Factors That Influence Earwitness Confidence

By

Kelly Thiessen

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfilment of the requirements of the degree of

MASTER OF ARTS

Department of Psychology

University of Manitoba

Winnipeg

Abstract

Researchers have examined the reliability of eyewitness testimony, but earwitness testimony has not received the same attention. The purpose of the current study was to examine whether biases exist that influence earwitness confidence regarding their target selection in a lineup. In this study, participants listened to a recording of a conversation (criminal, neutral, or vulnerable), completed a filler task, and then rated their confidence that each voice was the target in a six-voice lineup. The voices in the lineup varied in pitch (high, medium, and low) and in emotion (emotional or monotone). When the target voice was present in the lineup, participants reported the highest confidence in the target voice when there was a match in pitch and emotion, especially in the criminal context compared to the neutral or vulnerable contexts. In target-absent lineups, participants' confidence ratings for the lineup voices were influenced by a voice frequency heuristic such that participants were least confident in high-pitched voices in the criminal content condition and low-pitched voices in the vulnerable content condition. Confidence ratings also negatively correlated with response time. It is suggested that emotion, content, and pitch should be controlled for in future earwitness research.

Keywords: earwitness testimony, confidence, voice pitch, conversation content, emotion, response time, voice identification

Acknowledgements

I thank my advisor, Dr. Launa Leboe-McGowan, for her guidance in the development of this research. Her support during my time as a graduate student is greatly appreciated. I also thank my committee members, Dr. Jason Leboe-McGowan and Dr. James Hare, for their feedback on my proposal and my research. Finally, I thank my family and friends for their support throughout this process. In particular, I thank my mother for her unwavering support and encouragement of my studies.

This research was supported by a Canada Graduate Scholarship from the Social Sciences and Humanities Research Council of Canada and a Tri-Council Top-Up Award from the University of Manitoba.

Table of Contents

List of Tables

List of Figures

Factors That Influence Earwitness Confidence

Witness testimonies are one source of evidence that authorities use in criminal investigations. Thus, psychology researchers over the past several decades have examined the accuracy of these testimonies: are witness testimonies sufficiently accurate and reliable to use within the judicial system? The majority of research has examined eyewitness testimonies, where a witness of the crime visually observes a lineup of individuals and needs to select the accused. However, less research has been conducted in the area of earwitness testimonies, where witnesses identify a culprit based on their voice. This form of evidence is often used when there are limited visual cues at the crime scene, and sometimes in conjunction with eyewitness testimonies. For example, the criminal may have been masked, in a dark environment, or communicating over the telephone. In this case, authorities would provide a lineup of voices for the witness to listen to and identify the culprit through this method, rather than visual recognition. While earwitness testimonies are not used as frequently in legal cases, there have been instances where these testimonies contributed significantly to a conviction, and problematically, to at least 17 wrongful convictions (Sherrin, 2016).

Two examples of cases where individuals were wrongfully convicted based partially on earwitness testimony are those of Cathy Watkins and Kirk Odom, both in the United States. In September 1997, in New York State, Watkins was convicted of second-degree murder of Baithe Diop, a taxi driver (Possley, 2016). Watkins was brought in as a suspect due to her living at an address that Diop had visited prior to his death and the dispatcher for the company Diop worked for identified Watkins as the caller for the cab by her voice. In combination with eyewitness testimony, Watkins was sentenced to 25 years to life in prison. However, in 2012, the charges against Watkins were dismissed, due to new evidence based on phone records and interviews

that indicated two other individuals were responsible for Diop's death (Possley, 2016). The case of Kirk Odom also demonstrates the fallibility of earwitness testimony. In Washington, D.C., 1981, Odom was convicted of multiple crimes, including sexual assault and armed burglary (Possley, 2015). During the investigation, a police officer was having an unrelated conversation with Odom, and later believed he matched the description the victim had given officers earlier. While the victim had tentatively identified Odom in a lineup of 10 photographs, she wanted to hear his voice as well. She was presented with a live lineup where she was able to listen to their voices and picked Odom from the lineup as the attacker. Odom was later charged, convicted, and sentenced to 20 to 60 years in prison. In July 2012, Odom's charges were dismissed, based on DNA evidence that was evaluated during the previous year (Possley, 2015). In both cases, while earwitness testimony was used to identify a suspect, the witness was mistaken, and it resulted in several undeserved years in prison. Thus, it is critical that researchers of earwitness testimony identify the mechanisms that are used during earwitness testimonies and lineups to help prevent further wrongful convictions.

Despite the apparent usefulness of witness testimonies, psychological research has shown that biases exist when identifying a culprit. Several reviews of eyewitness studies have discussed the variables that can account for inaccurate identifications in eyewitness testimonies (e.g. Lloyd-Bostock & Clifford, 1983; Wells & Loftus, 1984; Wells, Memon, & Penrod, 2006). Wells et al. (2006) described three types of variables that may impact the accuracy of eyewitnesses, including estimator variables, system variables, and postdiction variables. Estimator variables are circumstances that typically occur at the scene of a crime and authorities of the criminal case have little to no control over. For example, the length of time the witness is able to view the culprit and the race of both the culprit and the witness can affect how accurate an eyewitness will

be when identifying the culprit at a later point. In a meta-analysis of facial identification and eyewitness studies, Shapiro and Penrod (1986) found that the longer a witness viewed a target person during the original event, the more accurate they were at identifying the person in an eyewitness lineup (r = .19). In a review of studies about facial identification and its relationship to race, Meissner and Brigham (2001) found that participants were 1.4 times more likely to correctly identify someone of their own race compared to someone of a different race. Furthermore, participants were 1.56 times more likely to make a false identification of a different-race face compared to a same-race face, suggesting that participants were more successful at identifying someone of the same race as themselves compared to someone of a difference race (Meissner & Brigham, 2001).

The second variable identified by Wells and colleagues (2006) that may influence witness identification is system variables, which are factors that members of the judicial system can influence (Wells et al., 2006). How a witness is interviewed following a crime and how the lineup is created and administered can have a wide range of effects on witness accuracy. For example, in a study conducted by Haw and Fisher (2004), the researchers told the lineup administrators which member of the lineup was the target and then manipulated how much contact the lineup administrator had with the participants who were trying to identify the target from the lineup. Participants were presented with a video and completed a lineup task 15 minutes later. There were two types of lineups, target present and target absent. In the target-present condition, the lineup administrator was accurately informed of who the target was. However, in the target-absent condition, the lineup administrator was told that the target-substitute was the actual target. As such, for the target-absent condition, the analysis was whether participants would be more likely than chance to select the target-substitute from the lineup depending on

how much contact the participants had with the lineup administrator. Haw and Fisher found that when the lineup administrator had high levels of contact with the participant by sitting across from them at a table and showing them the photos of the lineup members, participants were more likely to identify the target that the lineup administrator believed to be the target (proportion of selection = .31) compared to what was expected by chance (.17), despite the true target not being present in the lineup (target-absent lineup). In comparison, participants in the low contact group, where the lineup administrator only observed the participants completing the lineup task, selected the target substitute no more often than by chance (.19), suggesting that participants in the low contact group were not influenced by the lineup administrator's belief of who the target was (Haw & Fisher, 2004). This suggests the need for lineups to be conducted in a double-blind format, in that the person who is administering the lineup does not know who the suspect (target) is in the lineup, nor do they know if the target is present. This would help reduce the amount of influence the administrator has on the witness' decision.

Lastly, postdiction variables refer to factors that correlate with witness accuracy rather than having a causal effect (Wells et al., 2006). For instance, the relationship between confidence and accuracy has been studied extensively, with varying results. While it may be intuitive to believe that the more confident a witness is, the more accurate they are, eyewitness research has found a weak relationship between the two variables (Bothwell, Deffenbacher, & Brigham, 1987; Sporer, Penrod, Read, & Cutler, 1995; Wells et al., 2006). Sporer et al.'s (1995) meta-analysis found a positive correlation of .28, although it was stronger for participants who selected a target in a lineup (.37) than participants who rejected the lineup (.12). Despite the positive correlation, they were hesitant to support confidence's use in the court system as an indicator of accuracy, particularly because it was not a strong correlation and that other variables

such as viewing conditions at the time of the event could distort confidence levels (Sporer et al., 1995). Bothwell et al. (1987) came to a similar conclusion, with their meta-analysis indicating a correlation of .25 between witness confidence and accuracy. The authors also found the correlations improve the longer a participant has to view the target face initially (.52). Bothwell et al. (1987) did conclude by mentioning that confidence measures in real life are more subjective as they are determined by the people observing the witness, such as a jury, and thus the relationship between accuracy and confidence could be distorted.

While variables that influence eyewitness testimonies have been and continue to be well documented, the factors that play a role in voice recognition and how they impact a person's ability to correctly identify a voice are not as well known. When a witness selects an innocent person in a lineup as being the culprit, it can have serious implications to the falsely accused individual's future if the court accepts this as evidence. Despite the lack of research done in earwitness testimonies, courts often accept voice identification evidence. In both Canada and the United States, individuals have been wrongfully convicted in court cases that used earwitness testimony to establish the guilt of the accused (see Sherrin, 2016). In a study examining the ways courts determine whether earwitness testimony is admissible as evidence, it was found that of the 226 cases that were included in the analysis, only in 11 cases were voice identifications deemed as inadmissible (Laub, Wylie, & Bornstein, 2013). Between the wrongful convictions and the generosity of the courts to allow voice identification as evidence, understanding how people identify voices, the mistakes that can occur during the process, and the biases that exist when selecting a voice, are important areas that need to be investigated in order to help inform judicial systems how to best utilize earwitness evidence.

Earwitness studies have examined a large range of variables that may impact target identification. For example, Clifford (1980) conducted a review of earwitness studies and presented some of his own research. He examined a number of variables, including length of speech, voice disguise, and delay between the original event and the earwitness lineup. First, the length of speech that an earwitness heard during the original event has some influence on target identification accuracy. Clifford (1980) had adults listen to a voice speak either one, two, or four sentences, and then required the participants to identify the target in a lineup of six voices that included the target voice. Accuracy did not differ depending on how many sentences were spoken in the original event. In comparison, when he conducted a similar experiment with participants aged 12 to 16, Clifford found that participants were more accurate when two sentences were presented at the original event (49%) compared to one sentence (36%). The final experiment examining length of speech had adult participants listen to either a one or eight-word sample and had them identify the target from a lineup of five or eleven distractor voices. Results indicated that participants were more accurate with the eight-word sample (Clifford, 1980). Clifford concluded that accuracy is best when at least one sentence is spoken, but speech longer than that appears to only benefit children.

A second variable examined by Clifford (1980) was voice disguise and the difference in vocal characteristics between the event and the voice lineup. A culprit's purposeful distortion of their voice during the original event can reduce the accuracy of the earwitness during a later lineup (Clifford, 1980). For example, one of the reviewed studies had participants listen to an undisclosed length of speech by voices familiar to them in everyday contexts and had them identify the voice (Pollack, Pickett, & Sumby, 1954). They found that it took participants three times as long to identify a whispered voice compared to a voice of a normal volume (Pollack et

al., 1954). Emotion is another way a voice could be disguised, and the change in emotion between the original event and the lineup reduces accuracy (Saslove & Yarmey, 1980). Saslove and Yarmey (1980) presented participants with a recording of a female voice speaking in a hostile tone for 10 to 12 seconds. Participants were then presented with a five-voice lineup, where all the voices spoke in either a hostile tone or a non-hostile tone. Using a confidence and accuracy measure, with a maximum score of six, Saslove and Yarmey (1980) found that participants were more accurate at identifying the target when the lineup was presented in a hostile tone (M = 5.15) compared to if the lineup was presented in a non-hostile tone (M = 2.37). Clifford's (1980) own study presented participants with disguised voices, with the disguise being implemented however the voice actor desired. Participants then had to identify the target voice in an undisguised voice lineup consisting of four to eight alternative voices. Clifford found that participants were much less accurate in this study (26%) compared to his previous studies with undisguised voices (65%). The author concluded that disguising the voice during the original event reduced participants' accuracy at identifying the target voice in a lineup that included voices spoken in a normal tone (Clifford, 1980).

Lastly, the amount of time between the original event and the lineup, or the retention interval, has been shown to impact identification accuracy. While there is some variability in the results, Clifford's (1980) studies generally showed that the longer the delay is, the less accurate an earwitness will be in their identification. For example, one of his studies used one, two, and three-week retention intervals, and found that accuracy was significantly worse in the three-week condition (9%) compared to one (50%) and two weeks (43%). A follow-up study by Clifford (1980) used retention intervals of 10, 40, 100, and 130 minutes, finding no difference in accuracy rates, which ranged from 40 to 56 percent. In his final two studies examining retention intervals,

he found that a ten-minute delay was significantly more accurate (40.8%) compared to 24 hours (20%), seven days (23.25%), and 14 days (19.25%). Overall, Clifford (1980) concluded that earwitness testimonies are prone to errors to the same degree as eyewitness testimonies, and that the judicial system should be as cautious with earwitness testimony as they are with eyewitness testimony.

More recently, researchers have examined how the acoustic environment, in combination with retention interval and speech duration, affect earwitness accuracy (Kerstholt, Jansen, Van Amersvoort, & Broeders, 2004). Kerstholt et al. (2004) presented participants with an auditory clip of a man answering questions about aspects of his life. Participants were then required to identify the voice in a lineup of six voices, with half of the participants choosing from a lineup with the target present, and the other half choosing from a target-absent lineup. A correct response from a participant in a target-absent lineup was defined as stating that the target was not in the lineup. Manipulations to the acoustic environment were implemented through the original recording. One of the recordings was a single, uninterrupted monologue recorded inside of a building. The other recording took place in two environments. The first half of the recording was recorded indoors and the second half was recorded outdoors, with a female voice separating the two parts. It was expected that accuracy performance would be worse for the recording with the outdoor environment, as the auditory quality was reduced. All the voices used in the lineup were recorded indoors. The researchers also manipulated retention interval (immediate/one week) and the speech duration of the original auditory clip (30 seconds/70 seconds). While the results showed no significant difference between types of acoustic environment, there was an interaction between retention interval, exposure duration, and whether the target was present or absent from the lineup. Participants who were presented with a target-absent lineup, a 70 second speech

duration, and a retention interval of one week were more accurate in target identification compared to any other condition.

A follow-up study conducted in a similar manner examined the impact that voice accents and speaking over the telephone had on identification accuracy, as well as a retention period of either three weeks or eight weeks (Kerstholt, Jansen, Van Amelsvoort, & Broeders, 2006). Two lineups were created for this experiment, one consisting of voices with a standard accent and one consisting of voices with a strong regional accent, as determined by the researchers. The voices were evaluated by naïve individuals to ensure there were no voices that were noticeably different than the others so that that the lineup would be considered fair. There were also four experimental conditions that incorporated speaking over a telephone and accented voices. The first condition included both the original recording and the lineup voices were presented in a standard accent and recorded live. Second, the original recording and the lineup were spoken in the regional accent and recorded live. Third, both sets of voices were recorded over the telephone and were spoken in the regional accent. Finally, the fourth condition had the regional accent, but the original event was recorded live whereas the lineup voices were recorded over the telephone. The researchers justified this final condition as it could be possible that a witness had heard the target voice in a live situation and a live recording of the target voice was not available for the lineup, but a telephone recording was possible. Results showed that participants who heard a voice without a distinct accent were marginally more accurate (target-present = 41%; target-absent = 56%) at identifying the target voice within a lineup compared to participants who heard an accented voice (target-present = 34%, target-absent = 35%). There was no significant difference between a retention interval of three weeks or eight weeks. However, when comparing the three-week interval with the one week interval of the first Kerstholt et al. (2004) study,

participants in the three-week interval condition were significantly more accurate, but only when the target was absent from the lineup (one week = 35%; three weeks = 65%). Finally, there was no significant effect in regards to the telephone variable (Kerstholt et al., 2006).

There is some research investigating whether the sex of the speaker or the listener has an impact on target identification accuracy. In a study conducted by Wilding and Cook (2000), male and female participants listened to two samples of speech, one spoken by a male and one spoken by a female. One week later, the participants returned to the lab and listened to two lineups, one for each original speech sample, with all voices in each lineup matching the sex of the original speaker. There were no significant main effects in regard to listener and speaker sex, but there was a significant interaction between the two. When participants were identifying female voices, female participants were more accurate (51%) compared to their male counterparts (38%). In contrast, male and female participants were equally accurate when identifying male voices (43% and 41%, respectively). In Clifford's (1980) review article, some of his own studies touched on sex differences of speakers and listeners. When examining speech duration, Clifford had participants listen to six target voices, each one speaking one, two, or four sentences, depending on the experimental condition. For each target voice, participants were presented with a six-voice lineup that consisted of the target voice and five alternative voices. All voices in the lineup spoke the same single sentence. While there was no significant difference between the speech durations, Clifford (1980) did find that both male and female participants were more accurate when identifying female voices (85%) compared to male voices (75%), whereas Wilding and Cook's (2000) study found that only female participants were more accurate at identifying female voices. However, a later study conducted by Clifford (1980) did not show significant differences regarding sex. In this study, participants were presented with six trials, where each

trial consisted of hearing a disguised voice speaking one sentence which was immediately followed by a lineup of undisguised voices that included the target voice, all speaking the same sentence participants just heard. The number of alternative voices was four, six, or eight, as a between-subjects variable, and every participant heard three trials with female voices and three trials with male voices. Clifford (1980) states that there were little to no differences in accuracy rates between male and female participants and speakers, although no data were provided to support this claim.

