2), 212-215.

Trehub, S.F. (1973). Infant's sensitivity to vowel and tonal contrasts. Developmental Psychology, 9, 91-96.

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. Attention, Perception, & Psychophysics, 60(6), 926-940.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., ... & Gocayne, J. D. (2001). The sequence of the human genome. science, 291(5507), 1304-1351.

Võ, ML-H., Smith, T.J., Mital, P.K., & Henderson, J.M. (2012) Do the eyes really have it? Dynamic allocation of attention when viewing moving faces. Journal of Vision 12(13).

Yarbus, A.L. (1967). Eye movements and vision (Vol. 2, No. 5. 10). New York: Plenum press.

Yin, R. K. (1969). Looking at upside-down faces. Journal of experimental psychology, 81(1), 141.