Working Papers of the Cornell Phonetics Laboratory
2000, vol. 13, pp. 125–166

The Effect of Arousal on Earwitness Identification*

Traci L. Suiter

This study reports on the effect of arousal on earwitness identification. While performing a cover task, thirty-nine subjects (32 females and seven males) heard a series of utterances in one of three voice tones (positive, neutral, or hostile). Subjects were immediately presented with a six-voice lineup. All subjects were tested immediately after hearing the target voice, and a subgroup of the subjects (21: 14 females, seven males) were also tested after a one-week delay. In addition to the identification lineup, subjects gave the reason for their decision, their confidence level, their arousal level when initially hearing the target voice, height and weight estimates of the target voice, and any personality and physical characteristics they could ascribe to the voice they chose from the lineup. Eight correct identifications were obtained. The only significant results obtained were for height estimates between the Neutral Tone and Hostile Tone conditions and between the Neutral and Positive Tone conditions. The tone of the voice and the level of arousal had no significant effect on identification accuracy, and there was no significant correlation between the level of confidence and identification accuracy. Subjects reported using process of elimination to match the target voice to the voice they chose from the lineup, and they ascribed more physical than personality characteristics to the voice they chose from the lineup.

1 Introduction

Identification of suspects by witnesses or victims of crime can be one of the most important tools in law enforcement. Forensic science can be extremely effective, but a successful identification can elevate the status of a suspect from accused to convicted. In some instances, the ability to recognize a voice is extremely important, but there are several limitations on this ability: the ability to recognize a voice under violent and traumatic circumstances; the fact that the identification is based on incidental recognition; and the delay interval between first hearing the voice and later attempting to identify it.

This paper first reviews several factors of earwitness identification: voice lineups, incidental and intentional recognition, participation, the effect of delay, determining the age and sex of the voice, duration and nature of the voice sample, voice similarity, height and weight estimations, and the effect of stress and arousal. As will be seen below, many of the studies conflict and contradict with each other, and the present study seeks to clear up some of the confusion. Second, this study reports an experiment in earwitness

* I would like to acknowledge the valuable comments I received from Allard Jongman, Tobey Lynn Doeleman, and especially Michael Owren. Thanks to Eric Evans for technical assistance.

identification in which subjects were exposed to one of three tones of voice (positive tone, neutral tone, or hostile tone) and were asked to choose from a lineup which voice they heard. Subjects also gave height and weight measurements and were asked a series of questions about their decision-making process.

1.1 Voice lineup

Although many studies both in eyewitness and earwitness lineups have reported on the composition and possible biases of lineups, there has been no general consensus. Broeders and Rietveld (1995) not only explore the theoretical aspects of earwitness lineups, but also give explicit instructions for the construction of a voice lineup. The construction of the lineup in the present study patterns their suggestions.

While there are many factors in the construction of an earwitness lineup, Broeders and Rietveld note that one of the most important is the sample of the target voice. They state that "if possible, the material should obviously closely correspond in nature to the offender's speech to which the witness was exposed" (Broeders and Rietveld, 1995, p. 36). According to the Calgary Police procedure as described in Komulainen (1988), the argument is that the identification should be made "on the voice characteristics and *not* on how the suspect repeats the words spoken by the criminal during the offense." Broeders and Rietveld recommend recording 10 to 15 minutes of the suspect's speech while the suspect is describing a picture or some other neutral object. Then a number of foils should be recorded using the same procedure. About 60 seconds of the recorded material should be used. A pre-test with mock witnesses should be conducted to eliminate from the extended lineup any voice which is obviously dissimilar to the others. Order of presentation is determined by the witness, who selects the cassettes which have been assigned random numbers or letters.

Bull and Clifford (1984) found a significant difference in performance between subjects given four or six distractor voices (voices included in the lineup which are similar to the target voice, but which are not the target voice) in the lineup (68% and 48% correct identification, respectively). They also found no evidence that the position of the

target in the lineup influenced recognition performance unless the target was the first voice in the lineup.

1.2 Effect of incidental/intentional recognition on voice identification

Most studies agree that intentional recognition (being told to specifically remember a voice) is more accurate than incidental recognition (not being told to specifically remember a voice). While both kinds of recognition are forensically significant, the present study tested incidental recognition (with no surprising results).

Geiselman and Bellezza (1976) studied long-term memory for speaker's voice. In their study, 128 subjects were asked to recall 20 sentences. Some subjects were presented with the sentences in two different voices or by two loudspeakers located in different positions in the room. Some of the subjects were asked to remember not only the sentence, but the voice that uttered the sentence. Geiselman and Bellezza found a high degree of accuracy in subjects who were told to remember the voice.

Saslove and Yarmey (1980) also tested incidental recognition. Their subjects (all female college students) overheard from another room a taped telephone conversation with a female participant. The female voice spoke in an angry tone for 11 seconds. In the recognition test, the participants heard five taped recordings: one recording consisted of the original recording while four others consisted of different voices speaking the same words. The participants were asked to judge whether each recording was old or new; that is, whether the voice was one they had heard before or not. Some of the participants were told to remember the voice and some were not. Saslove and Yarmey found a 62% identification accuracy rate in their study for those who tested on intentional recognition. Those tested for their incidental recognition only scored slightly above chance level, which is consistent with other findings for incidental recognition.

1.3 Effect of participation on speech identification

Hammersley and Read (1985) conducted an experiment on the effect of participation on speech identification while at the same time producing more natural conditions of

voice perception. In their experiment, Hammersley and Read used thirteen pairs of subjects (all female). Each pair heard a conversation and then had a five-minute conversation on the same topic. The subjects heard and participated in the conversations through adjacent rooms connected with headphones and microphones. The topic of conversation was "how to plan the perfect bank robbery." They were not told to pay any specific attention to the voice of their partner. Forty-eight hours later, the subjects were presented with a set of voices and asked to select the voices they had heard. The recognition test consisted of all 26 voices.

Results from the experiment showed that subjects were better able to recall their own voices and the voices of their conversational partners than those voices to which they had passively listened. They concluded that "talking to someone leads one to recognize and identify their voice better than simply listening. In an active conversation voice carries more information and is attended to more than when the same conversation is heard passively" (Hammersley and Read, 1985, p. 79).

1.4 Effect of delay interval on speaker identification

The commonsense view is that the longer the time span between first hearing a voice and later identifying that voice, the less accuracy the subject will demonstrate. Yet many studies do not agree on how long the delay must be to significantly retard identification. In keeping with the timespan of most studies, subjects in the present study were tested both immediately and one week later, yet the accuracy was significantly below the levels reported in the studies below.