Campos and Alonso-Quecuty (2006) examined the type of information that witnesses remember about a conversation they heard; such as whether the witnesses remembered verbatim or gist information. The researchers were also interested in whether having both auditory and visual information during the conversation would aid in recalling the content of conversation, in comparison to only have auditory information. Participants in the audiovisual condition watched a 15-minute conversation about a planned theft between two individuals. The video was recorded to make it look like it came from a security camera on a building. As such, the recording took place at night with black and white footage. Participants in the auditory-only condition were presented with just the audio from the conversation. All participants were told to behave as if they were real witnesses to the crime and they would need to recall this information later. Participants then completed a 15-minute distractor task, either immediately following the conversation or four days later. After the distractor task, participants completed a free-recall task where they were asked to write down as much as they remembered about the conversation. To score the results, the dialogue was split into 374 units. The responses given by participants were similarly split and compared to the dialogue units to be scored as gist recall, verbatim recall, fabricated ideas, or distorted ideas. Campos and Alonso-Quecuty (2006) found that across all

conditions, participants recorded more gist recall items (M = 12.09 of 374 units) compared to verbatim recall items (M = .31). The researchers attributed this to the participants being informed that they would need to later recall the content of the message, but not that they would need to recall it verbatim, so participants may have been biased to rely on gist information. They also found that both gist and verbatim recall were more prevalent when the free-recall task was administered immediately after the conversation (M_{gist} = 14.90; $M_{verbatim}$ = .57), rather than four days later (M_{gist} = 9.28; $M_{verbatim}$ = .05). While there was no significant difference between the audiovisual and the auditory-only conditions in terms of verbatim recall, participants in the audiovisual conditions recalled more gist information (M = 13.68) than their auditory-only counterparts (M = 10.50), suggesting that there is a benefit to witnessing an event with both modalities rather than only hearing it. Finally, in regard to errors, the researchers examined fabricated information, or writing down information that was not present in the original conversation, and distorted information, where a detail of a statement recorded during the free-recall task was incorrect compared to the original conversation. They found that fabrications were the more common form of error (M = 1.98) compared to distortions (M = 1.39), and that participants in the auditory-only condition as well as participants in the delayed retention interval condition made more fabrications compared to the other conditions. The authors suggested that in real-life situations, earwitnesses would be more likely to provide details about the encounter after a delay. Therefore, the authors suggest that the use of witnesses who only have auditory information about the crime should be used cautiously when explaining details of the event, as they are more likely to recall fabricated information (Campos & Alonso-Quecuty, 2006).

The current experiments were designed to investigate a variety of factors that could influence earwitness confidence beyond what was investigated in the studies described above. In

the current study, participants indicated for each voice in an earwitness lineup how confident they were that the voice was the target voice from an earlier presented conversation. Differences in confidence ratings between voice frequencies were examined to see if participants would be biased towards voices of a certain pitch. Conversation content was also manipulated, with participants hearing a conversation of either criminal, neutral, or vulnerable content. It was anticipated that voice pitch would interact with conversation content, where participants would have higher confidence ratings for voices presented in a low-pitch, criminal condition and in a high-pitch, vulnerable condition. This would be based on preconceived expectations about characteristics of criminals and victims. Vocal emotion was incorporated into the experiment, examining whether it would be beneficial for earwitness' confidence to have a lineup with emotional voices compared to monotone voices. To determine if there was a relationship between response time and confidence, the amount of time it took for participants to respond with their confidence ratings was also recorded. Finally, these variables were incorporated into two experiments, one with a target-present lineup and one with a target-absent lineup, as the influence of these factors may present differently depending on whether the target is included in the lineup.

Correlation between Accuracy and Confidence

A major concern about the use of earwitness testimonies is the degree of confidence a witness has about their judgment and how this confidence relates to their accuracy. Should a jury accept the testimony of a witness if they have a high level of confidence? Previous research has indicated that there is a lot of variation in the correlation between accuracy and confidence, with results ranging from -.48 to .53, as well as studies that indicate no significant correlations (Öhman, Eriksson, & Granhag, 2011; Orchard & Yarmey, 1995; Yarmey & Matthys, 1992).

Researchers have strongly cautioned against using confidence as a determinant for earwitness accuracy, even going so far as saying that earwitness testimonies are less reliable than eyewitness testimonies (Deffenbacher et al., 1989; Kerstholt et al., 2006; Olsson, Juslin, & Winman, 1998). However, according to Sherrin (2016), the Supreme Court of Canada more strongly cautions the use of confidence for eyewitness accounts compared to earwitness accounts, suggesting a lack of knowledge of how confidence relates to earwitness testimonies.

A particularly concerning finding about the correlation between accuracy and confidence is that participants will express high levels of confidence in earwitness studies, despite having made an incorrect voice identification (Olsson et al., 1998). Moreover, overconfidence appears to be a more serious issue with earwitnesses than eyewitnesses (Olsson et al., 1998). For instance, Olsson et al. (1998, Experiment 1) had participants listen to a two minute criminal conversation between four individuals, with two of the people speaking for 40 seconds and the other two people speaking for 15 seconds. Participants were then tasked with identifying each individual from the criminal conversation in four separate earwitness lineups either one hour or one week later. Only half of the participants were informed that they would be presented with lineups during the second phase of the study. There were two lineup conditions that were administered between subjects, with the difference being how the distractor voices were selected. The first lineup consisted of distractor voices that fit the general description of the target voice. The target voice was male, aged 20 to 30, and did not have a noticeable dialect. The researchers had created a voice library from participants who did not participate in the lineup study. From this voice library, they selected seven voices that matched the description of the target voice. The other lineup condition selected distractor voices based on how similar they sounded to the target voice. Two independent judges indicated which voices in the voice library sounded the most similar to

the target voice, and the seven most similar voices were included in the lineup. The researchers predicted that given the greater variation of the distractor voices in the general description lineup, participants would be more accurate in identifying the target voice. Participants were asked to make an identification if they thought the target was in the lineup, and then indicate on an 11-point scale how confident they were in their identification. Following the identification task, participants were presented with the suspect voice and each of the distractor voices and rated how similar they were on a scale from 0 to 1000. While the voices in the similar condition were rated as more similar to the target voice (M = 282) compared to the general description condition (M = 222), participants were not significantly more successful at identifying the target voice in the general description condition (similar condition = 15%; general description condition = 20%). Olsson et al. (1998) found that across all conditions, participants who stated that they were 100% confident about their selection were only accurate 30-40% of the time. In other words, participants were overconfident in their selections. The researchers then compared their results to a previous eyewitness study (Juslin, Olsson, & Winman, 1996), where participants had viewed eyewitness lineups that were made up in similar ways as the earwitness lineups. In the eyewitness study, participants were only slightly overconfident when the lineup images were similar to the culprit, and slightly under confident when the lineup images had more variability. Thus, earwitnesses appear to be more likely to be overconfident in their target selection compared to eyewitnesses. The authors concluded that judicial systems should be more cautious with earwitness testimonies than they are with eyewitness testimonies if using confidence as a measure of accuracy (Olsson et al., 1998).

Further research provides more evidence for the cautionary use of confidence to predict the accuracy of earwitness testimonies. Orchard and Yarmey (1995) conducted a study

examining the influence that distinctive voices have, as well as whispering compared to normal tones of voice. To mimic being kidnapped, the participants were blindfolded and seated in a room. Participants then listened to either a 30-second or eight-minute monologue from the kidnapper, who was the target voice. The voice that spoke the monologue was either distinctive or non-distinctive and spoke in either a whisper or a normal tone. A distinctive voice was determined by a pilot study of ten undergraduate students who rated voices on a 10-point Likert scale, based on whether the voice was unique compared to the other voices and "highly striking" (Orchard & Yarmey, 1995, p. 252). The two highest rated voices were selected as distinctive voices, which had consistent ratings of eight or higher. Non-distinctive voices were randomly chosen from lower rated ones, with an average score of 4.5. Two days after hearing the monologue, participants were presented with two earwitness lineups, one with the target included and one without, with the order of presentation counterbalanced across participants. The participants' task was to identify the target and to rate how confident they were in their decision. Both lineups presented to participants were spoken in either a whispered or normal tone, resulting in three experimental conditions based on tone of voice between the monologue and the lineup; both were spoken in a normal tone, both were spoken in a whisper, or the monologue was whispered and the lineup was spoken normally. Results were analyzed using a 10-point combined accuracy and confidence measure, where higher scores indicated higher accuracy and confidence. Orchard and Yarmey (1995) found that in the eight minute monologue, target-present condition, participants were more accurate (M = 6.23) compared to those in the 30-second condition (M = 4.67). Moreover, participants who heard a normal spoken monologue and a normal spoken lineup performed better than participants in the whisper monologue, normal lineup condition. Inferred from the graph provided by the authors, participants in the normal-

normal condition had a mean score of six, whereas participants in the whisper-normal condition had a mean score of three. This finding was also observed in the suspect absent lineup in the eight-minute condition, with a mean score of 8.75 for the normal-normal condition and 4.50 for the whisper-normal condition. These results suggest that having a match in tone between the original event and the lineup facilitates participants' recognition of the target voice. In regard to confidence, when all conditions were combined other than target presence, there was a significant positive correlation between confidence and accuracy, $r = .25$ and $r = .36$ for target-present and target-absent conditions, respectively. However, when looking at individual conditions, the results varied. In particular, there was actually a negative correlation between accuracy and confidence when identifying distinctive voices ($r_{present} = -.48$ and $r_{absent} = -.18$) , suggesting that for this condition, the more confident an earwitness was, the less accurate they were at identifying the target voice within a lineup (Orchard & Yarmey, 1995).

Yarmey and Matthys (1992) also conducted an earwitness study that showed negative correlations between accuracy and confidence in particular conditions. Their variables included the duration of the original voice, the retention interval, and the number of exposures the participant had to the original voice. Additionally, there were male and female speakers and target-present or target-absent lineups. In terms of accuracy, there were no significant effects for any of the conditions in the target-absent lineups. In other words, when the target was not in the lineup, participants' accuracy rates did not differ significantly based on the duration of the original voice, the number of exposures they had to the original voice, or how long before they were presented with a lineup. In the target-present lineups, participants who heard longer durations of speech (120 seconds and six minutes) were more accurate compared to participants in the shorter duration conditions (18 and 36 seconds), with a mean proportion difference of

approximately .20 based on the figure provided by the authors. Additionally, participants who were exposed to the original voice twice were more accurate (50%) compared to participants who heard the voice once (34%) or three times (35%). The number of exposures was implemented so that there was a five-minute break between each exposure, and the authors suggested that this five-minute break may not be enough to improve accuracy rates with more exposures. This does not explain the difference in accuracy between two and three exposures, and the difference was not addressed by Yarmey and Matthys (1992). Overall, there was no significant correlation between accuracy and confidence. However, when broken down into the individual conditions, correlations ranged from -.42 to .53. For example, participants who listened to an 18-second voice were less accurate at identifying the target voice within an earwitness lineup if they were highly confident with their selection. In contrast, participants in the 120 second and six-minute conditions were more accurate when they were more confident in their target selection. The participants in the 36 second condition did not demonstrate a significant correlation between accuracy and confidence. Within each of the conditions, there was a range of significant and nonsignificant correlations. The authors concluded that confidence was not a reliable indicator of earwitness accuracy, and that even in ideal situations, voice identification is a difficult task (Yarmey & Matthys, 1992).

Other experiments described in the current paper also used measures of confidence that did not show a relationship between accuracy and confidence. For example, Kerstholt and colleagues (2004, 2006) did not find a significant relationship between accuracy and the confidence of the participant. Kerstholt et al. (2004) did find that the research assistant's confidence in the participant's accuracy was more predictive of accuracy than the participant's confidence rating. The research assistant who administered the task was unaware of whether the

target was included in the lineup being presented to participants. Prior to participants making a confidence judgment, the research assistant recorded how confident they were that the participant had selected the target voice. These results showed that after accounting for chance, 23% of participants' responses could be predicted by the research assistant's confidence judgment. In comparison, participants' confidence judgments did not reliably predict their accuracy. However, Kerstholt et al. (2006) were unable to replicate the research assistant results, and also found that participants' confidence did not predict their identification accuracy. Similarly, Öhman et al. (2011) conducted an earwitness study with children and adults and found overall that there was no significant correlation between confidence and accuracy for either age group.

Due to the lack of consistent, positive correlations between accuracy and confidence, it has been suggested, even by those in the field of law enforcement, that the judicial system should not consider confidence when evaluating the accuracy of an earwitness' testimony (Sherrin, 2016). While the judicial system should be careful when using confidence as a judgment of accuracy, the fact that witnesses feel confident about certain voices in a lineup could affect their target selection. If witnesses were to listen to a lineup of voices, it is likely that they would pick the person who they feel most confident about as being the criminal, even if that person is not the accurate target. Thus, knowing the factors that impact confidence levels is just as important as knowing what factors influence correct identifications. Due to this, the current study used confidence ratings as the dependent variable in the experiments and examined how certain biases may impact ear witness confidence ratings.

Voice Frequency

The limited research done on the impact of voice frequency suggests that lower and higher frequency voices can influence who a person chooses as the target voice within a lineup (Mullennix et al., 2010; Rodstrom & Neuhoff, 2003; Stern, Mullennix, Corneille, & Huart, 2007). For example, Rodstrom and Neuhoff's (2003) study had two voices, one male and one female, that spoke a single vowel sound. They then made copies of this voice and artificially changed the frequency of it by half semitones, resulting in 16 higher frequency voices and 16 lower frequency voices. Participants heard the original voice three times, then they were presented the altered voices one at a time, and the participants had to determine if the presented voice was the same as the original voice they heard. Participants were more likely to indicate that a higher frequency voice was produced by the same person as the original voice sample, compared to lower frequency voices. Since all presented trials were of the same voice, it appears that participants were more easily able to recognize a higher frequency voice compared to a lower frequency voice.

Two studies have looked at how the pitch of the original speech sample influences what distractor voice participants choose in an earwitness lineup (Mullennix et al., 2010; Stern et al., 2007). In a study conducted by Stern, Mullennix, Corneille, and Huart (2007), participants listened to a computer-generated voice sample of five unrelated words. Participants were then tasked with identifying the original voice from pairs of voices that spoke the same set of words as the original sample. One voice in the pair was always the target voice. The distractor voice was the same computer-generated voice but in an altered pitch (lowest, low, high, and highest). There was a total of eight trials, half with the target presented first and the other half with the target was presented second. Additionally, half the trials had the altered voice as a higher pitch and the other half as a lower pitch. As such, participants were presented with each distractor

pitch twice. Responses were scored based on the number of errors participants made for a particular distractor, resulting in a maximum score of two. Results showed that when the original voice was high pitched, participants were more likely to select a higher-pitched distractor voice ($M_{highest}$ = 1.00; M_{high} = 1.05) than a lower-pitched one (M_{lowest} = .71; M_{low} = .90). The reverse was true for a low-pitched original voice, where low-pitched distractors (M_{lowest} = .95; M_{low} = 1.25) were picked more frequently than high-pitched distractors ($M_{highest}$ = .65; M_{high} = .60). On the other hand, if the original voice sample was of moderate pitch, participants were equally likely to pick a higher- or lower-pitched distractor item ($M_{highest}$ = .63; M_{high} = 1.19; M_{lowest} = .88; M_{low} = 1.12). These findings held true for both male and female sample voices (Stern et al., 2007). In a follow-up study that examined both speaking rate and vocal pitch, it was found that while speaking rate did not have a significant influence on choice of distractor voice, vocal pitch impacted the participant's choice in the same way as the previous study (Mullennix et al., 2010). That is, when participants were presented with a high-pitched voice, they were more likely to pick a higher-pitched distractor voice, compared to a lower-pitched distractor. The opposite held true for the low-pitched original voice as well (Mullennix et al., 2010). These two studies indicate that the original speaker's pitch can have an impact on the errors listeners make, in that they are more likely to pick a particular pitch dependent on the original pitch.

In a study examining child and adult earwitnesses, researchers observed that children use pitch level and pitch variation, along with articulation rate, when determining the target voice (Öhman et al., 2011). To determine if children could perform earwitness tasks at the same level of accuracy as adults, Öhman et al. (2011) conducted an earwitness task with children aged 7 – 9 and 11 – 13, and with adults with a mean age of 30.26. Participants listened to one side of a phone call of a criminal conversation that was age appropriate for the children, and then returned

two weeks later to listen to a lineup, with half of the participants being presented with a target-present lineup and the other half with a target-absent lineup. Interestingly, the older children (11-13 years old) were the only group to accurately identify the target above chance (12.5%) in the target-present condition (27%). All age groups correctly rejected the lineup in the target-absent condition more often than chance (M = 33.83%). As for false identifications, adults were more likely to choose a distractor voice in both lineup conditions (target-present 50%; target-absent 60%) compared to the younger children (target-present 36%; target-absent 49%) and the older children (target-present 46%; target-absent 49%). However, when the children did select a distractor voice, they did so with a particular pattern in terms of articulation rate, pitch level, and pitch variation. The faster the distractor voice spoke and the more pitch variation there was in their voice, the more likely the children would incorrectly select that voice as the target. In addition, the lower the distractor's voice pitch, the more likely the children would incorrectly identify the distractor voice as the target. These correlations were not seen in the adult participants, suggesting that children are more likely to use acoustic cues when identifying voices, compared to adults. Taken together, these results suggests that the older children were performing as well as, if not better, than the adults in the earwitness task, although the authors strongly advised that more research be conducted with this age group before making recommendations to court systems (Öhman et al., 2011).

The current study aimed to provide more information about how vocal pitch is utilized in earwitness lineups. With the exception of Öhman et al. (2011), the previous studies altered the pitch of the voices using a computer program, which could make the voices sound less natural and therefore, less likely to be picked within a lineup (Mullennix et al., 2010; Rodstrom & Neuhoff, 2003; Stern et al., 2007). To address this potential confound, all speakers in the current

study were recorded speaking at a high, normal, and low frequencies, thus creating more natural sounding auditory clips for the earwitness lineup. Additionally, this study examined whether a frequency heuristic can be induced through the content of the message. A frequency heuristic refers to the likelihood that an earwitness will associate a particular pitch of voice with a certain event. For example, people may associate lower-pitched voices with criminality, and due to this, may be influenced to select a lower-pitched voice in an earwitness lineup. On the other hand, people may associate higher-pitched voices with vulnerability, and may be more likely to select a higher-pitched voice if the original conversation was of a vulnerable nature. Since the previous research looked at either single syllable sounds or criminal content only, it is unknown if voice frequency is impacted by the content of the message.

Conversation Content

Most research on the accuracy of earwitness testimony has relied on stimuli of either single sounds or words. When participants are presented with a longer dialogue, it is generally related to a criminal event, such as a kidnapping (e.g. Orchard & Yarmey, 1995). However, there might be something in particular about criminal events that can impact voice selection during the lineup. For example, it is possible that people would be more biased towards choosing a lower-pitched voice as a criminal. A criminal offender is generally seen as a male, especially for more serious crimes such as murder, armed robbery, and rape (Allison, Sweeney, & Jung, 2013). This representation matches with criminal statistics in Canada, in that women are less likely to be admitted to adult correctional services than men (Reitano, 2017). Since men typically have lower-pitched voices than women, it is possible that during an earwitness situation, if the content is criminal, there may be a bias to pick a lower-pitched voice from a lineup, even if the speakers are female. Furthermore, television shows more frequently display men as the perpetrators of

crime and violence, and women as the victim of crime (Parrott & Parrott, 2015). This differentiation could mean that if the content of the speech was vulnerable in nature, rather than criminal, an earwitness may be more likely to pick a higher-pitched voice, since women's voices tend to be higher pitched than men's voices.

To my knowledge, only one study has examined how the content of speech impacts voice recognition, and it was in combination with voice emotion. Read and Craik (1995) had participants listen to six statements and rate them on emotionality, with one statement being the target statement that was intended to be the most emotional of the six. Each statement was spoken by a different person and spoken in the appropriate emotion for the statement. For example, the target statement, "Help me, help me. Oh God, help me!" was spoken with more emotion than a filler statement, such as, "I went to the movie last Saturday night" (Read & Craik, 1995, p. 9). Participants' emotionality ratings of the statements did indicate that the target statement was more emotionally spoken than the filler statements. After a 17-day delay, participants were asked to identify the voice that had spoken the target statement within two sets of lineups of six speakers. The first lineup was comprised of the speakers reading lines from a play in a conversational tone, creating a lineup that was of different emotion and content compared to the target statement. The second lineup was of the same six speakers reading the target statement in the appropriate emotion, creating a same emotion and content lineup. These researchers found that when the lineup had the same characteristics as the original event, participant's accuracy was significantly better (66%), compared to when the lineup was different (20%). However, Read and Craik's (1995) study only tested for a differences of content and emotion between the event and the lineup. The current study was designed to examine differences between a criminal event and other events, such as when a person in a vulnerable

position is to be identified, or a neutral event that is less emotionally charged than a criminal event. The key interaction tested for was between conversation content and voice frequency, to determine if participants were using a frequency heuristic as described previously. Witnesses may bias their judgments based on the content of the conversation they heard, in that witnesses may be more inclined to select a lower-pitched voice from an earwitness lineup after observing a criminal event. In comparison, if they need to identify someone who was put in a vulnerable position, they may be biased to select a higher-pitched voice from the lineup. Support for this hypothesis would suggest that earwitnesses are biased when making judgments about a lineup, and that these biases are influenced by the type of event that they witnessed and their preconceived notions of perpetrators of crimes and vulnerable individuals.

Voice Emotion

Criminal events tend to be filled with emotion, with the perpetrator's voice reflecting this. However, perpetrators may attempt to speak in a more natural and calm voice during a police lineup, which could impact witness voice recognition (Saslove & Yarmey, 1980). Studies have shown that when there is a difference in tone between the original event and the lineup, accuracy levels are reduced significantly, in some cases to the point of chance (Mullennix, Bihon, Bricklemyer, Gaston, & Keener, 2002; Orchard & Yarmey, 1995; Read & Craik, 1995; Saslove & Yarmey, 1980). For example, Mullennix et al. (2002) demonstrated this in two of their experiments. In their first experiment, participants responded to several trials where they heard two names, Todd or Tom, and had to determine whether they were the same name or different names. The name was spoken in either an angry tone or a surprised tone. Additionally, it was spoken by the same voice or by different voices. When the two clips differed in emotion, participants had significantly more errors in their decision (same voice = 6.3%; different voice =

6.7%), compared to when the clips had the same emotion (same voice = 4.0%; different voice = 5.4%). These results were replicated in their second study, where the emotions used were anger and commanding. This suggests that the detriment to accuracy when comparing different emotions is robust, even if the emotions are similar. The authors concluded that it would be disadvantageous to compare voices of different emotional tones (Mullennix et al., 2002).

There is evidence that there is a detrimental effect to voice recognition when the voice lineup contains less emotion compared to the original event. Saslove and Yarmey (1980) had participants come into the lab and do a decoy task, and presented them with a short auditory clip of someone speaking in a hostile tone over the telephone. Half of the participants were informed ahead of time that they would be listening to this clip in addition to the other task and that they would need to identify the voice, either immediately following the event or 24 hours later. The other half of the participants were not informed about the voice recognition task, and the auditory clip was played without warning during their decoy task. Retention intervals were also incorporated into Saslove and Yarmey's (1980) study. In one condition, participants were presented with the voice recognition task immediately. In the other condition, participants returned to the lab after 24 hours to complete the voice recognition task. Thus, there were four conditions between the provided instructions and the retention interval. During the recognition task, participants were either presented with a lineup of voices speaking the conversation clip in a hostile voice or in a conversational voice. Participants had to identify the target voice and rate their confidence. The results were analyzed using a combined accuracy and confidence measure with a maximum score of six. The researchers found that when the tone of voice was consistent between the event and lineup, participants were more accurate in identifying the voice (M = 5.15) compared to when the tone of voice had changed (M = 2.37), even when the participants

were expecting to identify the voice and did so immediately after the event (Saslove & Yarmey, 1980). These results are consistent with a study described earlier by Read and Craik (1995), where participants listened to various sentences spoken in the appropriate tone of voice for their content. Read and Craik found that when the lineup contained the identical tone of voice, performance was significantly better compared to if the lineup was spoken in a conversation tone. It was also found that participants' confidence in their selection differed depending on whether the emotion was identical to that of the event and whether their selection was accurate. For the conversational lineup, participants confidence ratings, out of four, were 2.11 and 1.87 for correct and incorrect selections, and not significantly different from each other. In comparison, for participants presented with an identical emotion lineup, confidence ratings were significantly higher for correct selections (2.63) than for incorrect selections (1.88). This indicated that having an identical emotion in the lineup task improved both participants' accuracy and their confidence in their selection. These findings were also replicated when using voices that were familiar to the participants (Read & Craik, 1995).

In combination, these studies suggest that if police do not require the suspect to speak in an emotional tone of voice that matches the original event, the likelihood of the witness selecting the correct person is minimized. The goal of the current study was to expand on the literature with regard to matching and mismatching emotions, as well as how it relates to other variables.

Response Time

Compared to confidence, less research has investigated the relationship between target identification accuracy and the amount of time witnesses take to make an identification in a lineup. In an effort to identify other postdiction variables that predict accuracy, response time has been one method researchers have examined. There have been inconsistent results, with some

experiments indicating a positive relationship between accuracy and response time (Flowe, 2011; Sauer, Brewer, & Wells, 2008), whereas others have shown a negative relationship (Sporer, 1992).

With the development of eye-tracking devices, eyewitness researchers have started to use the technology to determine how long a participant looks at a face, and what features of the face they are focusing on. Using these eye-tracking devices, Flowe (2011) examined dwell time, which is the amount of time a witness spends looking at a face within a lineup. She had participants study 12 faces, and then presented them with 12 lineups, one for each studied face. Half of the participants were shown a simultaneous lineup, where all the faces were shown to participants at the same time. The other half of the participants were shown a sequential lineup, with faces being presented one at a time, and participants had to make a response after each image. There were three possible responses for this procedure that were analyzed separately, with each analysis comparing dwell time of simultaneous lineups to sequential lineups. When participants correctly identified the target, dwell time on the target face did not significantly differ between lineups. However, for the alternative faces, participants in the simultaneous lineup condition spent less time examining the face than in sequential lineups. This suggests that participants are conducting a more thorough examination of the faces when they are presented in a sequential format. This pattern was also observed when participants identified a face other than the target face. For both simultaneous lineups and sequential lineups, participants spent the most time examining the face they would eventually identify as the target. However, participants in the sequential lineup condition spent more time examining the other faces in the lineup, compared to the simultaneous lineup, again indicating that a more thorough examination was conducted in a sequential lineup. Finally, if participants opted to reject the entire lineup, dwell

times were longer in the sequential condition than in the simultaneous condition, supporting the conclusions in the other analyzes (Flowe, 2011)

While the possibility of studying dwell time has only recently been made possible by eye-tracking devices, another way to determine how long a witness spends examining a suspect within a lineup is by examining how long it takes for participants to make a decision for each face in a sequential lineup (response time). The key difference between the two types of measurements is that dwell time measures only the time spent looking at a face, not elsewhere on the screen, whereas response time only measures how long it takes for a participant to make a response to the face in the lineup. Sauer et al. (2008) conducted an eyewitness study that examined differences in response time for each face in a sequential lineup. To do this, they presented participants with a video of a simulated crime where a man steals a customer's credit card from a waitress. After 15 minutes, participants were shown an eight-person sequential lineup and asked to identify the thief, and then repeated the process to identify the waitress, recording how long participants spent on each face within the lineup. Sauer et al. (2008) found that participants spent a significant amount of time looking at the first face within both lineups. Then participants went through the rest of the lineup at a quick speed until they reached the face they identified as the target, where they again spent a significantly longer amount of time examining the face. For the waitress lineup, participants who correctly identified the waitress were faster at making their decision compared to participants who selected an incorrect face in the lineup. However, for the thief lineup, there was no significant difference in response time in terms of their accuracy in choosing the target face. Overall, Sauer et al. (2008) suggested that the reason for spending more time on the face that participants would eventually identify as the

target was due to the participants making sure that the image matched their memory of the target before making a response.

Contrary to the previous experiments, Sporer (1992) found a negative correlation between response time and target identification accuracy. Participants were brought into a lab to conduct a decoy task of rating faces for attractiveness and sympathy. During the procedure, a confederate walked into the lab insisting to the research assistant that he needed to take the projector being used in the decoy task. The interaction took place over approximately 20 seconds. Following the end of the experiment session, participants were asked to come back one week later as a follow-up to the decoy task. They were not informed about having to complete the lineup task until they returned to the lab for the second session. Participants recorded their pre-decision confidence, where they indicated how confident they were that they could identify the confederate in a lineup. A video lineup was presented next, with the beginning of the video showing all seven lineup members, which was then followed by a sequential display of each member's face. Participants were instructed to not make a judgment until they had viewed all the faces in the lineup. There were three possible responses participants could make, positively identifying a member in the lineup, rejecting the lineup, or indicating that they did not know but identifying the person they would choose if they were forced to. Finally, they indicated how confident they were that they had accurately identified the target (post-decision confidence). Results indicated that there was a positive correlation between accuracy and post-decision confidence (.45), indicating that the more confident a participant was after making an identification, the more accurate they were. Further, there was a negative correlation between accuracy and response time (-.36), in that participants were quicker to respond when they were correctly identifying the target. Finally, there was a negative correlation between response time

and post-decision confidence (-.55). This indicates that participants responded faster when they were more confident that their selection was accurate.

While dwell time is not possible with earwitness research, researchers in this field have not yet delved into response time. The majority of research does not indicate if the researchers allow participants to listen to an entire lineup more than once, or if they could listen to the individual voices multiple times. Amongst the studies that do provide this information, there are mixed methodologies. For example, Kerstholt and colleagues had their participants listen to the entire lineup once, before having them make a forced choice of whether or not the target was in the lineup (Kerstholt et al., 2006, 2004). In comparison, Öhman et al.'s (2011) participants listened to the entire lineup once, then listened to a shortened version of the lineup until they positively identified a target or rejected the lineup. Neither study discussed the reasoning or implications of their methodologies. To address the gap of information in earwitness literature, the current study examined response time by measuring how long it took participants to make a confidence response to each lineup member, a similar method to previous eyewitness research by Sauer et al. (2008).

Target Lineup Presence

A common method used in witness testimony research is to use two types of lineups; a target-present and a target-absent lineup. It is important to include both types of lineups in this area of research for ecological validity as well as witness testimony improvements (Wells & Turtle, 1986). For example, if police are attempting to identify the criminal, they may not have the correct suspect in custody, but rather an innocent person instead, which would result in a real-life situation of a target-absent lineup. Understanding the differences in target selection between the two types of lineups is necessary, since it is impossible to know which type of

lineup is being presented in a real-life line up situation. Additionally, some methods of improving witness testimony may only impact target-absent lineups, allowing for less misidentifications (Wells & Turtle, 1986). An example of this is using unbiased instructions when presenting a lineup to a witness (Malpass & Devine, 1981). Malpass and Devine (1981) tested the difference in target identification accuracy when using biased instructions or unbiased instructions. The study took place in a lecture hall with students from an introductory university course watching a demonstration. A male confederate came in and disrupted the lecturer by changing settings on an electronic device and eventually making a large commotion by pushing over the device and leaving the room. After being told that the man was a confederate, students were asked to volunteer to take part in a lineup study to identify the confederate. Participants were then randomly assigned to either a biased instructions condition or an unbiased instructions condition. In the biased instructions condition, participants were not informed that the confederate may not be in the lineup and the response form did not list an option to indicate that no one in the lineup was the target. In comparison, the participants in the unbiased instructions condition were explicitly told that it was possible that the confederate was not in the lineup and their response form had an option that allowed them to indicate the confederate was not present in the lineup. These researchers found that by using unbiased instructions, participants made significantly fewer errors in the target-absent lineup, while maintaining a high correct identification rate in the target-present condition. Overall, less misidentifications were made using unbiased instructions, which is a simple change to implement into real-life scenarios where it is unknown if the target is present or absent from the lineup.

Since previous research has indicated that there may be differences between target-present and target-absent lineups, the current research examined how voice frequency,

content, and emotion impacted both types of lineups. Experiment 2 utilized a target-absent lineup. Differences in pitch between members of the lineup can be more properly analysed with this method, as the presence of the target is not influencing confidence ratings. In other words, confidence ratings would not be skewed towards the target voice and suppressing the alternative voices. Experiment 3 did include the target in the earwitness lineup. This would allow for differences in pitch and emotion between the original conversation and the lineup to be investigated in relation to the target voice. Specific pitches and matching emotions may make it easier for participants to recognize the target voice and respond with higher confidence ratings.

The Current Research

The goal of this project was to expand the earwitness literature in the areas described previously. Specifically, it examined how the variables of voice frequency, content, and emotion impacted the confidence ratings of a witness' selection of a target. As discussed, while there does not appear to be a strong correlation between target selection accuracy and confidence, it is likely that participants will make their selection based on who they feel most confidence about being the target. Thus, the current study measured how confident a participant was that the voice in the auditory recording they were listening to was the voice they originally heard in the target event. In addition, by using this dependent measure, a more detailed analysis of participants' decisions was able to be conducted, as participants responded using an 11-point confidence scale rather than a two-option accuracy measure. This measure allowed for a within-subject analysis, to see how pitch influences confidence for each voice within the lineup, which would not be possible using an accuracy method where only one voice could be selected. In addition to confidence ratings, response time to each voice in the lineup was also examined, a novel analysis to earwitness research. If there is a relationship between response time and confidence, it may

eventually be an additional tool that experts could use to determine the legitimacy of an earwitness report.

The stimuli used in the current project are also novel in the field of earwitness research. Rather than having individuals record their voice and then altering the pitch through the use of a computer program, the stimuli used in this project were created by having professional actors record their voices at three different pitches, their normal pitch, and then in a lower pitch and a higher pitch. The purpose for creating the voices in this manner was to provide participants with more natural sounding alterations in pitch, while maintaining the same source of the voice. This study also incorporated a conversation as the original event, rather than hearing one voice in isolation. Previous earwitness work generally has participants listen to a single voice and then presents the participants with a voice lineup. There are a few exceptions to this. For example, Olsson et al. (1998) presented participants with a conversation among four people, each speaking an identical statement at different points throughout the conversation. Participants would then have to identify all four voices in separate lineups. The current study differs from this, in that there are three speakers, but only one target voice. Thus, participants are only required to identify one voice, although they are unaware of which voice they need to identify until they are presented with the lineup. Campos and Alonso-Quecuty (2006) also used a conversation as their original event, but the goal of their study was to see if earwitnesses remember the content of the message verbatim or just the gist of the message, rather than identifying voices. People generally do not hear voices in isolation of other voices, although one exception to this would be a phone call with no background noise. Thus, by having participants listening to multiple voices in one event, it better represents the experience of earwitnesses in real life.

The current project includes three experiments. The first experiment was a manipulation check of the content of the three conversations (see Appendices A through C) that would be used in Experiments 2 and 3. The manipulation check consisted of participants listening to the three conversation types (criminal, neutral, and vulnerable) and rating them on their aggressiveness. Since the hypotheses of Experiment 2 and 3 depended on reliable differences in semantic content, it was necessary to ensure that participants would interpret these conversations as being different from each other. It was hypothesized that the Criminal Content conversation would be the most aggressive, whereas the Vulnerable Content conversation would be the least aggressive.

Experiment 2 examined factors that influence ratings of confidence when the target is absent from the earwitness lineup. This allows for an analysis that is not related to the participant's memory of the target voice but rather the biases that occur when trying to identify a target voice in a lineup scenario. For this study, the primary analysis was a mixed-model ANOVA with the pitch level as a within-subjects variable and emotion and content as between-subjects variables. It was hypothesized that there would be a main effect of Lineup Emotion, in that participants would have higher confidence ratings for voices presented in an emotional tone of voice compared to a monotone voice. The original conversation recording was spoken in an emotional tone that was appropriate for the content of the conversation. Following the findings of Orchard and Yarmey (1995), it is expected that the match in emotional tone between the original conversation and voice lineup will facilitate the recognition of the target voice, thus increasing confidence ratings despite not having a target in the lineup. Additionally, it was hypothesized that there would be an interaction between Target Emotion and Conversation Content. The scripts for the Criminal Content and Vulnerable Content conditions were more emotional in nature. The criminal script involved the perpetrator threatening two individuals and

demanding money from them. The vulnerable script had a person looking for their lost cat who they have not seen in a day. The Neutral Content condition, in comparison, involved the target voice returning a dropped hat. Due to this difference in emotional content, it was hypothesized the having a match in emotion between the original conversation and the voice lineup would facilitate participants' recollection of the original voice to a greater extent in the Criminal and Vulnerable Content conditions, compared to the Neutral Content condition. In contrast, it was predicted that Conversation Content would not influence confidence ratings for the Monotone lineups.