McGehee (1937, 1944) found that after 5 months recognition accuracy was no better than chance (20%). She found identification accuracy at 83% for two days, 68% for two weeks, 35% at three months, and 13% accurate identification at five months.

Clifford, Rathborn, and Bull (1981) conducted two studies on the effects of delay on voice recognition accuracy. In their first experiment, 176 listeners heard male and female voices and attempted to identify them from a lineup containing two target voices (one male, one female) and 20 distractors (equal amounts of male and female voices). The

subjects (all female nurses) were tested after either 10, 40, 100, or 130 minutes. Recognition accuracy after 10 minutes was 56%; after 40 minutes recognition accuracy was 41%; after 100 minutes, recognition accuracy was 41%; and recognition accuracy after 130 minutes was 44% (chance was 9%).

Experiment Two tested recognition accuracy after 10 minutes, 24 hours, 7 days, and 14 days. Again the subjects were female nurses (though not the same nurses as in Experiment One). The method and procedure were identical as before. Recognition accuracy after 10 minutes was 55%; accuracy after 24 hours was 32%; after 7 days, accuracy was 30%; and after 14 days, accuracy was 38% (again, chance was 9%). Clifford et al. state that the testing was constructed so that optimal performance was ensured: voices were presented by a tape recorder to rule out intravoice variability; they told subjects that the target was included in the lineup (a target-present lineup), and that they would be tested on their ability to identify the voice later.

Clifford and Denot (1982) also studied the effects of the delay interval. In a series of experiments they staged a live incident where a confederate entered the room and had a neutral or aggressive conversation with the experimenter. After a delay of either one, two, or three weeks the witness' ability to identify the voices was tested. Recognition accuracy was 50% after one week, 43% after two weeks, and 9% after three weeks.

1.5 Age and sex of listener and voice

Few studies agree on which sex is better not only at identifying, but at being identified. While the present study does not test any explicit assumption in this area, male voices were used and female subjects participated since the majority of violent crime is committed by males. A small number of male subjects also participated.

Thompson (1985) used six male voices and six female voices in separate same-sex lineups. A panel of judges pretested the lineup by eliminating all voices which had a distinctive accent or other distinctive characteristic. All subjects initially heard a voice sample of a bank robbery (82 words). After a retention interval of one week, they were presented with the same six voices reading a passage about anemia (72 words). The

lineup was repeated three times. Subjects had three options for response to the lineup: the position of the target voice in the lineup; the target was not in the lineup; or, the subject was not sure whether the voice was in the lineup. In addition, the subjects were asked to give a confidence rating with their answer.

Results show that male target voices were given a higher confidence rating for accuracy than female target voices. There was no effect of sex of subject. The male target voices were more accurately identified than female voices, yet there was no effect of sex of subject. Thompson, however, does not provide any interpretation of why male voices might be identified with more accuracy than female voices. Clifford (1980), however, found that female voices were better recognized than male voices, only Clifford treats this as an additional finding and provides no further interpretation of his results.

1.6 Duration and nature of the voice sample

The present study, like Bricker and Pruzansky (1966), used whole sentences, but did not find close to their identification accuracy. Bricker and Pruzansky (1966) found that their subjects had 56% correct identification when vowel excerpts were provided, 84% accuracy for syllables, and 93% accuracy for whole sentences (chance is 10%).

In a study by Saslove and Yarmey (1980), female college students overheard a taped female voice answer a telephone call and engage in an angry conversation. In the voice identification task, the target voice was presented talking in either the original hostile tone or in a more conversational tone. Two of the distractor voices were hostile and two were conversational. The change in tone reduced identification results to chance level. Identification accuracy did not improve after a 24 hour delay.

1.7 Acoustical measurement of voice similarity

In the present study, acoustical measurements of voice similarity were performed, based on Walden, Montgomery, Gibeily, Prosek, and Schwartz (1978), who suggested that in previous studies "a priori assumptions about the nature of the perceptual dimensions had to be made. That is, the investigators selected in advance what attributes

or characteristics of speech production the listeners would judge" (Walden et al., 1978, p. 266). They felt a better approach would match acoustic measures to perceptual judgments of voice similarity. In their study, twenty adult male speakers recorded the word "beans." They chose the word since it was a four-phoneme monosyllabic word, it has a relatively short duration, and it simplified the acoustic measurements. Furthermore, previous research (Pollack, Pickett, and Sumby, 1954; Bricker and Pruzanksy, 1966) showed that listeners had a high identification accuracy rate (as high as 80%) on a monosyllabic speech sample. Walden et al. measured vowel duration, nasal duration, fricative duration, word duration, first, second, and third formant frequencies, as well as mean fundamental frequency (the rate at which a sound source compiles its vibratory cycle). They found that word duration and mean fundamental frequency as dimensions are consistent with the findings of prior research in the area of voice perception and identification.

1.8 Height and weight estimation

Although strictly speaking not a variable in earwitness identification, estimating a person's height and weight by only listening to their voice is relevant to any forensic situation in which the victim could not visually identify the perpetrator.

Gunter and Manning (1982) studied listener estimations of speaker height and weight in unfiltered and filtered voice conditions. In their study, twenty speakers (10 male, 10 female) produced four steady state vowels (/i/, /u/, /a/, /ae/). Four tapes were created. One tape consisted of the unfiltered vowels. In the second tape the fundamental frequency was eliminated by filtering (F0 condition). The third tape consisted of the first formant bandwidth removed (F1 condition), and the fourth tape consisted of the second formant bandwidth removed (F2 condition). Forty listeners (20 male, 20 female) were asked to estimate the height and weight by listening not only to the unfiltered signal, but each of the three filtered signals. Gunter and Manning reported significant differences between actual and estimated heights and weights for individual listener estimates of individual speakers. In the unfiltered condition, estimated heights averaged .83 inches over the

actual heights, and the estimated weights averaged 7.75 pounds over the actual weights. In the F0 filtered condition, the estimated height averaged +.52 inches, while the estimated weight averaged +7.13 lbs; in the F1 filtered group, the estimated height averaged +.69 inches, while the estimated weight averaged +6.95 lbs; and finally in the F2 filtered condition, the estimated height averaged +.78 inches and the estimated weight averaged + 5.81 lbs. Gunter and Manning concluded that "since the listeners did not closely estimate speaker measurements when provided with the full spectrum, it is not surprising that providing the listener with less information did little to change his/her ability to perform the task. It would appear, therefore, that information contained in the speech spectrum does not reflect the characteristics of speaker height and weight, at least in an obvious manner" (Gunter and Manning, 1982, p. 256).