Another hypothesized main effect was for Lineup Pitch. Specifically, it was hypothesized that participants would provide higher confidence ratings towards higher-pitched voices, compared to medium or low-pitched voices. Rodstrom and Neuhoff (2003) found that with single vowel sounds, participants were more accurate identifying the original voice at a higher frequency compared to lower frequencies. Furthermore, Stern et al.'s (2007) study indicated that when the original voice was of a higher frequency, participants remembered the voice as higher pitched than it actually was and were more likely to pick a distractor voice of a higher pitch. Between this previous research and the utilization of female voices in the current studies, it was predicted that participants would be more confident in high-pitched voices compared to medium- and low-pitched voices.

It was also predicted that there would be an interaction effect between Lineup Pitch and Conversation Content based on the frequency heuristic. This is due to evidence that males, who typically have lower-pitched voices, are more likely to be viewed as criminals, in real life and on television, and are more frequently admitted to adult correctional services (Allison et al., 2013; Parrott & Parrott, 2015; Reitano, 2017). In contrast, females, who normally have higher-pitched

voices, are usually portrayed as victims of crime, thus being more vulnerable compared to males (Parrott & Parrott, 2015). Based on these findings, it was predicted that participants would utilize a frequency heuristic when listening to the earwitness lineup. If participants are biased to believe that lower-pitched voices are more likely to be criminals, participants who are in the Criminal Content condition would have higher confidence ratings towards low-pitched voices, compared to medium- or high-pitched voices. The opposite would be true for participants in the Vulnerable Content condition, in that if they are biased to believe that higher-pitched voices are more likely to be victims, then they will have higher confidence ratings towards high-pitched voices.

Finally, it was hypothesized that the voices that participants rate with the highest degree of confidence would be the voices they spend the most time listening to. Previous research in eyewitness identifications has shown that participants will spend the longest amount of time on the face they will indicate as the target face from the original event (Flowe, 2011; Sauer et al., 2008). While Sporer (1992) found a negative relationship between confidence and accuracy, the methodology used in the current experiments more closely aligns with Sauer et al. (2008), where response time was recorded for each member of the lineup. As such, it was anticipated that the results found in the current study would replicate those of Sauer et al. (2008). Specifically, it was hypothesized that there would be a positive correlation between response time and confidence, where participants would be more confident in voices they spend more time examining.

Experiment 3 investigated lineups that included the target voice. This experiment was analyzed using a between-subjects ANOVA, with confidence ratings for the target voice as the dependent variable. While the analysis was different, it was still anticipated that the expected results from Experiment 2 would hold true for the target voice. Instead of examining all the voices in the lineup, the purpose of Experiment 3 was to identify the factors that influence target

identification. As such, it was expected that participants would provide the highest confidence ratings when the target voice was presented in a high pitch, compared to a medium or low pitch. It was also hypothesized that Target Pitch would interact with Conversation Content. Specifically, it was expected that participants would be the most confident in the target voice when it was presented in a low pitch in the Criminal Content condition, and in a high pitch in the Vulnerable Content condition. As with Experiment 2, a match in emotion was hypothesized to produce higher confidence ratings toward the target voice, compared to a monotone target voice. Lineup Emotion was also expected to interact with Conversation Content, with higher confidence ratings for the target voice when it was spoken emotionally in the Criminal and Vulnerable Content conditions. As for response time, it was expected that there would be a positive correlation between confidence and response time, as predicted in the target-absent experiment. Finally, a further analysis was conducted comparing the target voice to an alternative voice that matched in pitch and emotion (matching, non-target voice), although this was a post-hoc analysis, so no hypotheses were made prior to the experiment.

Experiment 1

The first experiment was conducted to ensure that there were semantic differences between the scripts used for the original conversations in Experiment 2 and 3. It was necessary that the content of these conversations differed in aggressiveness so that differences in content in the following experiments could be attributed to the material participants listened to. Participants in Experiment 1 listened to three conversations (criminal, neutral, vulnerable) and rated them on a Likert scale ranging from 1 to 9, with low scores indicating the target voice was vulnerable and high scores indicating the target voice was aggressive. It was anticipated that participants would

rate the Criminal Content conversation as the most aggressive and the Vulnerable Content conversation as the most vulnerable.

Method

Participants

Seventy-seven participants (53 females, M_{age} = 22.92) were recruited from the introductory psychology course at the University of Manitoba and were compensated with research credits required for their course. The final sample after data exclusion included 64 participants (45 females, M_{age} = 22.32).

Materials

The auditory stimuli used in this study were recorded by professional female actors who read three scripts (see Appendices A through C). Three actors recorded the lines of a scripted conversation. The actors spoke in their natural pitch, which would be considered a medium pitch for the purposes of Experiment 2 and 3, as well as spoke in an emotional tone of voice that was appropriate for the content of the conversation. For example, in the Criminal Condition, the actor recited the target voice's lines in an aggressive tone, whereas in the Vulnerable Condition, the actor sounded desperate. Each line of the script was recorded individually from each actor. The individual lines were then combined to form the recorded conversation between the three actors, reducing the amount of silence between each line to allow for a natural sounding conversation. The lines were combined in a manner that allowed for all possible combinations of roles to be fulfilled by each actor, creating six separate counterbalances (see Table 1). The purpose of this counterbalance was to eliminate the possibility that a particular actor's voice was distinctive, confounding the results of the study. Each conversation recording was approximately 60 seconds

in duration, with slight variations due to actors' articulation rate. Within each recorded conversation, the target voice, Person C, was introduced at approximately the 40 second mark, again with slight variations resulting from the articulation rate of the speakers. Additionally, the target voice spoke for 35-37 words, depending on the conversation content condition.

The recordings were then uploaded to Soundcloud (2008) and embedded into Qualtrics (2005). Using this method, the title of each clip would be presented to participants. As such, the files were named "Clip A", "Clip B", … and "Clip R", for a total of 18 clips. This naming system was used as to give no indication to the participant of what the content of the recording contained.

Procedure

Participants took part in this study through Qualtrics (2005) using their personal internet devices. After providing informed consent, participants were informed that they would be listening to a conversation between two people with a third person interrupting the conversation and that their task would be to rate how aggressive or vulnerable the third person was. Each participant listened to three conversation recordings, one for each Conversation Content condition. The order of the voices within the recording was counterbalanced between participants (Table 1). The three recordings were presented in a random order. The page following the instructions presented the first of the conversation recordings. Participants were asked to listen to the recording. To start the auditory clip, participants had to click on the embedded sound file. Below the sound file was a nine-point scale with 1 indicating *vulnerable* and 9 indicating *aggressive* with a slider that could be moved in either direction to indicate the level of vulnerability or aggressiveness. The slider on the scale started in the middle to prevent bias in either direction. Participants were required to stay on the page for the duration of the

conversation recording. Once the duration of the recording had passed, a button with an arrow appeared at the bottom, right-hand corner of the screen that participants would press to move on to the next screen. The following page presented a multiple-choice question that was related to the content of the conversation they just heard to ensure that participants were paying attention (see Appendix D). Participants were required to answer the question before they were allowed to progress in the study. If participants clicked the grey arrow button at the bottom, right-hand corner of the screen to progress before answering this question, the text of the question was highlighted in yellow and the statement "Please answer this question" appeared above the question. The next screen would then present another conversation recording of a different content condition. After responding to the three conversation recordings, participants were asked to provide their age and gender for demographic purposes, as well as asked whether they had any technical difficulties.

Results

Data Exclusion

Participants were excluded from data analysis for two reasons. First, if participants indicated that they had technical difficulties listening to the auditory recordings, then they were excluded as they were unable to listen to the recordings as intended. Two participants indicated that they had technical difficulties and were subsequently removed from data analysis. Second, if participants had answered any of the three multiple-choice content questions incorrectly, they were removed from the data analysis, as they had failed to successfully pass the attention checks within the study. Ten participants answered one or more of the multiple-choice questions incorrectly. One participant was removed for both reasons. This resulted in 13 participants being excluded prior to conducting statistical analyses.

Aggressiveness Ratings

Participants' aggressiveness ratings were submitted to a one-way within-subjects ANOVA, with Conversation Content being the independent variable. This analysis indicated that there was a significant difference in aggressiveness ratings between Conversation Content conditions, $F(2, 126) = 91.57$, $MSE = 295.42$, $p < .001$, $\eta_p^2 = .592$. Bonferroni comparisons revealed that the Criminal Content condition ($M = 6.91$) was significantly more aggressive compared to the Neutral Content condition ($M = 3.55$, $p < .001$) and the Vulnerable Content condition ($M = 2.91$, $p < .001$). Additionally, the Neutral Content condition had higher aggressiveness ratings compared to the Vulnerable Content condition, $p = .029$.

Discussion

As predicted, there were semantic differences between the different conversation recordings. Participants provided higher aggressiveness ratings to the Criminal Content recording, suggesting that they felt the target voice was more aggressive when asking the two other voices to give them their money. Furthermore, participants did indicate that the target voice in the Vulnerable Content recording was vulnerable. These results indicate that the manipulation of Conversation Content was being interpreted as intended. As such, these conversation recordings were utilized for the Study Phase in Experiments 2 and 3.

Experiment 2

Experiment 2 examined factors that influence earwitness identifications in target-absent lineups. In these lineups, the target voice from the original conversation was not present, allowing for a more thorough examination of the influence of differences in pitch in participants' confidence ratings. When the target was present in the lineup, provided that participants are able

to identify who the target is, confidence ratings would be skewed towards the target, suppressing confidence ratings for the alternative voices in the lineup. By having a target-absent lineup, an analysis could be conducted without the influence of a matching target voice, isolating the biasing effects of pitch, content, and emotion. Additionally, real-life scenarios may not have the correct suspect in the lineup. Knowing how the variables in this study influence witnesses' confidence may help with the process of developing lineups that are fair towards an innocent suspect. If witnesses are biased to select certain voices based on pitch, then lineups can be adjusted to ensure that an innocent suspect is not more likely to be selected compared to the other voices in the earwitness lineup. These biases may also encourage earwitnesses to incorrectly select a voice from a lineup instead of rejecting the lineup, as they may feel confident due to characteristics of the voice rather than hearing the voice previously.

Method

Participants

A total of 252 participants were recruited from the Introduction to Psychology Participant Pool at the University of Manitoba (173 females, M_{age} = 19.83). Participants were required to have normal or corrected-to-normal hearing. Following completion of the study, 52 participants were excluded from data analysis due to pre-determined exclusion criteria described in the results section. The final sample consisted of 200 participants (140 females, M_{age} = 19.89). All participants received research credits for their introductory psychology course as compensation.

Materials

The auditory stimuli used in the Study Phase of Experiment 2 was identical to that of Experiment 1. The second set of auditory stimuli for the Test Phase was recorded by nine female

actors, including the three actors that voiced the stimuli for the Study Phase. Each of these actors recorded auditory files containing the lines of Person C in each of the scripts. Each actor recorded multiple versions of these lines, altering their voice naturally through pitch (high, medium, low) and emotion (emotional, monotone). Importantly, the changes in pitch were produced by the actors, which made the low and high frequency auditory recordings sound more natural than if they had been altered electronically, ensuring ecological validity of the earwitness lineups. The auditory clips were then analyzed for their fundamental frequency, to ensure that there were differences in pitch. The average frequency for high-pitched voices was 251.57 Hz; medium-pitched voices were 209.42 Hz; and low-pitched voices were 194.50 Hz. The emotional recordings were spoken in an emotion that matched the content, as was done for the conversation recordings. Each auditory clip contained the entire set of lines spoken by Person C, with pauses between each line. The duration of each auditory clip ranged from 10 to 25 seconds and had an identical word count to the target voice in the original conversation. The variation in duration was largely due to differences between actors in articulation rate and differences in emotion, with monotone recordings having a longer duration.

In between the Study and Test Phase, a filler task of simple arithmetic questions was administered. All questions were comprised of single digits terms, with two or three terms per question. The operations used in the questions were addition, subtraction, multiplication, division, and exponentiation by 2. The solutions to these questions were always between 1 and 50. For example, "$8 + 7 + (5)^2 = 40$" and "$3 + 7 – 4 = 6$". Answers were inputted in a textbox located below the question. The experiment was programmed to present 10 questions on a page and participants were required to answer all questions correctly before proceeding to the next set of questions. Additionally, the experiment presented questions until the total duration of the

study was greater than 35 minutes. This total duration included both the study phase and the filler task. The filler task was intended to last for approximately 30 minutes, allocating five minutes for the study phase. This method was employed to reduce variability due to participants' math ability, allowing for the filler task to last for approximately that same duration across all participants. One disadvantage of this method was that some participants would load the survey on their device and come back to it an hour later, and thus, would not be presented with any arithmetic questions. These participants were excluded from data analysis, as they experienced a different procedure than the majority of participants.

Participants completed the study on their personal internet devices using the online program Qualtrics (Qualtrics, 2005). Participants were informed when they were recruited for the experiment that they would be required to use headphones during the experiment. All auditory files were uploaded privately to an external website and then embedded into the Qualtrics survey as required. As with Experiment 1, the auditory clips used in the Study Phase were labelled as "Clip A" and continued alphabetically, for a total of 18 recordings. For the Test Phase, the files were labelled with an uppercase and lowercase letter, such as "Voice Df" or "Voice Hg". The letter combination was coded so that the researcher knew which voice was in the clip, the content of the message, the pitch of the voice, and if it was emotional or monotone. This was implemented to prevent the participant from knowing the features of the voice before they listened to the recording.

Variables

Conversation Content. There were three types of conversations that a participant could have been presented with, Criminal (Appendix A), Neutral (Appendix B), and Vulnerable (Appendix C). The Criminal Content Condition involved a person threatening two people and

demanding money; the Neutral Content Condition had a person giving another person a hat that they dropped; and the Vulnerable Content Condition was a person looking for their lost cat. The content was consistent throughout both phases of the study, in that if a participant heard a criminal conversation during the Study Phase, then the auditory clips they were presented with during the Test Phase were of the lines of Person C in the criminal script.

Lineup Emotion. An emotional tone of voice was used for the entire script of the conversation recordings used during the Study Phase. The Test Phase consisted of either a matching Emotional Lineup or a mismatching Monotone Lineup. All auditory recordings in the lineup were spoken with the same emotional style of voice.

Lineup Pitch. All actors used their natural pitch when recording the Study Phase conversation recordings, which was labelled as their medium-pitched voice. During the Test Phase, participants were presented with a lineup that consisted of six voices, one of which was the target voice from the original conversation recording in the Study Phase. Two voices for each pitch type (high, medium, low) were included in the lineup. The voices were counterbalanced between participants so that each of the six voices was presented in all three pitch types. Same-pitch voices were averaged, giving each participant three confidence rating scores for high, medium, and low-pitch voices.

Confidence. Participants were asked how confident they were that the voice being presented to them was the target voice in the original conversation. Confidence ratings were given on an 11-point scale (0 to 10) and were collected for each of the six voices within the lineup.

Response Time. Response Time measurements were collected during the Test Phase of the study for each voice within the lineup. The voices were presented individually on separate web pages within the experiment, and the amount of time the participant spent on the page was recorded. A total of six response times were recorded for each participant, one for each voice within the lineup.

Procedure

Study Phase. Participants were provided with a link to the Qualtrics survey that they could access with any electronic device. After agreeing to the informed consent, participants were presented with instructions stating that they would be listening to a simulated conversation and that they should listen to this auditory clip one time and then move on to the next screen. At this point, participants were reminded to connect headphones to their device. The following screen had the statement "Please listen to this auditory clip." Directly below this statement, the embedded sound file was displayed that listed the title of the clip (e.g. "Clip A") and had a play button that participants clicked on to play the recording. To move on to the next page, participants clicked on a square, grey button labelled with a right arrow at the bottom, right-hand corner of the screen. The program recorded in seconds how long the participant stayed on the page. This provided a method of determining whether the participant listened to the entire conversation clip, or the possibility of listening to the recording more than once. On the next screen, a multiple-choice question was presented to determine if the participant had paid attention to the content of the conversation (Appendix D).

Filler Task. Following the Study Phase, participants were given the arithmetic filler task. Instructions informed participants that they would be answering math questions for approximately 30 minutes and that they were required to answer all questions on the page

correctly before they could move on to the next page. Two example questions were provided that demonstrated all of the arithmetic operations that would be used throughout the task. Participants clicked on the arrow button at the bottom, right-hand corner of the screen to move on to the next page where the arithmetic questions began. Each page of arithmetic questions listed 10 questions, with a textbox for participants to enter their response. When participants finished a page, they would click on a square, grey arrow button at the bottom, right-hand corner of the screen to move on to the next page. If they incorrectly answered any questions, the incorrect arithmetic question and the corresponding textbox would be highlighted in yellow and present the statement "Incorrect Answer. Please input the correct number." After the total duration of the study reached 35 minutes, the next time participants finished a set of questions and clicked on the arrow button, they would be presented with the test phase, rather than another set of arithmetic questions.

Test Phase. At the beginning of the Test Phase, participants were instructed that they would need to identify the person who interrupted the conversation they had previously heard by listening to several voice clips. Participants were again reminded to connect headphones. Participants were then presented with a sequential lineup of voices, using the auditory clips of Person C from the script. In this experiment, participants were presented with six voices. None of these voices were the target voice or either of the other voices that spoke in the original conversation. The screen presented an auditory clip of a single voice, as well as a confidence question. Once the participant had listened to the auditory clip, they responded to the question "how confident are you that this is the voice you heard?" Participants responded using an 11-point scale, with 0 indicating this voice was *definitely not* the target, and 10 indicating that this voice was *definitely is* the target. A sliding scale was used, with the marker starting at 0.

Participants would click and drag the marker to the point on the scale that represented their confidence rating. The scale was notched so that participants could only respond using whole numbers. Participants were required to move the marker before they could move on to the next voice. Additionally, the arrow button that allowed participants to proceed to the following page did not appear until the duration of the voice clip had finished. For example, if the voice clip was 12 seconds, then the arrow button did not appear until 12 seconds had passed. Participants completed this process six times, once for each voice in the lineup. In addition to the confidence ratings, the survey recorded the amount of time a participant spent on a single screen. This allowed for Response Time measurement, or how long the participant required to make a response. An exclusionary criterion was based on this measurement as well, such that participants who spent more than five minutes on the page were excluded from data analysis. This cut-off was implemented to ensure that participants who were included in the analysis would be on task and had listened to the voice. If a participant required more than five minutes on the page, it was possible that the participant had either stopped participating in the experiment, or that they listened to the voice recording multiple times within the time limit. Once the participants finished the six-voice lineup, they were asked about their age and gender for demographic purposes. Additionally, they were asked if they had any technical difficulties listening to the auditory recordings throughout the study. This was implemented through a multiple-choice question with the options of yes or no. Participants were required to answer this question before they could proceed to the end of the study where they automatically received their research credits for compensation.

Results

Data Exclusion

Prior to running any data analyses, data were subjected to pre-determined exclusion criteria. First, any participants who had indicated that they had technical difficulties listening to the auditory recordings were removed. Participants may not have heard the voices they were supposed to identify, thus any confidence ratings they may have made would be uninformative. In this experiment, two participants were excluded for this reason. Second, due to the way that the filler task was implemented, two participants had to be excluded, as they were not presented with arithmetic questions. The reason for this is that the filler task was programmed to show arithmetic questions until the total duration of the study reached 35 minutes. The participants excluded for this reason would have opened the experiment on their device but did not participate until after 35 minutes had passed. These participants would have heard the original conversation and then were immediately presented with the lineup of voices. Previous research, such as Clifford (1980), has shown that retention interval can influence earwitness testimonies, so participants who did not experience the filler task were excluded to maintain the same retention interval between all analyzed participants. A third exclusion criterion was based on the multiple-choice question that inquired about the content of the conversation. Participants who answered this question incorrectly were excluded from data analysis, as it was unlikely that they were paying attention to the conversation, or they may not have listened to the entire conversation as the question referred to a statement from Person C. There were 25 participants who incorrectly answered the question in this study. The fourth exclusion criterion was the duration of time participants spent on the target conversation page. Participants were instructed to listen to the conversation once and then move on to the next page. Since the target

conversation was approximately 60 seconds in duration, participants were expected to stay on the page between 60 and 160 seconds. This range accounted for participants who had slower internet speeds to load the auditory clip, as well as those who did not start the recording immediately due to reading the instructions or having to connect headphones. However, participants who stayed on the page for longer than 160 seconds may have listened to the auditory clip more than once or left the experiment for an extended period of time. Thus, it was decided to remove these participants from data analysis. Any participant who left the page in less than 60 seconds was also excluded, as they did not listen to the entire recording and would not have an equal amount of exposure to the target voice as the other participants. Eleven participants were excluded for this reason. Finally, participants were also excluded if during the Test Phase, they remained on one voice for more than five minutes. While response time is a variable being examined, participants who spent more than five minutes on one voice were considered to no longer be paying attention to the experiment. With this cut-off, participants would still be able to listen to a voice multiple times before making their confidence response. Four participants were excluded due to listening to a lineup voice for over five minutes. Finally, eight participants were excluded for a combination of the above criteria. Four participants did not listen to the target conversation for the required amount of time and incorrectly answered the multiple-choice question. Two participants indicated they had technical difficulties and answered the multiple-choice question incorrectly. One participant did not meet the time limit for the target conversation and indicated they had technical difficulties. One participant also did not listen to the target conversation for the required time and was not presented the filler task. In total, 52 participants were removed prior to data analysis.

Confidence Ratings

Across participants and voices, the average confidence rating was low, at 2.53, indicating that generally, participants were not confident that any of the voices in the lineup were the target voice. Each participant had six confidence ratings, one for each voice in the lineup. Three mean confidence ratings were calculated for each participant corresponding to each level of pitch. These ratings were submitted to a 3x(3x2) mixed-model ANOVA, treating Lineup Pitch (high, medium, low) as a within-subject variable and Conversation Content (criminal, neutral, vulnerable) and Target Emotion (emotional, monotone) as between-subject variables. The mean and mean square error for each cell is displayed in Table 2.

The analysis revealed a significant main effect of Lineup Pitch, $F(2, 388) = 8.11$, $MSE = 32.49$, $p < .001$, $\eta_p^2 = .040$. Bonferroni pairwise comparisons found that participants were more confident in medium-pitched voices ($M = 3.00$) than high-pitched ($M = 2.30$, $p = .004$) and low-pitched voices ($M = 2.30$, $p = .003$). There was no significant difference between high-pitched and low-pitched voices ($p = 1$). There was also a significant main effect of Conversation Content, $F(2, 194) = 5.22$, $MSE = 48.33$, $p = .006$, $\eta_p^2 = .051$. Confidence ratings were significantly higher for the Neutral Content condition ($M = 3.09$) compared to the Criminal Content condition ($M = 2.16$, $p = .007$) and the Vulnerable Content condition ($M = 2.35$, $p = .048$). Additionally, the Criminal Content and Vulnerable Content conditions did not differ from each other ($p = 1$). There was no significant main effect for Target Emotion, $F(1, 194)$ p= .221, $MSE = 2.04$, $p = .639$, $\eta_p^2 = .001$.

The analysis also revealed a significant interaction effect between Lineup Pitch and Conversation Content, $F(4, 388) = 3.17$, $MSE = 12.72$, $p = .014$, $\eta_p^2 = .032$. Participants in the Criminal Content condition were significantly less confident in high-pitched voices ($M = 1.45$) compared to medium-pitched ($M = 2.70$, $p = .002$) and low-pitched voices ($M = 2.33$, $p = .016$).

For participants in the Vulnerable Content condition, confidence ratings were lower for low-pitched voices (M = 1.68) compared to medium-pitched (M = 2.90, p = .003) and high-pitched voices (M = 2.47, p = .037). There were no significant differences between pitches in the Neutral Content condition. These results are displayed in Figure 1. No other interactions were statistically significant.

Response Times

This experiment was designed to examine the relationship between confidence ratings and response times. Since the voice recordings differed in their duration, response times were established by taking the duration of time spent on the page and subtracting the duration of the recording of the voice. Since participants were unable to leave the page until the lineup voice clip had finished, all participants would have listened to the voice at least one time and this measurement is indicative of the time spent re-examining the voice and making a decision about their confidence levels.

Pearson correlation analyses were conducted between confidence ratings and response times. The mean confidence rating across all voices was 2.53. The mean response time was 12.37 seconds. A partial correlation controlling for participant indicated that there was a significant negative correlation, $r(1197) = -.08$, $p = .006$. In other words, participants spent less time on voices that they were more confident about. Following this, a Pearson correlation was conducted for each presentation order within the lineup. Results are shown in Table 3. The first voice presented in the lineup had a positive correlation between response time and confidence (.18). In comparison, the voices that were presented second through sixth in the lineup had negative correlations. As such, two separate partial correlations were conducted, one for the first

order voice and one for the remaining voices. These results indicated that there was a significant positive correlation between confidence and response time for the first voice participants listened to, $r(197) = .18$, $p = .010$. This means that participants spent more time examining the first voice in the lineup when they had higher confidence that it was the target voice. There was also a significant negative correlation for the second through sixth voice order, $r(997) = -.12$, $p < .001$. That is, for the remaining voices in the lineup, participants would spend less time examining a voice when they were more confident that it was the target voice.

Discussion

Confidence ratings across all voices in the lineup were low, averaging at 2.53. While the full range of confidence ratings were provided, participants generally were not confident that the voices in the lineup matched the target voice, a correct assessment for a target-absent lineup. However, the data analysis did reveal an effect based on the variables manipulated. For Lineup Pitch, participants had the highest confidence ratings towards the medium-pitched voices, compared to the high- and low-pitched voices. This is contrary to the original hypothesis, where it was predicted that high-pitched voices would have the highest confidence ratings. This hypothesis was based on the results of Rodstrom and Neuhoff (2003), who observed that participants were more accurate in identifying high frequency voices and Stern et al. (2007), who observed that participants more often chose a higher-pitched distractor voice when the original voice was high pitched. However, the current study's results indicated that instead, a matching effect may be occurring. The original conversation took place in what was considered the actor's medium-pitch voice. Therefore, it follows that the closest matching voice in a target-absent lineup would be the medium-pitched voices. Participants appear to have been most confident in

the voices that were of equivalent pitch due to the similarity, rather than being most confident in the high-pitch voice as the previous voice identification research indicated.

A main effect of Conversation Content was also revealed in this experiment, although it was not anticipated. Participants provided the highest confidence ratings for the Neutral Content condition. One possible explanation for this is that in the Neutral Content condition, there was less at stake for identifying the target. In other words, participants perceived less negative consequences for incorrectly identifying a target when the target picked up a dropped hat, whereas wrongly identifying a criminal could send an innocent person to jail. Similarly, victims may not want to be identified, especially if it has the potential of further harm, and so participants may have been less willing to pick out the victim in a lineup. However, in the current study, participants are only providing their confidence in selecting a voice rather than asking them to identify a target with a yes or no statement. There should be less risk in commenting on their confidence compared to physically picking someone from a lineup and identifying them as the target. Furthermore, the victim in the Vulnerable Content condition in the current study is missing their cat, rather than being a victim of a crime, so identifying the target should be of little risk to the victim. It is recommended that this effect should be replicated before exploring this difference in more detail.

The results for the Target Absent study did reveal an interaction effect between Lineup Pitch and Conversation Content. The original hypothesis stated that participants would be most confident in the low-pitched voices in the Criminal Content condition and in the high-pitched voices in the Vulnerable Condition. Instead, this study found that participants were significantly less confident in certain voice pitches depending on the content of the original conversation. Specifically, participants in the Criminal Content condition were least confident in the

high-pitched voices, and participants in the Vulnerable Content condition were least confident in the low-pitched voices. While the hypothesis referred to the conditions with the highest confidence ratings, the results are consistent with what conditions would have the lowest confidence ratings. These results support the idea that there is a frequency heuristic that is influencing confidence levels of earwitnesses. It was reasoned that earwitnesses may be more confident in a criminal, low-pitched voice since it is typically men who are portrayed as criminals. As such, earwitnesses would be less confident in high-pitched voices, which is supported by the results in this experiment. Likewise, since women are more often seen as victims, earwitnesses may be more confident in high-pitched voices, and less confident in low-pitched voices, as seen in the Vulnerable Content condition of this experiment. One possibility of why there were no significant differences between the content predicted pitch and the medium-pitch voices is because the original conversation presented the target voice in a medium pitch. Even though none of the voices in the lineup were the target voice, there was a match in pitch, making them seem more similar to the target voice compared to the low or high-pitched voices. Thus, in the Criminal Content condition, participants were equally confident in the medium-pitched voice, which matched the original conversation, and the low-pitched voice, where confidence levels were influenced by a frequency heuristic. The opposite would be true for the Vulnerable Content condition, in that participants were just as confident in the matching-pitch voice as they were in the high-pitched voice.

Contrary to the original hypothesis, there was no significant difference between Emotional Lineups and Monotone Lineups. It was expected that the match in emotion between the original conversation and the lineup would facilitate participants' recollection of the target voice and thus produce higher confidence ratings. It appears that a match in emotion between the

original target voice and non-target voices does not increase confidence ratings. It may be that even if the emotion matches the original event, if an earwitness does not recognize the target voice, the presence of emotion may not increase confidence towards other voices. This would indicate that emotion is utilized in a different manner than voice pitch, as the match in voice pitch did increase confidence ratings. A match in emotion may be more advantageous when the target is present in the lineup, as it would more closely resemble the original conversation. There was also no interaction effect between Lineup Emotion and Conversation. Confidence ratings did not differ depending on the Conversation Content when voices were presented in an emotional tone. It appears that a match in emotion is not more beneficial for conversations that are emotional in nature, compared to a neutral conversation. It was possible that the match in emotion would be more beneficial when the target voice was implemented into the lineup and this was examined in Experiment 3.

In regard to the response time results, the correlational analysis revealed a negative correlation when examining all voices in the lineup. In other words, the more confident a participant was that the voice was the target voice, the less time they spent reviewing the voice. This contradicts the original hypothesis of a positive correlation between response time and confidence. This was based on the sequential lineup research conducted by Sauer et al. (2008) that utilized a similar methodology to the one used in the current experiment. However, there is previous research that support these results (Sporer, 1992, 1993). In Sporer's (1992) research, as described in the introduction, a negative correlation was found between eyewitness confidence and response time (r = -.55). His results were replicated in a later study (r = -.33) using a methodology where participants responded after each member of the lineup rather than at the end of the lineup, which more closely matches the methodology of the current experiment (Sporer,

1993). The earwitness results of the current study also showed a significant negative correlation (r = -.08), although not as strong as the ones found in Sporer's research.

One thing to note is that the first voice in the lineup had a positive correlation between confidence and response time, whereas the remaining voices maintained the negative correlation. This may be due to the methodology used, where participants needed to make a confidence judgment before moving on to the next voice and were not allowed to revisit a voice. Therefore, since it is the first voice that participants were encountering, it is likely that participants were taking longer to examine the first voice when making a confidence judgment while they were getting used to the methodology and had no other voice to compare it to. This occurred in Sauer et al.'s (2008) research, where participants spent more time on the first face in the lineup compared to the rest of the faces, with the exception of the person they would positively identify as the target. As such, if using response time to judge participants' confidence or accuracy, the order the voice appeared in within the lineup should be taken into account.

Experiment 3

The second experiment was conducted without the target included in the lineup, examining whether biases existed that influenced participants' confidence ratings towards certain voices. Experiment 3 presented participants with a target-present lineup, which allowed for a direct test of participants' memory of the target voice in conjunction with biases that make it either easier or more difficult to identify the matching voice. This was done by comparing the pitch and the emotion of the target voice in the lineup between subjects. In addition, a within-subjects analysis was conducted comparing the target voice to a matching, non-target voice, where the pitch and the emotion were identical for both voices.

Method

Participants

Undergraduate students in the Introduction to Psychology course at the University of Manitoba were recruited to participate in this study. There were 611 participants who completed the study and received research credits towards their introductory psychology course (447 females, M_{age} = 19.52). Following data exclusion, there was a final sample of 471 participants (349 females, M_{age} = 19.69).

Materials

The materials used in Experiment 3 are identical to those used in Experiment 2.

Variables

Experiment 3 examined the same variables as Experiment 2, although pitch was changed to a between-subjects variable. While the lineup consisted of six voices across three pitch types, the main focus of this experiment was on the target voice, thus the variable used is Target Pitch, which could be high, medium, or low. The target voice spoke in a medium-pitched voice in the original conversation. The other voices in the lineup did vary in pitch, in that there was always one alternative voice that matched the pitch of the target voice, and then two voices for each of the other two pitches.

Procedure

The study phase and the filler task were identical to that of Experiment 2. For the test phase, the target voice was included in the lineup. In other words, the actor who spoke Person C's lines from the script appeared in the earwitness lineup. Participants were also presented with

the five alternative voices that were also used in Experiment 2. All voices were presented in either an emotional or monotone voice and matched the content of the original conversation.

Results

Data Exclusion

The exclusion criteria used in Experiment 2 were also used in the current experiment. In total, 140 participants were excluded from data analysis. Thirteen participants were excluded due to having technical difficulties, nine for not receiving the filler task, 39 for incorrectly answering the multiple-choice question, 27 for not meeting the time requirements of the target conversation, and 11 for exceeding five minutes for at least one voice in the lineup. There were also 41 participants who were excluded for a combination of the described exclusion criteria. Nineteen participants did not listen to the target conversation for the required duration and incorrectly answered the multiple-choice question. Eleven participants were excluded for those two reasons as well as indicated that they had technical difficulties. There were three participants who incorrectly answered the multiple-choice question and indicated they had technical difficulties. Three participants were excluded due to the target conversation time restrictions and technical difficulties. Two participants were excluded due to the target conversation time restrictions and not being presented with the filler task. Two participants exceeded the time limits for both the target conversation and the lineup voices. Finally, one participant was excluded because they did not meet the time requirements for the target conversation, incorrectly answered the multiple-choice question, and did not receive the filler task.

Confidence Ratings

The first analysis for this experiment was conducted on the confidence ratings for each of the voices in the lineup, including the target voice. In this analysis, the alternative voices are labeled as "Alternative Voice A" to "Alternative Voice E". Each of these voices was portrayed by a different actor, for a total of five actors who were not included in the original conversation. The average confidence rating for the target voice was 4.25, although the full range of responses were received. The average confidence rating for each voice in the lineup is displayed in Figure 2. A within-subjects ANOVA was conducted on the confidence ratings of each voice in the lineup, resulting in a significant difference among the voices, $F(5, 2350) = 60.13$, $MSE = 453.56$, $p < .001$, $\eta_p^2 = .113$. Participants' confidence ratings for the target voice were significantly higher compared to all other voices, $p < .001$. Additionally, Alternative Voice C had significantly lower confidence ratings than all other voices in the lineup, $p < .001$. This suggests that participants were less likely to believe that this voice was the target voice from the original conversation. Finally, Alternative Voice D and E were significantly different from each other, with the Alternative Voice D having higher confidence ratings, $p = .002$.

Confidence ratings of the target voice were then submitted to a 3x3x2 between-subjects ANOVA, Target Pitch (high, medium, low), Conversation Content (criminal, neutral, vulnerable), and Lineup Emotion (emotional, monotone). The mean and standard error for each cell is presented in Table 4. This analysis revealed a significant main effect of Target Pitch, $F(2, 453) = 3.79$, $MSE = 48.59$, $p = .023$, $\eta_p^2 = .016$. Bonferroni comparisons indicated that confidence ratings for high-pitched voices ($M = 3.68$) were significantly lower than confidence ratings for medium-pitched voices ($M = 4.78$), $p = .018$. Confidence ratings for low-pitched voices ($M = 4.26$) were not significantly different from the other pitches. The interaction between

Target Pitch and Conversation Content was not significant, $F(4, 453) = .163$, $MSE = 2.09$, $p = .957$, $\eta_p^2 = .001$.