The most controversial studies in this area have been the studies of Norman Lass and his colleagues. Lass, Beverly, Nicosia, and Simpson (1978) studied speaker height and weight identification by means of direct estimate. Their purpose was to first "determine if listeners were capable of making accurate direct estimations of speakers' heights and weights from recorded speech samples" (Lass et al., 1978, p. 69) and secondly to determine the effect that the sex of the listener and speaker might have on these determinations.

A prose passage was read by 30 speakers (15 male and 15 female). A tape subsequently made of all the voices (randomized) was played to 40 judges (20 males and 20 females). All subjects then participated in two experiments, one to judge height and one to judge weight. In one group, half the subjects made height and then weight estimates, while in the other group the subjects first made judgments of weight and then of height. During the judgments, the middle sentence from the passage was played to the subjects again.

Results indicate that the listeners were capable of accurately identifying the approximate height and weight, while the sex of the speaker and the listener did not significantly affect the identification judgments. Male speakers' weights were overestimated by both male and female listeners while female weights were

underestimated by both males and females, the mean difference for male speakers by male and female listeners being +3.31 lbs and the difference between actual and estimated weight of female speakers was -3.64 lbs. Statistical analysis revealed no significant difference in weight judgments for the sex of the listener and the sex of the speaker.

Results from the height judgments indicate that male speakers' heights were underestimated for both male and female listeners, and that the same held true for the female speakers. The mean difference again was small: the mean difference for male speakers by male and female listeners was -.25 inches, and the mean difference for female speakers was -1.35 inches. As with the weight measurements, there was no statistical difference. Lass et al. recommend that further investigation should attempt to "isolate and define the important acoustic cues in the voice which may reflect speakers' heights and weights" (Lass et al., 1978, p.75).

Lass and Brown (1978) built upon earlier work, this time with a correlational study of speakers' heights, weights, body surface areas, and speaking fundamental frequencies. In this study, 30 college-aged speakers (15 male and 15 female) participated in the study. Lass and Brown employed the Fundamental Frequency Indicator (FFI) to obtain speakers' mean speaking fundamental frequencies. To determine the relationship between height, weight, body surface areas, and speaking fundamental frequencies, Lass and Brown conducted a Pearson product-moment correlation coefficient.

Results indicate that height and weight are significantly positively related, but height and speaking fundamental frequency, SFF, weight and body surface, and SFF were not significantly related. Lass and Brown state that "these findings do not support the hypothesis that, in general, taller-than-average and/or heavier-than-average speakers should have lower-than-average speaking fundamental frequencies" (Lass and Brown, 1978, p. 1219). They do suggest that there might be other variables which could work either alone or in conjunction with the speaking fundamental frequency that could provide cues for the speaker's perceptions.

Lass, Philips, and Bruchey (1980) studied the effect of filtered speech on speaker height and weight identification. A total of 30 speakers (15 male, 15 female) recorded a prose passage, which was subjected to three different conditions: unfiltered, 225 Hz low-pass filtered, and 225 Hz high-pass filtered. A total of 30 judges participated in three sessions, one for each of the tapes. In each session they were asked to make estimations of the height and weight for each speaker. The subjects were also asked for confidence ratings on their judgments. Results indicate that speaker height and weight identification was not significantly affected by the filtering of speech. The average difference between actual and estimated heights and weights in the unfiltered condition was 1.03 inches and 4.14 lbs, respectively. The average differences in the filtered condition was 1.79 inches for height, and 1.88 lbs for weight. Lass et al. conclude that "apparently different portions of the broadband speech spectrum contain adequate acoustic cues for accurate height and weight identification. The speaker's vocal tract resonance characteristics (i.e., formants) appear to play an equally important role in these tasks as their laryngeal fundamental" (Lass et al., 1980, p. 98).

Cohen, Crystal, House, and Neuburg (1980) critiqued Lass' work. They mainly focused on his studies concerning weight, and stated that Lass and his colleagues misrepresented their data by providing the "average differences" and not the actual deviations from the weight of the speaker. By taking the average difference, Lass and his colleagues were taking both positive and negative deviations, grouping them together, and then taking that average, instead of taking the average of only the positive and only the negative deviations. Cohen et al. state that "if the panel's accuracy is judged on the basis of how well it estimates the weight of a given individual, the results are not nearly as encouraging as suggested in the research reports" (Cohen et al., 1980, p. 1885).

Van Dommelen (1993) also weighed in with criticism of Lass' work, specifically Lass, Philips, and Bruchey (1980d). Van Dommelen's criticism stemmed from the basic hypothesis of Lass' work: that if listeners can truly estimate the height and weight of speakers from voice samples, there should exist a positive statistical correlation between the actual heights and weights and those heights and weights estimated by listeners.

Since Lass et al. (1980d) gave only histograms but no raw data, Van Dommelen reconstructed the raw data. Pearson product-moment correlations were calculated separately for male and female speakers. The results of the correlation for weight yielded statistically non-significant results (as did the present study--and contrary to Lass) for both male and female speakers. Van Dommelen states that "this finding implies that the listeners were not in the least able to identify the speakers' weights" (Van Dommelen, 1993, p. 338). In addition, Van Dommelen's analysis revealed that listeners were highly consistent in their inaccurate judgments. Similar results were obtained for the identification of height. Van Dommelen states that the positively and statistically significant correlations obtained by Lass and his colleagues were due to lumping the male and female speakers together, which is unacceptable since rather than give accurate estimations, listeners might give estimations based on the knowledge that generally men are both taller and heavier than women. In summary, then, subjects cannot accurately identify speakers' heights and weights from voice samples.

1.9 Effect of stress and arousal on speaker identification

Also one of the most disputed areas of earwitness identification is the effect of stress and arousal on speaker identification. Although each study has its own particular catalyst for introducing stress and arousal, the present study introduced by voice tones three different conditions of arousal.

Clifford and Hollin (1981) found that the testimony of witnesses to a violent incident was significantly poorer than that given by witnesses to a nonviolent incident. A violent incident could generate arousal or stress in the witness and, as such, could cause a narrowing of attention to a limited range of information. Kuehn (1974) studied 100 cases of police files, and hypothesized "that completeness of report decreased as a function of the increasing emotionality of the crime."

Deffenbacher (1983, 1991) appealed to the Yerkes-Dodson Law when explaining the effects of arousal on identification. The Yerkes-Dodson Law states that stress or arousal demonstrates an inverted U-shaped relationship with the identification accuracy. Low

levels of arousal, such as when waking up, produce low attentiveness; moderate levels of arousal serve to heighten perceptual and attentiveness skills; and, higher levels, such as that felt by an individual under extreme danger or duress, debilitates perceptual skills.