There was also a significant effect of Lineup Emotion, $F(1, 453) = 5.64$, $MSE = 72.23$, $p = .018$, $\eta_p^2 = .012$. Confidence ratings for Emotional Lineups were higher than for Monotone Lineups ($M = 4.64$ vs. $M = 3.85$). The interaction between Lineup Emotion and Conversation Content was not significant, $F(2, 453) = .242$, $MSE = 3.10$, $p = .785$, $\eta_p^2 = .001$. No other significant main effects or interaction effects were found from the between-subjects ANOVA.

The current experiment was set up so that there was always a non-target voice that matched in pitch and emotion. After observing the differences in confidence ratings in Table 4, a post-hoc analysis was run to ensure that participants were able to distinguish between the target voice and the matching non-target voice. To investigate this further, a 2x(3x3x2) mixed ANOVA with Voice (target, non-target) as a within-subjects variable and Conversation Content, Target Pitch, and Lineup Emotion as between-subjects variables. This resulted in a main effect of Voice, $F(1, 453) = 97.58$, $MSE = 969.14$, $p < .001$, $\eta_p^2 = .177$. Confidence ratings were higher for the target voice ($M = 4.24$) compared to the matching, non-target voice ($M = 2.21$). This reflects the results discussed at the beginning of this section, where participants were able to distinguish the target voice from the other voices within the lineup. Additionally, there was a significant 4-way interaction revealed, $F(4, 453) = 3.108$, $MSE = 30.87$, $p = .015$, $\eta_p^2 = .027$. Figure 3 displays the results of this interaction.

To further examine this interaction, a 2x(3x2) ANOVA was conducted on each level of Conversation Content. A significant main effect of Voice was maintained across all levels of Conversation Content. In the Criminal Content condition, the target voice had higher confidence ratings ($M = 4.60$) than the matching, non-target voice ($M = 1.99$), $F(1, 159) = 57.17$, $MSE =$

560.37, $p < .001$, $\eta_p^2 = .264$. The target voice in the Neutral Content condition also had higher confidence ratings ($M = 3.84$) than the matching, non-target voice ($M = 1.91$), $F(1, 154) = 34.30$, $MSE = 295.97$, $p < .001$, $\eta_p^2 = .182$. The same results were observed in the Vulnerable Content condition, with the target voice ($M = 4.29$) having higher confidence ratings than the matching, non-target voice ($M = 2.73$), $F(1, 140) = 15.53$, $MSE = 178.82$, $p < .001$, $\eta_p^2 = .100$.

There was a 3-way interaction between Voice, Target Pitch, and Lineup Emotion in the Criminal Content condition, $F(2, 159) = 57.17$, $MSE = 107.97$, $p = .005$, $\eta_p^2 = .065$. This interaction was not significant in the Neutral or Vulnerable Content conditions ($p > .05$). The 3-way interaction in the Criminal Content condition was then broken down by Lineup Emotion. There was a significant main effect of Voice in the Emotional condition, $F(1, 81) = 31.52$, $MSE = 365.96$, $p < .001$, $\eta_p^2 = .280$, with higher confidence ratings for the target voice ($M = 5.15$) compared to the matching, non-target voice ($M = 2.19$). A main effect of Voice was also found in the Monotone condition, $F(1, 78) = 26.15$, $MSE = 207.13$, $p < .001$, $\eta_p^2 = .251$, again with higher confidence ratings for the target voice ($M = 4.05$) than for the matching, non-target voice ($M = 1.78$). The interaction between Voice and Target Pitch was significant in the Emotional condition, $F(1, 81) = 3.15$, $MSE = 365.96$, $p = .048$, $\eta_p^2 = .072$, but it was not significant in the Monotone condition, $F(1, 78) = 2.419$, $MSE = 19.16$, $p = .096$, $\eta_p^2 = .058$. All other interactions were nonsignificant.

The interaction was decomposed by examining the 2-way interaction in the Criminal Content, Emotional Lineup condition. Participants gave significant higher confidence ratings for the target voice compared to the matching, non-target voice in the High Pitch condition ($M = 4.00$ vs. $M = 1.79$, $p = .012$) and in the Medium Pitch condition ($M = 6.62$ vs. $M = 1.83$, $p <$

.001). There was no significant difference in confidence ratings in the Low Pitch condition (M = 4.82 vs. M = 2.96, p = .084).

Response Time

As with Experiment 2, response times recorded in the experiment were subtracted by the duration of the voice clip to account for differences in durations. As such, the remaining response time represented the amount of time the participant spent re-examining the voice in the lineup after the first time they listened to it.

A partial correlation was conducted between confidence and response time, while controlling for participant, as each participant had six confidence ratings and corresponding response times. Across all participants and voices, the average confidence rating was 2.47 and the average response time was 12.46 seconds. The correlation analysis revealed a significant negative correlation between confidence ratings and response time, $r(2787) = -.06$, $p = .003$. In other words, the more confident a participant was that the voice they heard was the target voice, the less time they spent on that page of the survey. An analysis was also conducted on the order of the voices. Six participants were excluded from this specific analysis due to the order of voices not being properly recorded during the experiment. This analysis showed a pattern of a positive correlation for the first voice and a negative correlation for the second through sixth voice (Table 3). As with Experiment 2, two partial correlations were conducted, controlling for participant. The first analysis examined the voice that appeared first in the lineup, which revealed a nonsignificant correlation $r(462) = .05$, $p = .287$. Unlike Experiment 2, there was no relationship between response time and confidence for the first voice participants listened to. The second partial correlation was conducted on the second through sixth presented voices. This

revealed a significant negative correlation, $r(2322) = -.08$, $p < .001$. For these voices, participants who were more confident that the voice was the target from the original conversation took less time to respond.

To see if participants responded quicker to the target voice, a partial correlation was conducted between confidence and response time for each voice within the lineup, controlling for participant to remain consistent with the previous analyses. In other words, it was examined whether there was a correlation between confidence and response time for each voice spoken in the lineup, resulting in six correlations. The only voice of the six that had a significant correlation was the target voice, $r(468) = -.16$, $p < .001$. When participants were presented with the target voice, they spent less time examining the voice when they were more confident that it was the target voice. In comparison, the alternative voices in the lineup did not have a significant correlation between confidence and response time. These correlations ranged from -.08 to .02, $p < .05$.

Discussion

Across participants, confidence ratings in identifying the target were relatively low. In examining Table 4, the highest average confidence rating was 6.62, although the average across all conditions was lower. This indicates that participants are relatively uncertain about their ability to discriminate the target voice from a lineup of voices. This is particularly noteworthy, as there was only a 30-minute delay between encountering the target voice and having to identify that voice in a lineup. In many real-life instances, the retention interval is much longer. While this study did not include varying retention intervals, previous earwitness research suggests that longer time delays between the original event and identifying the voice lead to a decrease in

performance (e.g. Clifford, 1980). The uncertainty observed in this study suggests that people are not confident in identifying voices, even in a context where there are no consequences for incorrect identifications, which raises concern for the field of earwitness testimony as a whole. These participants were bystanders to the original event, suggesting that earwitnesses who are not directly involved in an event may be poor sources of evidence, especially since previous research has indicated that even if earwitness confidence is high, accuracy is relatively poor (Olsson et al., 1998).

An analysis was conducted on just the target voice, to see if factors such as pitch, emotion, and content would influence participants' confidence ratings towards the target voice. These results indicated a general matching effect of the target voice, in that confidence ratings were highest when the pitch of the voice matched that of the original conversation as well as when the emotion of the voice matched. This suggests that it was easier for participants to recognize the target voice when it was presented in the same manner as the original conversation. Another way to look at this is that participants were less confident that the target voice was the voice from the original conversation when it deviated from how the participants had originally heard it. This makes sense in regard to confidence ratings, as even if the participant is able to identify the voice as the target, they would not be as confident when the voice was presented differently, such as a different pitch or emotion.

To relate back to the original hypotheses, the matching effect was predicted for the emotion variable, but not for pitch. It had initially been predicted that participants would be most confident in the high-pitched version of the target voice. This was based on previous research conducted by Stern et al. (2007) where presenting a high-pitched voice led participants to more frequently select high-pitch distractor items. However, a distinction between their study and the

current one is that Stern et al. (2007) did not present participants with a matching pitch distractor voice. Based on the results found in the current study, it is possible that when participants did pick a distractor voice instead of the target voice, they would do so more frequently with a matching pitch distractor than a low or high-pitched distractor. Further, Stern et al.'s (2007) results are relevant to when participants select the distractor voice rather than the target voice. The main effect of Target Pitch does not include other voices from the lineup. Increased confidence for the medium-pitched target voice makes sense, due to the increased similarity it has to the original conversation.

As for Rodstrom and Neuhoff's (2003) study, they had found that the target voice was more likely to be identified when it was presented at a higher frequency than at a low frequency. This was not evident in the current study, in that there was no difference in target voice confidence ratings between the low and high pitch conditions. Had results in accord with those of Rodstrom and Neuhoff (2003) been obtained here, participants would have been more confident in the high-pitched target voice than in the low-pitched target voice. One reason for this discrepancy could be the difference in stimuli. The current study had actors naturally alter their voice pitch, whereas Rodstrom and Neuhoff (2003) used a computer program to alter the pitch of the original voice. It may be that using a computer program distorts a voice more noticeably when lowering the pitch. Having the actor alter their voice naturally may prevent or lessen this problem. Another possibility for the discrepancy is that the methodology was drastically different. The current study had participants listen to a conversation and then listened to six voices in a lineup, of which only one voice was the target voice. In comparison, Rodstrom and Neuhoff (2003) presented participants with only two voices across 132 trials, although altered in pitch. The repeated exposure to the target voices may impact the ease of which participants

recognized the voice, and pitch may have impacted recognition in a different manner than seen in the current study. However, these possibilities cannot be confirmed with the current study's methodology and would have to be examined further.

While it was hypothesized that there would be interactions of Lineup Emotion and Target Pitch with Conversation Content, neither were observed in the current study. It is suggested that the matching effect between the original conversation and the lineup voice overwhelmed any other possible biases that could occur, such as those seen in the target-absent experiment. Specifically, if the emotional tone of voice or the voice pitch of the target matches that of the original conversation, then the content of what is being said has little effect on earwitnesses recognition of the target voice. As such, it may be most beneficial for people conducting earwitness lineups to ensure that there is some degree of resemblance in regard to voice pitch and emotion between the original event and the voice lineup. It would be near impossible to conduct a lineup that exactly matches that of the original event in real life, but the current results indicate that there would be greater success in identifying the target when conditions are more similar than dissimilar.

While it is important to know the factors that influence confidence ratings of the target voice itself, knowing if earwitnesses are able to distinguish the target voice from other, non-target voices is imperative. Using a within-subject ANOVA, it was determined that participants were able to distinguish the target voice from all the other voices in the lineup, regardless of the other variables. Participants had the highest confidence ratings for the target voice, compared to each of the alternative voices. The ability to discriminate the target was further tested by comparing the confidence ratings of the target voice to those of the matching, non-target voice. The design of the current experiment ensured that there was always a non-

target voice that matched the target voice in terms of pitch, emotion, and content. Essentially, the only difference was the actor who spoke the lines. The purpose of this analysis was to see if there were certain conditions in which it was easier to distinguish between the target voice and the matching, non-target voice. The analysis that compared the target voice to the matching, non-target voice also allowed for an examination of whether participants were able to identify the target voice or if they were just more confident in voices that were presented in particular conditions. For example, if participants were generally more confident in medium-pitched, emotional voices, or if they were best able to identify the target when it closely matched the original conversation.

The analysis comparing the target voice and the matching, non-target voice indicated that there are certain instances where participants find it easier to differentiate between the target voice and the matching, non-target voice. Specifically, participants were best at distinguishing the target voice from the matching non-target voice in the Criminal Content condition, when the emotion matched that of the original conversation, and when the target pitch was either medium or high. Participants were more confident in the target voice than the matching non-target voice in the Vulnerable and Neutral Content conditions. However, different levels of pitch and different emotional tones seemed to impact confidence levels in the Criminal Content condition to a greater extent, as a three-way interaction between Voice, Lineup Emotion, and Target Pitch only occurred in the Criminal Content condition. As such, it appears that participants' ability to recognize the target voice after a criminal event may be more susceptible to differences in the lineup, compared to witnessing a neutral or vulnerable event.

The three-way interaction was then further broken down to compare Emotional and Monotone Lineups. Again, participants were able to distinguish between the target voice and the

matching, non-target voice in both conditions. However, only the Emotional Lineup had an interaction between Voice and Target Pitch. Participants were more confident in the target voice compared to the matching, non-target voice when both voices were presented in an emotional tone and in either a high or medium pitch. From this, it seems as though participants have a more difficult time identifying the target voice when it is presented in a lower pitch to what they had originally heard. One possibility for this is that the current study used female voices, which are generally higher pitched than male voices. People may be less experienced with distinguishing low-pitched female voices, which could account for the non-significant result found in the low-pitch, emotional, criminal condition. If this is the case, it would be expected that participants would have a more difficult time distinguishing between a male target voice and a matching, non-target voice, in a high-pitch, emotional, criminal condition.

Finally, the correlational analysis provided some insight into how confidence and response time are related to each other. These results replicated those of Experiment 2, in which there was a negative correlation between response time and confidence across all lineup voices together. The more confident a participant was in a voice being the target voice, the quicker they responded with their confidence judgment. This also replicates the results of Sporer (1992, 1993), although it is a much weaker correlation. In contrast to Experiment 2 and Sauer et al. (2008), the first voice in the lineup did not have a significant positive correlation. It is suggested for Experiment 2 that there was a positive correlation for the first voice presented due to participants getting used to the methodology and not having heard other options in the lineup. If this is the case, it appears that it would not be a consistent finding, although additional research would need to be conducted to confirm that.

Experiment 3 also examined the correlation between confidence and response time for each voice in the lineup. Since the target voice was included in the lineup, it was possible that participants may interact with that voice differently compared to the alternative voices, and this would be reflected in how long it took participants to respond with their confidence judgment. As such, this analysis found that participants who were more confident that the target voice was actually the target voice spent less time examining the voice. This negative correlation was not found for any of the alternative voices. This suggests that the overall negative correlation between response time and confidence was largely driven by how participants interacted with the target voice. Sporer (1992) did find that response time and accuracy also correlated negatively with each other (r = -.36), although this again was a stronger correlation compared to the one found for the target voice specifically (r = -.16). An area of further research could compare eyewitness and earwitness research, to first see if these results can be replicated, but also whether the strength of the relationship differs between the two types of testimonies. However, this is a promising start for earwitness research, and indicates that response time may be a useful tool in measuring the legitimacy of earwitness testimonies.

General Discussion

The majority of witness testimony research has been conducted using eyewitnesses, despite the utilization of earwitness testimony in the justice system. The goal of the experiments described here was to expand on earwitness research and examine variables such as voice pitch, emotion, and content to see how they may bias earwitness confidence ratings. To that extent, this research was successful, finding that the participants would respond with differing levels of confidence depending on the pitch of the voices, the emotion of the voices in the lineup, and the

content of the message heard in the original conversation. Furthermore, there were differences observed based on whether the target was present or absent from the lineup.

In real-life scenarios, it is unknown whether the perpetrator is present in the lineup, thus witness testimony research often utilizes both target-present and target-absent lineups to ensure greater ecological validity. This was implemented in the current study, allowing for the observation of biases with and without the presence of the voice participants were instructed to identify. Indeed, there were differences amongst the two experiments that were relevant to the hypotheses. For instance, in the target-present experiment, participants were more confident in the target voice when it was presented in the same pitch as the original conversation, compared to a high or low pitch. For the target-absent experiment, participants were also most confident in medium-pitched voices, even though they were not the target voice. However, Experiment 2 also found an interaction between Lineup Pitch and Conversation Content, which was not present in the target-present experiment. In Experiment 2, participants in the Criminal Content condition were least confident in high-pitched voices, whereas participants in the Vulnerable Content condition were least confident in low-pitched voices. This could be accounted for by a frequency heuristic, where preconceived notions about who a criminal is and who a victim is biased people to be less confident in voices that did not fall in line with their previous beliefs. For example, the general population is more likely to describe a criminal as male (Allison et al., 2013). Men typically have lower-pitched voices, so participants in the current study may have been disinclined to believe that a high-pitched voice would be a criminal. Low confidence in the high-pitched voice may have been exaggerated due to the lineup containing low-pitched voices, due to a frequency heuristic, and medium-pitched voices, which matched the pitch of the original conversation. However, in Experiment 3, the interaction between Target Pitch and Conversation

Content did not occur. It is suspected that because participants heard the actual target voice, the content of what the voice was saying did not impact their confidence ratings as hypothesized. Whether the target voice matched the conditions of the original conversation appeared to be more influential on participants' confidence ratings. The difference in how pitch influences earwitness confidence depending on whether the target is present or absent indicates that extra caution should be taken with regard to pitch. If it is unknown whether the target is present in the lineup, it may be difficult to pinpoint the effect voice frequency would have when conducting an earwitness lineup. While there needs to be further research done in the area of voice frequency, the current experiments suggest that not controlling for voice pitch in a lineup may lead earwitnesses to be biased to select certain voices, depending on the content of the event previously experienced.

There is also the issue of voice disguise related to pitch that would need to be taken into account when developing an earwitness lineup. Very little research has been conducted on voice disguise in relation to earwitness lineups. One of the studies described in Clifford's (1980) review discusses how accuracy decreased when the original event occurred in a disguised voice and the lineup did not. However, in Clifford's (1980) experiment, the actors who created the voice recordings were allowed to disguise their voices by their own choosing, so there was little control on how the voice was disguised. Of concern to the current topic, a perpetrator could disguise their voice through pitch during the original encounter. If they were to speak in their regular pitch during an earwitness lineup, it would likely decrease the ability for an earwitness to identify the perpetrator, since it no longer matches the conditions of the original event. While the purpose of the current study was to see if people were biased towards selecting certain voice pitches depending on the content of the message, future research could also examine how

altering the voice pitch during the original event influences earwitness confidence for lineup voices when they are presented in their natural pitch.

Another difference that was observed between Experiment 2 and 3 was how emotion affected confidence ratings. In the target-absent experiment, having a lineup with emotional voices did not promote higher confidence ratings compared to a monotone lineup. In contrast, when examining emotional differences in the target voice as well as comparing the target voice to a matching, non-target voice, results indicated having a match in emotion between the original conversation and the lineup was beneficial, as it increased target voice confidence ratings and the ability to distinguish between the target voice and the matching, non-target voice. This supports previous research which indicates that accuracy of identifying the target voice is improved when the emotion in the lineup matches the original event (e.g. Read & Craik, 1995; Saslove & Yarmey, 1980). This finding provides further evidence that there are differences between target-present and target-absent earwitness lineups. Many witness testimony studies include both types of lineups to ensure generalization to real-life situations where there is an innocent suspect (Wells & Turtle, 1986). Thus, finding this difference between the two types of lineups has practical implications for developing earwitness lineups. A reason for this difference in emotion could be that no matter the emotion of the voice, if an earwitness does not recognize the target voice, the match in emotion will not increase their confidence across a lineup, as seen in the target-absent experiment. In regard to the target voice, by having a matching emotion, recognition of the target voice may increase, which would result in the increase in confidence seen in Experiment 3. As such, it seems to be beneficial to match the emotion in the lineup to that of the original event even if it is unknown whether the target is present or not.

The current experiments examined variables that have had little attention in earwitness literature. The previous research in voice frequency has indicated that participants are more successful at identifying target voices when they are presented at a high pitch compared to low pitch (Rodstrom & Neuhoff, 2003). Further, if the target voice is originally presented at a high pitch, participants are more likely to select a high-pitched distractor than a low-pitched distractor (Mullennix et al., 2010; Stern et al., 2007). The current study found that when the pitch of the target speaker in the original event matches the pitch of a voice in the lineup, participants are more confident that the voice is the target voice, even if it was a different speaker. This was seen with higher confidence ratings for medium-pitched voices in Experiment 2 and for the medium-pitched target voice in Experiment 3. Based on these results, it appears that deviations of pitch between the original event and the earwitness lineup may be detrimental and should be avoided in order for a witness to successfully identify the target. However, as discussed earlier, the perpetrator may have disguised their voice, which raises practical concerns for developing an earwitness lineup that matches the pitch of the original event if the lineup administrators do not have an example of the voice.

The current experiments were also novel in that they took into account the content of the conversations that participants heard. With earwitness research, the content of the messages ranges from neutral to criminal. For example, Orchard and Yarmey (1995) used a monologue of a kidnapper, whereas Kerstholt et al. (2004) had the speaker answer questions related to his life. In the current research, Criminal, Neutral, and Vulnerable Content were compared to see if the difference in topics would affect earwitness confidence. In Experiment 2, it was found that Conversation Content and Voice Pitch interacted, a finding novel to earwitness research. Participants were most confident in the pitch that matched the original conversation, as well as

the pitch that is supported by stereotypes in media and crime statistics. Real-life implications of this are relevant to when an innocent person is a suspect in a case. Of importance, if the innocent person is suspected of a crime and has a low-pitched voice, or one that matches the pitch of the true perpetrator, they may be more likely to be identified by an earwitness despite being innocent.

It was also noted in Experiment 3 that a particular group of circumstances make it easier for an earwitness to distinguish between the target voice and the alternative voices, another novel finding. When participants were presented with a criminal situation and provided with a target that matched in both pitch and emotion, they were most confident in that voice compared to the matching, non-target voice. However, there is also a negative side to this, where any deviation from an exact match deteriorates confidence in the target voice when presented in a criminal context. This goes back to the idea of voice disguise at the original event. If the disguised voice is difficult to match, earwitnesses may not be able to identify the target. Further, even if the lineup is presented in an emotional tone of voice, if it does not match that of the original circumstance, confidence may be diminished as well. This conclusion is supported by findings from Mullennix et al. (2002) where even if the two emotions used are similar (e.g. anger and commanding), participants still had a more difficult time identifying the target voice when the emotions were different between the original event and the lineup. It appears that the best-case scenario for earwitnesses identifying a criminal is when the conditions of the original event and the lineup are an exact match, and there is a decline in performance as soon as the lineup deviates from the original event.

A final aspect of the current studies that provide novel insight to earwitness testimony is that confidence and response time are negatively correlated. To my knowledge, no earwitness

research has examined this relationship before. While exciting to see that a relationship exists, these results should be interpreted with caution. Although the correlation was significant, it was weak, and may not replicate in future earwitness research. Additionally, these results may not correspond to accuracy findings and thus could not be used as a diagnostic tool. Despite these limitations, it can be inferred that participants who were confident in a voice as being the target spent less time listening to the voice and making a decision. This was also seen when examining the target voice as opposed to the alternative voices. Further examination of this relationship may lead to helpful tools to help determine the legitimacy of an earwitness testimony.

There were a few limitations of the experiments described. One limitation of this study is the amount of data that needed to be excluded. Across all experiments, approximately 20 percent of participants were excluded based on pre-determined exclusion criteria. Despite the seemingly simple nature of the task, it appeared that participants had a difficult time completing the task. This may have been a consequence of conducting the study online, rather than in a lab, as participants may have been less inclined to pay attention to the study and technical difficulties could not be controlled for. However, there were justifications for all exclusion criteria to ensure that participants had equivalent experiences. As such, it should not diminish the findings of these experiments.

Another limitation of the study is that the voices used in the lineup were not controlled for distinctiveness. Research conducted by Orchard and Yarmey (1995) indicated that there were differences in earwitness accuracy and confidence depending on whether they heard a distinctive or non-distinctive voice. They defined distinctiveness as "…highly striking and not likely to be confused with other voices" (Orchard & Yarmey, 1995, p. 252). While the target voices were controlled for by having counterbalances of who portrayed the target voice and the other two

individuals in the original conversation, the voices in the lineup were not tested for distinctiveness. In Experiment 3, Alternative Voice C had significantly lower confidence ratings compared to all the other voices in the lineup. While it cannot be known for sure based on the methodology of this study, it is possible that this voice was distinctive enough from the target voice that participants were disinclined to believe it was the target voice. Future research could control for this by conducting a pilot test asking participants about the distinctiveness of voices, using the definition outlined above.

To conclude, the current research demonstrated that earwitness testimonies can be influenced by several factors, including pitch, emotion, and content. The results of the experiments presented indicate that the closer a lineup voice is to the original event, the more confident earwitnesses are. For the target voice, a match in emotion and pitch increases earwitness confidence. For voices that match the pitch of the original event, earwitnesses are more confident that those voices are the target voice, even if they were presented with a target-absent lineup. Furthermore, if the perpetrator is not in the lineup, earwitnesses are likely to be influenced by a frequency heuristic, where previous biases about the characteristics of criminals and victims influence confidence in certain voices. Specifically, earwitnesses who are identifying a perpetrator in a criminal context will be more confident in a low-pitched voice. In a vulnerable context, earwitnesses are more confident in a high-pitched voice. Due to the presence of these biases, it is suggested that the variables examined in these experiments should be controlled for in earwitness lineups and in future research.

References

Allison, M., Sweeney, L., & Jung, S. (2013). A comparison of Canadian and American offender stereotypes. *North American Journal of Psychology*, *15*(3), 589–607.

Bothwell, R. K., Deffenbacher, K. A., & Brigham, J. C. (1987). Correlation of eyewitness accuracy and confidence: Optimality hypothesis revisisted. *Journal of Applied Psychology*, *72*(4), 691–695. https://doi.org/10.1037/0021-9010.72.4.691

Campos, L., & Alonso-Quecuty, M. L. (2006). Remembering a criminal conversation: beyond eyewitness testimony. *Memory*, *14*(1), 27–36. https://doi.org/10.1080/09658210444000476

Clifford, B. R. (1980). Voice identification by human listeners: On earwitness reliability. *Law and Human Behavior*, *4*(4), 373–394. https://doi.org/10.1007/BF01040628

Deffenbacher, K. A., Cross, J. F., Handkins, R. E., Chance, J. E., Goldstein, A. G., Hammersley, R., & Read, J. D. (1989). Relevance of voice identification research to criteria for evaluating reliability of an identification. *The Journal of Psychology*, *123*(2), 109–119.

Flowe, H. (2011). An exploration of visual behaviour in eyewitness identification tests. *Applied Cognitive Psychology*, *25*(2), 244–254. https://doi.org/10.1002/acp.1670

Haw, R. M., & Fisher, R. P. (2004). Effects of administrator-witness contact on eyewitness identification accuracy. *The Journal of Applied Psychology*, *89*(6), 1106–1112. https://doi.org/10.1037/0021-9010.89.6.1106

Juslin, P., Olsson, N. O. E., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence-accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and*

Cognition, *22*(5), 1304–1316. https://doi.org/10.1037/0278-7393.22.5.1304

Kerstholt, J. H., Jansen, N. J. M., Van Amelsvoort, A. G., & Broeders, A. P. A. (2006). Earwitnesses: Effects of accent, retention and telephone. *Applied Cognitive Psychology*, *20*(2), 187–197. https://doi.org/10.1002/acp.1175

Kerstholt, J. H., Jansen, N. J. M., Van Amersvoort, A. G., & Broeders, A. P. A. (2004). Earwitnesses: Effects of speech duration, retention interval and acoustic environment. *Applied Cognitive Psychology*, *18*(3), 327–336. https://doi.org/10.1002/acp.974

Laub, C. E., Wylie, L. E., & Bornstein, B. H. (2013). Can the courts tell an ear from an eye? Legal approaches to voice identification evidence. *Law & Psychology Review*, *37*, 119–158. https://doi.org/10.1525/sp.2007.54.1.23.

Lloyd-Bostock, S. M. A., & Clifford, B. R. (Eds.). (1983). *Evaluating Witness Evidence*. Chichester, GB: John Wiley & Sons, Ltd.

Malpass, R. S., & Devine, P. G. (1981). Eyewitness identification: Lineup instructions and the absence of the offender. *Journal of Applied Psychology*, *66*(4), 482–489. https://doi.org/10.1037/0021-9010.66.4.482

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3–35.

Mullennix, J. W., Bihon, T., Bricklemyer, J., Gaston, J., & Keener, J. M. (2002). Effects of variation in emotional tone of voice on speech perception. *Language and Speech*, *45*(3), 255–283. https://doi.org/10.1177/00238309020450030301

Mullennix, J. W., Stern, S. E., Grounds, B., Kalas, R., Flaherty, M., Kowalok, S., … Tessmer, B.

(2010). Earwitness memory: Distortions for voice pitch and speaking rate. *Applied Cognitive Psychology, 24,* 513–526. https://doi.org/10.1002/acp

Öhman, L., Eriksson, A., & Granhag, P. A. (2011). Overhearing the planning of a crime: Do adults outperform children as earwitnesses? *Journal of Police and Criminal Psychology, 26*(2), 118–127. https://doi.org/10.1007/s11896-010-9076-5

Olsson, N., Juslin, P., & Winman, A. (1998). Realism of confidence in earwitness versus eyewitness identification. *Journal of Experimental Psychology: Applied*, *4*(2), 101–118. https://doi.org/10.1037/1076-898X.4.2.101

Orchard, T. L., & Yarmey, A. D. (1995). The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification. *Applied Cognitive Psychology, 9,* 249–260.

Parrott, S., & Parrott, C. T. (2015). U.S. television's "Mean World" for white women: The portrayal of gender and race on fictional crime dramas. *Sex Roles, 73,* 70–82. https://doi.org/10.1007/s11199-015-0505-x

Pollack, I., Pickett, J. M., & Sumby, W. H. (1954). On the identification of speakers by voice. *The Journal of the Acoustical Society of America*, *26*(3), 403–406. https://doi.org/10.1121/1.1907349

Possley, M. (2015). Kirk Odom. Retrieved from https://www.law.umich.edu/special/exoneration/Pages/casedetail.aspx?caseid=3943

Possley, M. (2016). Cathy Watkins. Retrieved from https://www.law.umich.edu/special/exoneration/Pages/casedetail.aspx?caseid=4073

Qualtrics. (2005). Qualtrics. Provo, Utah, USA: Qualtrics. Retrieved from http://www.qualtrics.com

Read, D., & Craik, F. I. M. (1995). Earwitness identification: Some influences on voice recognition. *Journal of Experimental Psychology: Applied*, *1*(1), 6–18.

Reitano, J. (2017). Adult correctional statistics in Canada , 2015 / 2016. Retrieved from http://www.statcan.gc.ca/pub/85-002-x/2016001/article/14318-eng.htm

Rodstrom, M., & Neuhoff, J. G. (2003). Increased pitch increases accuracy of voice identification. *Perceptual and Motor Skills*, *97*(2), 665–670. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14620259

Saslove, H., & Yarmey, A. D. (1980). Long-term auditory memory: Speaker identification. *The Journal of Applied Psychology*, *65*(1), 111–116. https://doi.org/10.1037/0021-9010.65.1.111

Sauer, J. D., Brewer, N., & Wells, G. L. (2008). Is there a magical time boundary for diagnosing eyewitness identification accuracy in sequential. *Legal and Criminological Psychology*, *13*, 123–135. https://doi.org/10.1348/135532506X159203

Shapiro, P. N., & Penrod, S. (1986). Meta-analysis of facial identification studies. *Psychological Bulletin*, *100*(2), 139–156. https://doi.org/10.1037/0033-2909.100.2.139

Sherrin, C. (2016). Earwitness evidence: The reliability of voice identifications. *Osgoode Hall Law Hournal*, *52*(3), 819–862.

Soundcloud. (2008). Soundcloud. Retrieved from https://soundcloud.com/

Sporer, S. L. (1992). Post-dicting eyewitness accuracy: Confidence, decision-times and person

descriptions of choosers and non-choosers. *European Journal of Social Psychology*, *22*, 157–180.

Sporer, S. L. (1993). Eyewitness identification accuracy, confidence, and decision times in simultaneous and sequential lineups. *Journal of Applied Psychology, 78*(1), 22–33.

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin*, *118*(3), 315–327.

Stern, S. E., Mullennix, J. W., Corneille, O., & Huart, J. (2007). Distortions in the memory of the pitch of speech. *Experimental Psychology*, *54*(2), 148–160. https://doi.org/10.1027/1618-3169.54.2.148

Wells, G. L., & Loftus, E. F. (Eds.). (1984). *Eyewitness Testimony: Psychological Perspectives*. New York, NY: Cambridge University Press.

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest*, *7*(2), 45–75.

Wells, G. L., & Turtle, J. W. (1986). Eyewitness identification: The importance of lineup models. *Psychological Bulletin*, *99*(3), 320–329. https://doi.org/10.1037/0033-2909.99.3.320

Wilding, J., & Cook, S. (2000). Sex differences and individual consistency in voice identification. *Perceptual and Motor Skills*, *91*, 535–538.

Yarmey, A. D., & Matthys, E. (1992). Voice identification of an abductor. *Applied Cognitive Psychology*, *6*, 367–377.

Table 1

Summary of Actor Counterbalances

	Role in Script		
Counterbalance	Person A	Person B	Person C
1	Actor 1	Actor 2	Actor 3
2	Actor 1	Actor 3	Actor 2
3	Actor 2	Actor 1	Actor 3
4	Actor 2	Actor 3	Actor 1
5	Actor 3	Actor 1	Actor 2
6	Actor 3	Actor 2	Actor 1

Table 2

Mean Confidence Ratings as a Function of Pitch, Content, and Emotion

	Criminal		Neutral		Vulnerable	
	Emotional	Monotone	Emotional	Monotone	Emotional	Monotone
High	1.72 (.34)	1.17 (.34)	3.05 (.52)	2.93 (.45)	2.56 (.42)	2.39 (.38)
Medium	2.55 (.36)	2.86 (.45)	3.14 (.54)	3.66 (.39)	2.75 (.45)	3.04 (.44)
Low	2.31 (.37)	2.34 (.42)	2.73 (.50)	3.04 (.38)	1.47 (.34)	1.90 (.33)

Note. Results from Experiment 2. Standard error is provided in parentheses.

Table 3

Correlation between Confidence and Response Time for Experiment 2 and 3

	Experiment	
Voice Order	2	3
1	.183*	.050
2	-.106	-.043
3	-.171*	-.117*
4	-.097	-.078
5	-.145*	-.046
6	-.094	-.096*

Note. Partial correlations conducted based on the order the voice appeared in the lineup. Experiment 2, N = 200. Experiment 3, N = 465. *$p < .05$.

Table 4

Target Voice Mean Confidence Ratings as a Function of Pitch, Content, and Emotion

	Criminal		Neutral		Vulnerable	
	Emotional	Monotone	Emotional	Monotone	Emotional	Monotone
High	4.00 (.70)	3.70 (.59)	3.68 (.63)	3.32 (.68)	4.36 (.78)	3.04 (.60)
Medium	6.62 (.65)	3.89 (.65)	4.96 (.60)	3.75 (.65)	4.46 (.85)	5.04 (.72)
Low	4.82 (.72)	4.54 (.74)	3.92 (.72)	3.41 (.73)	4.91 (.92)	3.96 (.71)

Note. Results from Experiment 3. Standard error is provided in parentheses.