Hollien (1990) studied whether the victims of crime can perform better in an aural-perception experiment than those people who are not aroused in the same manner. To discover if stress or arousal influenced accuracy levels, Hollien screened (by use of a polygraph) female subjects for their potential to sensitivity to stressors. He accepted the 20 most susceptible to stress and the 20 least susceptible as subjects. The first group (the 20 most susceptible) he subjected to a video of 10 minutes of violent stimuli. During that time, a male voice read a threatening commentary. Those in the control condition (the 20 least susceptible) watched a video of a pastoral scene (horses being exercised). During the presentation of that video, a male voice read a neutral commentary. Hollien monitored the subjects' responses by means of a polygraph test. The identification task (of the male voice) revealed that the females who were more susceptible to stress were better at the identification task than those who were not aroused. Furthermore, the emotionally aroused women sustained better identification scores over time. Hollien states that "other things being equal, fear/stress/arousal can improve a person's ability to make the relatively complex judgments required for auditory speaker recognition" (Hollien, 1990, p. 202).

Although Hollien's study does provide a starting place for further research, the study itself is flawed. For example, the subjects should not have been split up into those most susceptible and those not; the subjects should have been randomly assigned. In a true forensic situation, the victim could be in either of these two groups. Also, Hollien does not state in his brief summary if anyone in the two groups had ever been a victim of violent crime. A person who had been a victim of violent crime would be more likely to be susceptible to stress than someone who had never been a victim. Furthermore, Hollien does not provide information as to exactly what stimuli was presented on the video for the stress group.

Read and Craik (1995) investigated the effects of emotionality in voice recognition. They recorded male voices speaking an emotional utterance (a call for help) and several non-emotional utterances of the same length. Each sentence was uttered in a tone appropriate to its content. The subjects were not informed of the memory test so incidental memory was also tested. One week later, the subjects were again tested on their ability to choose from a lineup of six voices which one had spoken the emotional utterance. Their recognition accuracy was 20% (chance was 17%). Their data offered "no suggestion that voices were differently memorable depending on whether they made emotional rather than unemotional statements" (Read and Craik, 1995, p. 14).

The current experiment sought to integrate many aspects of the above previous research. In studying the effect of arousal on earwitness identification, I presented a target voice uttering a series of phrases to a group of subjects in one of three voice tones (Positive, Neutral, and Hostile). All subjects were tested immediately after hearing the voice, and a subgroup of subjects were also tested one week later (the delayed recall subcondition) to ascertain if and how their accuracy on the identification task and other tasks might change with time. I used the lineup procedure given in Broeders and Rietveld (1995) since they gave extensive instructions for the lineup construction, and their procedure is used by certain police forces. To insure a forensically relevant experience, I tested incidental (as opposed to intentional) recognition.

Subjects were involved in a cover task when they heard through a loudspeaker the target voice uttering a series of phrases in one of the three tones. A loudspeaker was used instead of a live target so the subjects would not have any visual cues to interfere with their identification task. Subjects were then asked to identify the voice from a voice lineup consisting of the target voice and five foil voices, which corresponded to the recommendations given by Broeders and Rietveld. All subjects were asked to make height and weight estimates of the voice they heard. Voice similarity was also measured. Eight accurate identifications were made. The only significant results which were obtained were for height estimates between the Neutral Tone and Hostile Tone conditions

and between the Neutral and Positive Tone conditions. The tone of the voice and the level of arousal had no significant effect on identification accuracy.

2 Methods

2.1 Participants

Thirty-two undergraduate females and seven undergraduate males from Cornell University participated in this experiment. They were recruited from various linguistics and psychology classes and received extra credit for their participation.

2.2 Materials

2.2.1 Target voice

The target voice recorded a series of 12 groups of utterances in tones ranging from neutral to hostile. A group of 15 disinterested listeners (nine male and six female) from a Cornell beginning composition class were asked to place each group of utterances on a continuum ranging from positive to hostile. The group which was judged most neutral were chosen for the neutral utterances, those that fell on the positive end of the scale were taken as the positive group and finally those that were judged most hostile were used for the hostile group. Table 1 presents the results.

Utterance	Rating						
	-3	-2	-1	0	1	2	3
	(hostile)						(non-hostile)
hostile	4	3	4	2	2	0	0
neutral	1	1	5	5	2	1	1
positive	2	1	2	3	1	3	5

Table 1. Voice tone ratings.

2.2.2 Voice Lineup

Thirteen speakers (target and foils) were recorded in an IAC (Industrial Acoustic Corporation) sound-proof booth. The speakers' descriptions were recorded on a Carver TD-1700 tape deck. They were told to describe a postcard (*Montahago Valley Farm* by an anonymous 19th century artist) which depicted a rural farm scene. The descriptions were then digitized from a Panasonic SV-3200 tape deck into the XWaves 5.2 program. The first 30 seconds of their descriptions were used in the lineup since it was this part of all the descriptions that contained the fewest hesitations and the most fluent descriptions. Five seconds of silence was placed between each speaker's description. Finally, five different lineups were constructed and were recorded onto a DAT tape from a Sony DAT DTC-790 player.

The initial lineup consisted of 13 voices, including the target voice. These 13 voices were pretested to 15 disinterested listeners (seven male and eight female) in an introductory composition class at Cornell University. Before the pretest, the listeners were told to write down the voice or voices of those they thought might have committed a crime. Any voice which was chosen by three or more students was discarded from the lineup (two voices were chosen three times and three voices were chosen five times--the students could make more than one choice). The other voices were either not chosen at all, or chosen only once or twice.

In the end, five voices were discarded and eight voices were chosen for the lineup. The eight voices were then given to a group of disinterested listeners (another composition class at Cornell University) who were asked to rate the distinctiveness of each voice on a 1-10 continuum, with 10 as the most distinctive. They were not given any definition of distinctiveness, but were told to use their own definition. The listeners were asked only to rate each individual voice, not compare them. Table 2 presents these distinctiveness ratings (S1 is the target voice).

Rating	1	2	3	4	5	6	7	8	9	10
Voices				S8/S3	S7/S6	S5	S2		S1	S4

Table 2. Distinctiveness ratings for lineup voices.

Five different lineups were made from the eight voices, each lineup containing six voices (including the target voice). The voices were randomly chosen from the eight available and randomly assigned a place in the lineup (with the only stipulation being that the target voice would not be first or last to avoid initial or recency effects). Table 3 presents the distribution of voices in the five lineups (A1, B1, C1, D1, and E1 refer to the re-randomized lineups constructed for those in the delayed recall subcondition).

Lineup	Voices in lineup (S1 = target voice)					
A	S5	S4	S8	S7	S1	S6
A1	S4	S6	S5	S1	S8	S7
B	S3	S7	S5	S8	S1	S6
B1	S7	S1	S3	S6	S5	S8
C	S3	S4	S8	S6	S1	S5
C1	S4	S1	S8	S5	S3	S6
D	S8	S1	S3	S4	S6	S7
D1	S6	S3	S1	S7	S8	S4
E	S2	S3	S1	S4	S6	S5
E1	S3	S6	S5	S1	S4	S2

Table 3. Distribution of voices in lineups.

2.3 Procedure

Subjects were tested either individually or in groups no larger than six. Subjects were seated in positions approximately equidistant from the speaker over which they heard the target voice.

After the subjects signed a consent form, they were given an Emotional Intensity Survey, developed by Bachorowski and Braaten (1994). Subjects were told that this part of the experiment was part of a larger study and that they were among many who were given the surveys.

After the subjects finished the survey, they were given a one-paragraph passage from *Jane Eyre* and told to cross off every letter "l" as quickly as they could. They were aware that they were being timed. The purpose of the timing test was to ascertain if the voice the subjects heard had any emotional effect on them. If there was a significant difference between the initial timing test and the one which occurred after they heard the target voice, then some emotional effect might be inferred.

As a cover task, the subjects were given a page of anagrams to solve. They were told they had five minutes to solve these word puzzles. At the end of their allotted time, they heard the target voice (through a speaker) utter the following phrases:

1. Where are you?
2. I need to see you.
3. It's urgent.
4. It can't wait.
5. I need to see you right away.

The subjects were in one of three groups. Subjects heard the voice speaking the above utterances in either a positive tone, a neutral tone, or a hostile tone.

Subjects in each condition were further split into two subgroups (see Table 4). One group of subjects was tested immediately after they heard the voice, and the other group was tested both immediately after they heard the voice, and also one week later (the delayed recall subcondition).

Positive Tone				Neutral Tone				Hostile Tone			
No Delay		Delay		No Delay		Delay		No Delay		Delay	
F	M	F	M	F	M	F	M	F	M	F	M
6	0	4	2	6	0	3	3	6	0	7	2

Table 4. Division of subjects into conditions and groups.

Immediately after the subjects heard the voice, the experimenter entered the room and informed the subjects of the true nature of the experiment. This was done because while some subjects may have guessed the real purpose of the experiment, some had not, and informing them of the true nature of the experiment would bring all the subjects to relatively the same level of knowledge. They then were given another timing test (the same task but a different passage) and then presented with the lineup. The purpose of the timing test was to ascertain if the voice the subjects heard had any emotional effect on them.

After the timing test, subjects were asked to estimate the height and weight of the person whose voice they had heard previously. They were not given another opportunity to hear the voice. After they completed that task, subjects were given one of the five previously created lineups (the experimenter chose the lineup). In addition, they were asked to give the reasons for their decision, their confidence of their judgment, and finally they were asked to list any personality or physical characteristics they wanted to ascribe to the voice they chose from the lineup. Table 5 lists the distribution of lineups to subjects.

Lineup	Positive Tone	Neutral Tone	Hostile Tone
A	1	5	2
B	1	3	2
C	4	2	4
D	1	1	4
E	6	1	2

Table 5. Lineup distribution per condition for participants.

In the next task, the subjects were asked to choose from a number of adjectives which described their feelings at the time they heard the voice, and then were asked to write any adjectives not on the list. This task simulated the Multiple Adjective Affective List developed by Zimmerman (1960). The purpose of this was to gather more information about the subjects' arousal state. Finally, subjects were asked to rate their arousal state on a continuum from 1-10.

After a retention interval of one week, the subjects in the delayed recall subgroup returned to the lab. They again were asked to give height and weight information about the voice they heard, their decision-making process, their confidence of their judgment, any personality or physical characteristics of the voice, and their emotional state at the time they heard the voice. The lineup they heard the second time consisted of the same voices as the first lineup, but the voices were again randomized (this lineup is designated by the number 1 after the lineup, such as lineup A1). The target voice still did not appear in the first position in the lineup.

Acoustical measurements of voice similarity were performed, based on Walden et al. (1978). In the current study, each speaker recorded five tokens of the word "beans." For the middle three tokens, word duration, vowel duration, nasal duration, fricative duration, the first three formants, and the mean fundamental frequency were measured.

Finally, perceptual judgments were made on paired voice samples. All possible combinations of the target and foil voices saying "beans" were given to 17 listeners (a third composition class at Cornell University), who were asked to rate (on a 10 point scale) how similar the voices were to each other. The listeners heard only one repetition.

3 Results

3.1 Identification accuracy

Eight correct identifications or "hits" were obtained out of a possible 60 identifications (see Table 6). A hit is defined as any time a subject correctly identified the target voice from the lineup.

Condition	Lineup	Hits	Lineup	Hits
Positive	E	1	E1	2
Neutral	B	1	----------	----------
	D	1	D1	0
	C	1	C1	0
Hostile	D	1	D1	1

Table 6. Hit number per condition and lineup.

As can be seen from Table 6, hits occurred with all lineups except A, and most of the hits occurred in the Neutral Tone condition.

3.2 Timing test

The means for each variable in the experiment were calculated for each condition as well as within condition. For the timing test, the term "hit" refers to every instance in which the subject crossed off an "I" from the page. The means did not differ much across conditions. For the Positive Tone condition, the mean of the pretest timing test was .49

hits per second and for the posttest the mean was .57 hits per second. For the timing test, the mean for the Neutral Tone condition was .52 hits per second (pretest), and .53 hits per second (posttest). In the Hostile Tone condition, the mean of the pretest timing test was .55 hits per second, and the posttest mean was .57 hits per second. A 3 x 2 (tone, delay) analysis of variance (ANOVA) was performed on subjects' pre- and post-timing tests. The effect was not significant [F(1,36)=.93, p=.405].

3.3 Height estimates

The mean height estimates (see Figure 1) for the three conditions were very similar across the Positive, Neutral, and Hostile Tone conditions, 71.0 inches (SD=2.63), 69.0 inches (SD=2.13), and 71.0 inches (SD=1.74) respectively. Height estimates also were very similar in the delayed recall subcondition (see Figure 2), the mean height estimate in the Positive Tone condition was 70.1 inches (SD=2.23) in the no delay condition, and with a one-week delay 70.1 inches (SD=2.22). The mean height estimate for the Neutral Tone condition was 70.6 inches with no delay (SD=1.21) and 71.0 inches (SD=1.79) in the delayed recall subcondition. For the Hostile Tone condition, the mean height estimate was 70.5 with no delay (SD=1.67), and 70.3 inches (SD=2.13) in the delayed recall condition.