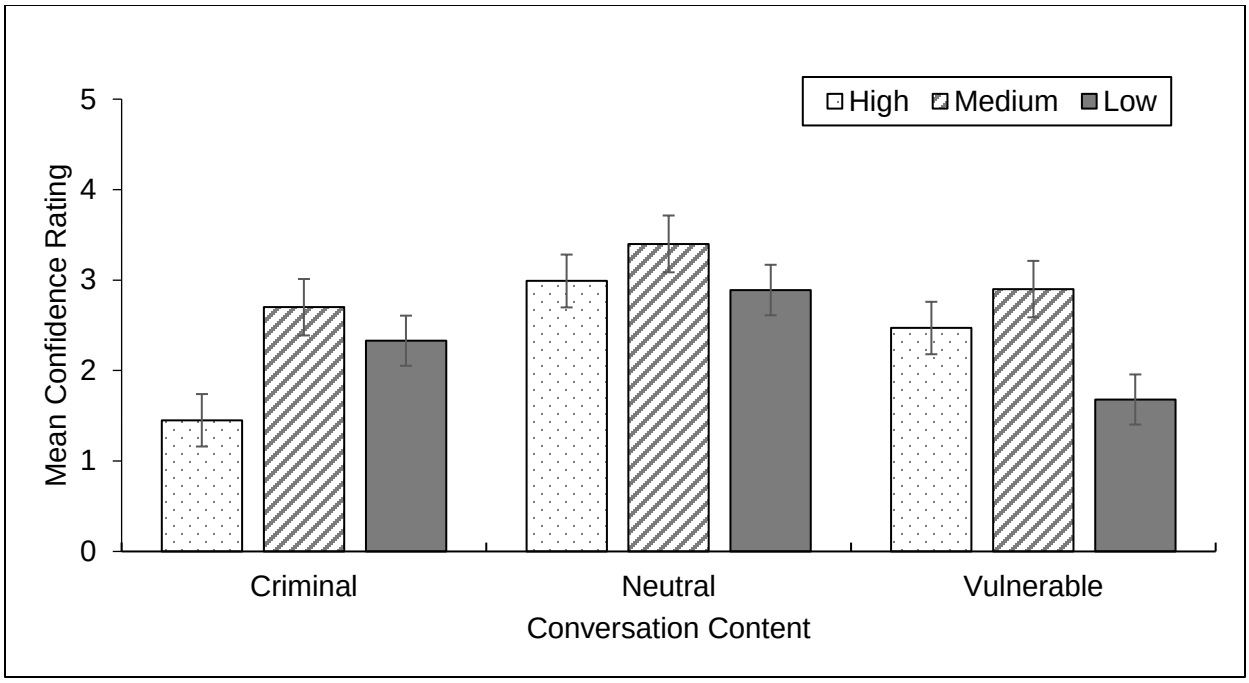

Figure 1. Interaction between Conversation Content and Voice Pitch in Experiment 2. Error bars indicate standard error.

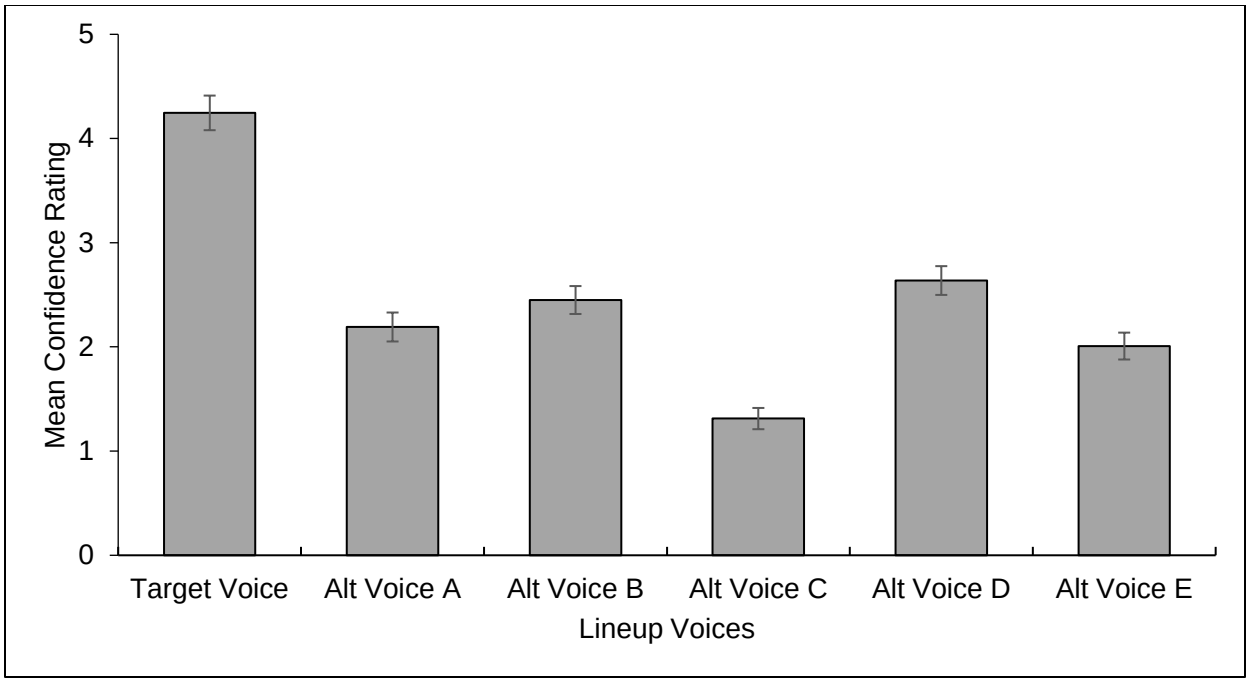

Figure 2. Differences in Confidence Ratings between Voices in Experiment 3. Alternative voices are labelled as “Alt Voice A”, “Alt Voice B”, etc. Error bars indicate standard error.

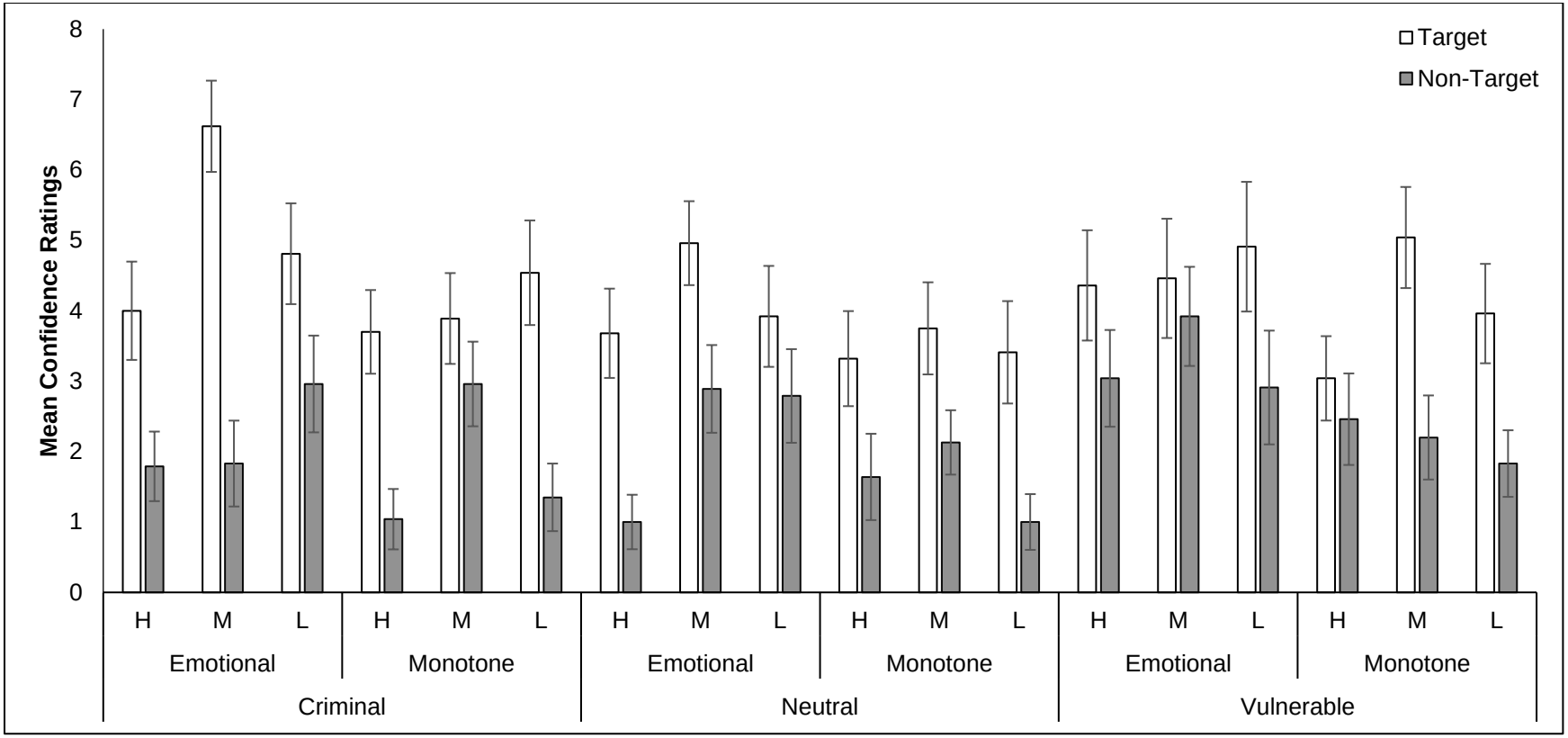

Figure 3. Four-way interaction between Voice, Conversation Content, Emotion, and Target Pitch. Data from Experiment 3. Error bars represent standard error.

Appendix A

Criminal Condition Script

Person A: So, did Angela ever end up calling you about filling in for her next Monday?

Person B: No, she didn't.

Person A: Huh… she mentioned it to me just the other day. She was pretty stressed out because she couldn't find anyone to cover for her and I think she has to move out of her apartment that day or something like that.

Person B: Well I guess she found someone else. Which is good, because I couldn't cover for her Monday anyways.

Person A: Why not? You said you had Monday off which is why I suggested she call you.

Person B: I do have it off, but my parents will be in town that day. They just let me know this morning.

Person A: Ah, that's pretty short notice!

Person B: I know! And my place is a mess because we are still redoing the kitchen. And you know my mom, she's going to be all like-

Person C: Hey you- stop right there.

Person A: Uh.... what?

Person C: Give me all of your money, right now. I have a knife. Hurry up and no one will get hurt.

Person B: … Okay I… just have to get it out of my wallet. Here- this is all I have. Please, we don't want any trouble.

Person A: I don't have any money.

Person C: Open up your wallet NOW.

Person A: Here, there's nothing.

Person C: Give me your bag! Get out of here both of you and be quiet.

Note: Person C is the target.

Appendix B

Neutral Condition Script

Person A: So, did Angela ever end up calling you about filling in for her next Monday?

Person B: No, she didn't.

Person A: Huh… she mentioned it to me just the other day. She was pretty stressed out because she couldn't find anyone to cover for her and I think she has to move out of her apartment that day or something like that.

Person B: Well I guess she found someone else. Which is good, because I couldn't cover for her Monday anyways.

Person A: Why not? You said you had Monday off which is why I suggested she call you.

Person B: I do have it off, but my parents will be in town that day. They just let me know this morning.

Person A: Ah, that's pretty short notice!

Person B: I know! And my place is a mess because we are still redoing the kitchen. And you know my mom, she's going to be all like-

Person C: Excuse me, I think you dropped your hat. I just saw it on the ground- I'm not sure if it's yours or someone dropped it earlier?

Person A: Oh! I didn't notice I dropped it- yes that is mine!

Person C: Here you go.

Person A: Thank you so much!

Person C: No problem. Have a nice day!

Note: Person C is the target.

Appendix C

Vulnerable Condition Script

Person A: So, did Angela ever end up calling you about filling in for her next Monday?

Person B: No, she didn't.

Person A: Huh… she mentioned it to me just the other day. She was pretty stressed out because she couldn't find anyone to cover for her and I think she has to move out of her apartment that day or something like that.

Person B: Well I guess she found someone else. Which is good, because I couldn't cover for her Monday anyways.

Person A: Why not? You said you had Monday off which is why I suggested she call you.

Person B: I do have it off, but my parents will be in town that day. They just let me know this morning.

Person A: Ah, that's pretty short notice!

Person B: I know! And my place is a mess because we are still redoing the kitchen. And you know my mom, she's going to be all like-

Person C: Excuse me, I'm so sorry to bother you but I was wondering if you've seen an orange cat around this area? He usually comes home in the early evening and I haven't seen him since yesterday.

Person A: No, we haven't, sorry!

Person B: Hope he turns up!

Person C: Okay, well thank you. Have a nice day!

Note: Person C is the target.

Appendix D

Multiple-Choice Questions

Criminal Content Condition:

"In the conversation you just listened to, what was asked for?"

- Money
- Paper
- Clothing
- Nothing

Neutral Content Condition

"In the conversation you just listened to, what was dropped?"

- Hat
- Purse
- Book
- Nothing

Vulnerable Content Condition

"In the conversation you just listened to, what was missing?"

- Cat
- Book
- Purse
- Nothing

Appendix E

Informed Consent (Experiment 1)

LETTER OF INFORMATION
FACTORS THAT INFLUENCE VOICE IDENTIFICATION

We invite you to participate in this study. Our goal is to better understand people's memory for voices. Participants will listen to multiple short audio recordings of a simulated conversation and then answer some questions related to the recordings. Normally, participants complete the study in 15 minutes or less. All of the information collected is confidential, meaning that only the Principle Investigator and Co-Investigators will have access to the anonymous data files which will be stored on a password protected computer in an office in the Duff Roblin Building. The University of Manitoba Research Ethics Board(s) and a representative(s) of the University of Manitoba Research Quality Management/Assurance office may also require access to your research records for safety, and quality assurance purposes. In reporting the results, we will only report our summary findings and not individual data. Your name will be collected and used only for the purpose of assigning credits. The electronic data collected from the study will be kept confidential, and will be disclosed only in forms consistent with the standards for the publication/dissemination of research. Some basic demographic information (age, gender) will also be collected. If you feel uncomfortable providing those details, or feel those details could be used to identify you with your anonymous data file, you may omit entering them.

If you consent to participate, you will earn one research credit. If you withdraw partway during the study, you will still receive the credit.

The potential risks associated with participating in this study are no greater than those which people likely experience in everyday life as a consequence of using online platforms provided by companies based in the United States. By answering questions in our survey, there is also some risk that the United States government could access your answers because our survey is hosted by an American-based company (i.e. Qualtrics) and all such companies are subject to American laws such as the Patriot Act. People frequently use services of such companies, such as Facebook, Google, and Hotmail.

Principal Investigator
Dr. Launa Leboe-McGowan
Assistant Professor
Depart of Psychology
University of Manitoba
Office Telephone:
Department FAX:
Email:

Co-Investigators
Dr. Jason Leboe-McGowan

Professor, Associate Dean of Arts
Department of Psychology
University of Manitoba

Dr. Doug Alards-Tomalin
Post-Doctoral Fellow
Department of Psychology
University of Manitoba

Kelly Thiessen
Master's Student
Department of Psychology
University of Manitoba
Email:

INFORMED CONSENT
FACTORS THAT INFLUENCE VOICE IDENTIFICATION

The Psychology/Sociology Research Ethics Board has approved this research. If you have any concerns or complaints about this project you may contact the researchers by email as listed above; alternative, you may contact the Human Ethics Secretariat at _____, or email _____. A copy of this consent form can be printed off for your records and reference, or you can receive an electronic (PDF) version from Kelly Thiessen upon request (submitted by email).

If you wish to participate in this study, please indicate that you have read and understood the above information to your satisfaction by clicking "**I agree**" below. In no way does clicking "**I agree**" waive your legal rights nor release the researchers, sponsors, or involved institutions from their legal and professional responsibilities. You are free to withdraw from the study at any time, and/or refrain from answering any questions you prefer to omit, without prejudice or consequence. You should only click "**I agree**" if you agree to participate with full knowledge of the study presented to you in this information and consent form and of your own free will.

Please select "**I agree**" if you wish to participate.
If you would like to withdraw and would like to exit the survey, please select "**I disagree**".

- I agree
- I disagree

Appendix F

Informed Consent (Experiment 2 and 3)

LETTER OF INFORMATION
FACTORS THAT INFLUENCE VOICE IDENTIFICATION

We invite you to participate in this study. Our goal is to better understand people's memory for voices. Participants will listen to a short audio recording of a simulated conversation and answer several arithmetic questions. Next, participants will listen to different short audio recordings of simulated conversations and then answer some questions about those recordings. Normally, participants complete the study in 50 minutes or less. All of the information collected is confidential, meaning that only the Principle Investigator and Co-Investigators will have access to the anonymous data files which will be stored on a password protected computer in an office in the Duff Roblin Building. The University of Manitoba Research Ethics Board(s) and a representative(s) of the University of Manitoba Research Quality Management/Assurance office may also require access to your research records for safety, and quality assurance purposes. In reporting the results, we will only report our summary findings and not individual data. Your name will be collected and used only for the purpose of assigning credits. The electronic data collected from the study will be kept confidential, and will be disclosed only in forms consistent with the standards for the publication/dissemination of research. Some basic demographic information (age, gender) will also be collected. If you feel uncomfortable providing those details, or feel those details could be used to identify you with your anonymous data file, you may omit entering them.

If you consent to participate, you will earn two research credits. If you withdraw partway during the study, you will still receive these credits.

The potential risks associated with participating in this study are no greater than those which people likely experience in everyday life as a consequence of using online platforms provided by companies based in the United States. By answering questions in our survey, there is also some risk that the United States government could access your answers because our survey is hosted by an American-based company (i.e. Qualtrics) and all such companies are subject to American laws such as the Patriot Act. People frequently use services of such companies, such as Facebook, Google, and Hotmail.

Principal Investigator
Dr. Launa Leboe-McGowan
Assistant Professor
Depart of Psychology
University of Manitoba
Office Telephone:
Department FAX:
Email:

Co-Investigators
Dr. Jason Leboe-McGowan
Professor, Associate Dean of Arts
Department of Psychology
University of Manitoba

Dr. Doug Alards-Tomalin
Post-Doctoral Fellow
Department of Psychology
University of Manitoba

Kelly Thiessen
Master's Student
Department of Psychology
University of Manitoba
Email:

INFORMED CONSENT
FACTORS THAT INFLUENCE VOICE IDENTIFICATION

The Psychology/Sociology Research Ethics Board has approved this research. If you have any concerns or complaints about this project you may contact the researchers by email as listed above; alternative, you may contact the Human Ethics Secretariat at _____, or email _____. A copy of this consent form can be printed off for your records and reference, or you can receive an electronic (PDF) version from Kelly Thiessen upon request (submitted by email).

If you wish to participate in this study, please indicate that you have read and understood the above information to your satisfaction by clicking "**I agree**" below. In no way does clicking "**I agree**" waive your legal rights nor release the researchers, sponsors, or involved institutions from their legal and professional responsibilities. You are free to withdraw from the study at any time, and/or refrain from answering any questions you prefer to omit, without prejudice or consequence. You should only click "**I agree**" if you agree to participate with full knowledge of the study presented to you in this information and consent form and of your own free will.

Please select "**I agree**" if you wish to participate.
If you would like to withdraw and would like to exit the survey, please select "**I disagree**".

- I agree
- I disagree