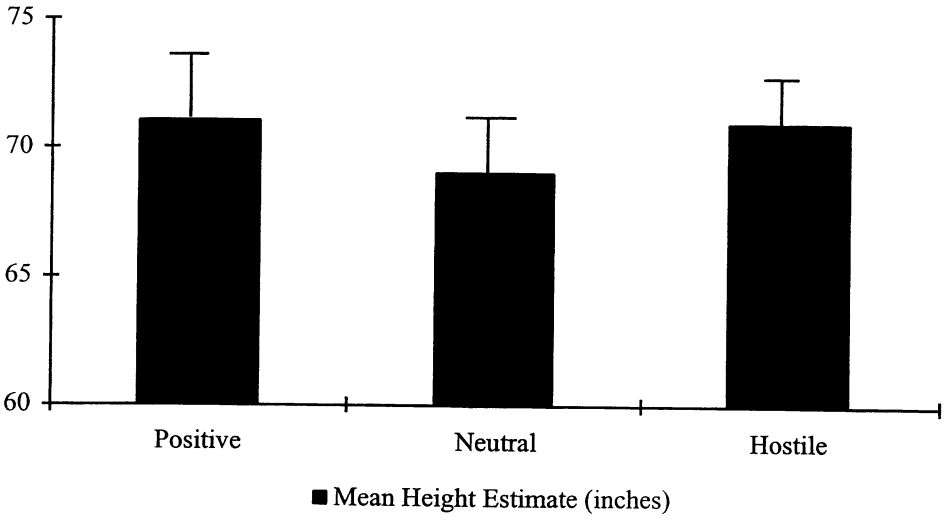

Figure 1. Mean height estimates per condition.

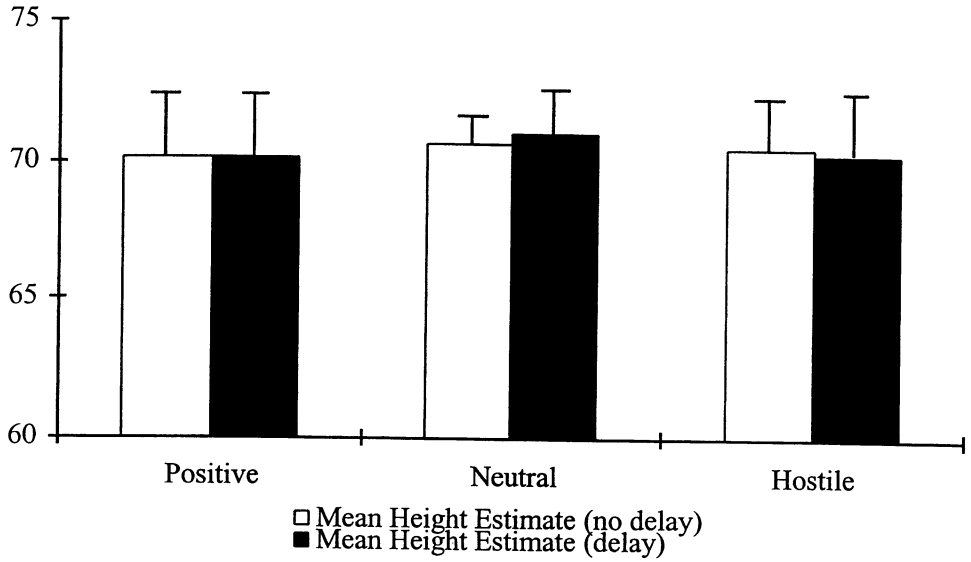

Figure 2. Mean height estimates in delayed recall subconditions.

For height, an ANOVA performed between all tone conditions yielded significant results [F(2,36)=3.98,p=.0273]. A post-hoc Student-Newman-Keuls test revealed a significant difference between not only the Neutral Tone and Hostile Tone conditions, but also the Neutral Tone and Positive Tone conditions.

A 3 x 2 (tone, delay) analysis of variance was performed which compared immediate estimates of height with the estimates given a week later. No significant results were obtained [F(2, 18)=.53,p=.598.] No significant results were obtained in comparing only the delayed recall subconditions (meaning those subjects tested not only immediately but also a week later) for height: Positive Tone subcondition, t(5)=.00,p=1.0; Neutral Tone subcondition, t(5)=-.79,p=.465; and Hostile Tone subcondition, t(8)=.55,p=.594. (See Figure 2.)

Measurements of the height difference between the estimates and actual height of the target yielded similar results across the Positive, Neutral, and Hostile Tone conditions, being +6.0 inches, +4.0 inches; and +5.9 inches respectively. Across the board in the delayed recall subconditions, the mean height differences were also very similar. The mean height difference for the Positive Tone subcondition was +5.1 inches in the no-

delay test, and +5.1 inches after one week; for the Neutral Tone subcondition, the height difference was +5.6 inches in the no-delay test, and +6.0 inches after one-week; for the Hostile Tone, +5.5 inches in the no-delay test and +5.3 inches after one week.

3.4 Weight estimates

Mean weight estimates (see Figure 3) across the Positive, Neutral, and Hostile Tone conditions were again similar, being 175 lbs (SD=19.93), 176.5 lbs (SD=20.08), and 182 lbs (SD=17.23), respectively.

In the delayed recall subconditions, the mean weight estimates did not change significantly. For the Positive Tone subcondition, the mean weight estimate was 169.3 lbs (SD=22.15) and after one week the mean was 165.0 lbs (SD=27.39); for the Neutral Tone subcondition, the mean weight estimate was 185.8 lbs (SD=24.58) in the no delay test and 188.7 lbs (SD=24.87) in the delayed recall subcondition; finally, for the Hostile Tone subcondition the mean weight in the no delay condition was 183.0 lbs (SD=18.70), and in the delayed recall subcondition the mean weight was 172.0 lbs (SD=14.17). (See Figure 4.)

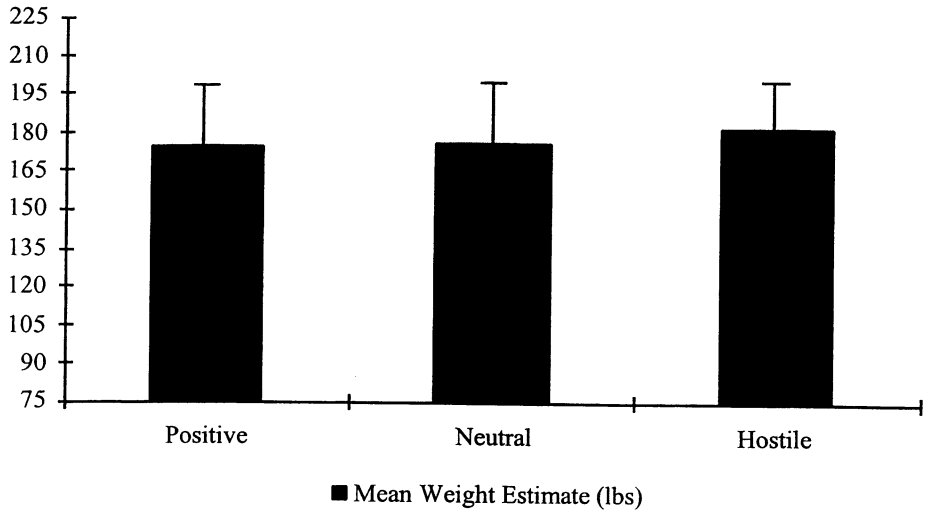

Figure 3. Mean weight estimates per condition.

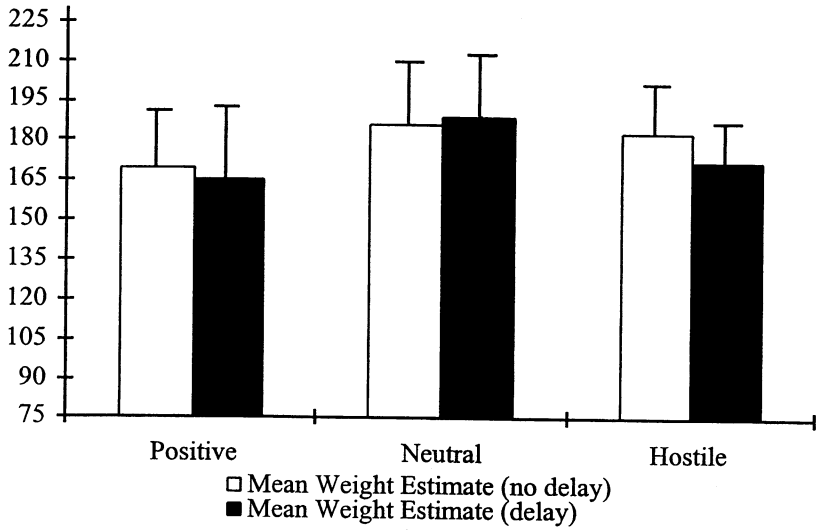

Figure 4. Mean weight estimates in delayed recall subconditions.

For weight, a one-way analysis of variance was performed for all tone conditions, and no significant results were obtained [$F(2,36)=.2510, p=.77$]. A 3 x 2 (tone, delay) analysis of variance was performed on weight estimates given immediately and those given one week later, and again no significant results were obtained [$F(2,18)=2.86, p=.084$].

Mean weight differences were similar for the Positive, Neutral, and Hostile Tone conditions, being 19.1 lbs, 21.0 lbs, and 16.6 lbs, respectively. For the Positive Tone subcondition the no-delay absolute mean weight difference was 24 lbs, and for the delayed recall group, 28.5 lbs. For the Neutral Tone subcondition, the absolute mean weight difference was 19.2 lbs, and after one week 18.67 lbs. Finally, for the Hostile Tone subcondition, the absolute mean weight difference was 15.5 lbs in the no-delay condition, and 17.8 lbs in the delayed recall condition. Table 7 reports both positive and negative mean weight differences for all conditions and subconditions.

Subject Group	Positive		Neutral		Hostile	
	+	-	+	-	+	-
all immediate	8.33	22.66	22.50	20.07	16.25	18.40
subcondition immediate	8.33	20.57	40.00	20.00	20.00	15.00
subcondition delayed recall	10.00	32.00	6.00	25.00	------	20.00

Table 7. Mean weight differences for subjects in all conditions.

3.5 Arousal rates

The mean arousal rate (see Figure 5) was similar across the Positive, Neutral, and Hostile Tone conditions, being (on a 10 point scale) 5.5 (SD=2.54), 5.25 (SD=2.70), and 6.1 (SD=2.76), respectively. The mean arousal rate for the Positive, Neutral, and Hostile Tone delayed recall subconditions was also very similar. In the Positive Tone the mean arousal difference was 5.5 (SD=2.59) in the no-delay test and 5.67 (SD=2.50) in the delayed recall test. For the Neutral Tone subcondition, the mean arousal difference was 5.33 (SD=3.08) in the no-delay test and 5.16 (SD=2.64) in the delayed recall test, and for the Hostile Tone subcondition, the mean arousal score was 6.0 (SD=3.00) in the no-delay test and 6.7 (SD=1.40) in the one-week delayed recall test (see Figure 6).

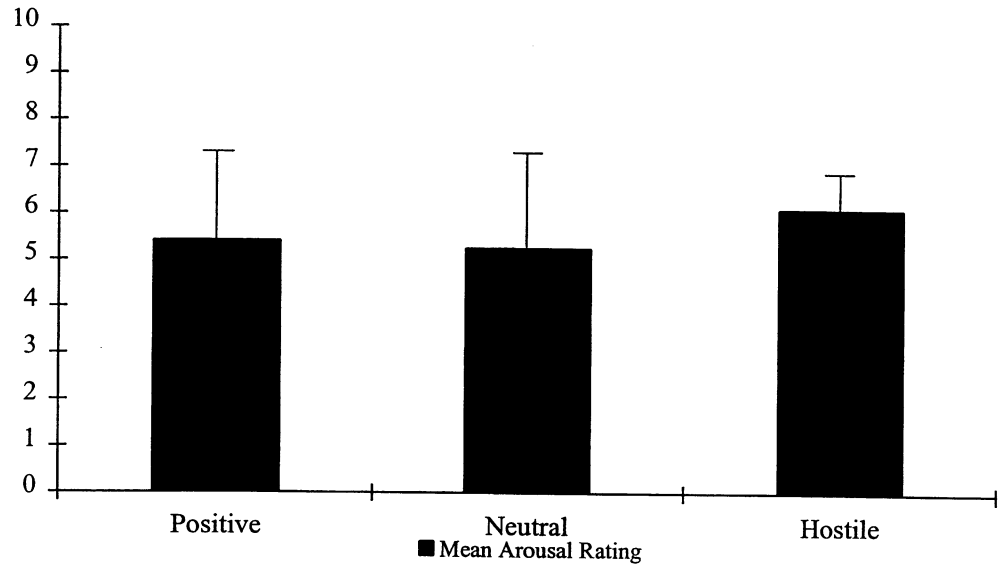

Figure 5. Mean arousal rating per condition.

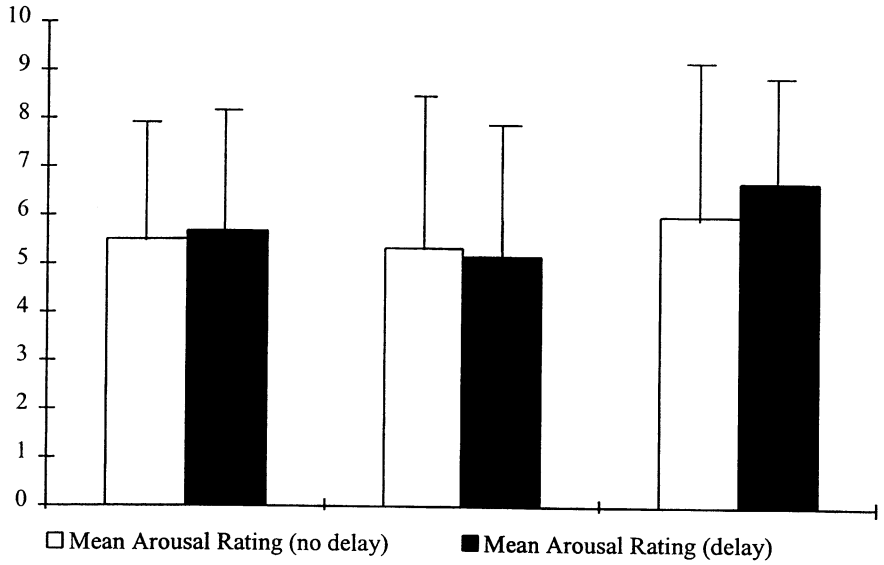

Figure 6. Mean arousal rating in delayed recall subconditions.

A one-way ANOVA was performed, and no significant results were found between all Tone conditions and subjects' arousal rating, $F(2,36)=.1427, p=.86$]. A 3 x 2 (tone, delay) analysis of variance was performed which compared arousal ratings for the delayed recall subcondition. No significant results were obtained [$F(2,18)=2.00, p=.165$]. Paired t-tests

were performed for all delayed recall subconditions and their arousal ratings, with no significant results obtained in any subgroup: Positive Tone subcondition, $t(5)=-.42, p=.695$; Neutral Tone subcondition, $t(5)=.31, p=.77$; Hostile Tone subcondition, $t(8)=-2.04, p=.076$.

3.6 Confidence rates

Mean confidence ratings (see Figure 7) differed little across the Neutral and Hostile Tone conditions, but both differed from the Positive Tone condition, being 5.25 (SD=2.09), 5.27 (SD=1.80), and 3.42 (SD=1.93), respectively.

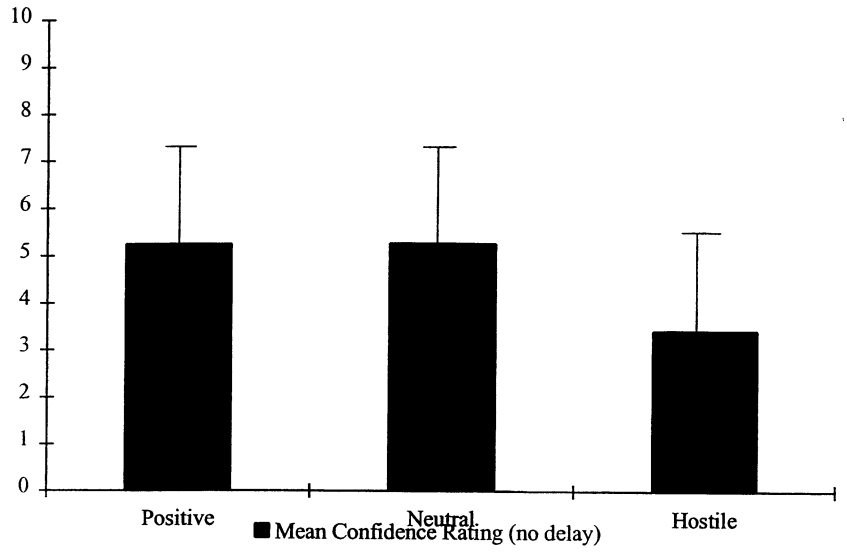

Figure 7. Mean confidence rating per condition.

For all subconditions, confidence ratings dropped in the delayed recall subconditions (see Figure 8). In the Positive Tone subcondition, the mean confidence rating was 5.67 (SD=2.34) in the no-delay subcondition, and 4.5 (SD=2.07) in the delayed recall group. For the Neutral Tone subcondition, the mean confidence rating was 5.5 (SD=2.07) in the no-delay test and 5.3 (SD=2.58) after one week. Finally, for the Hostile Tone subcondition, the mean confidence rating was 5.2 (SD=1.80) in the no-delay test, and 4.2 (SD=1.92) after one week.

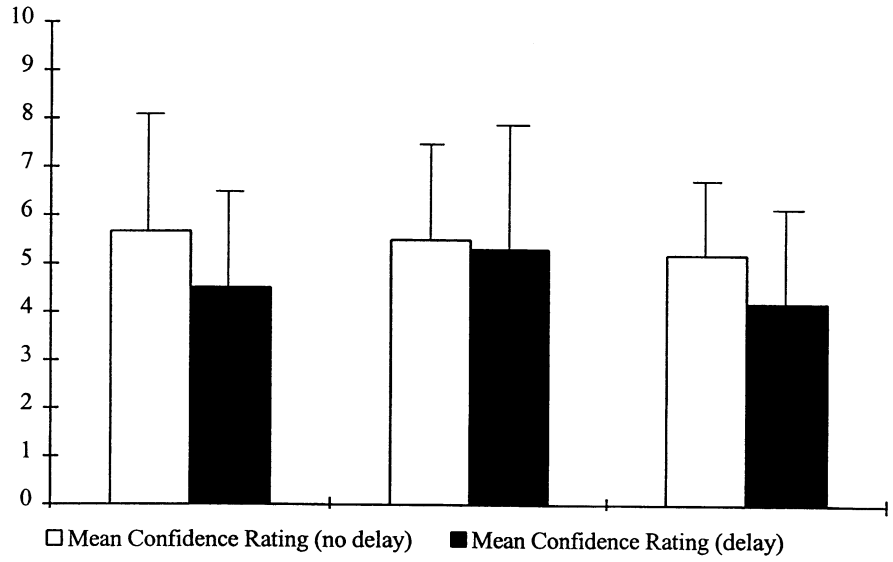

Figure 8. Mean confidence rating in delayed recall subconditions.

A one-way ANOVA was performed for all groups and their confidence ratings. For all groups, no significant results were obtained, [$F(2,36)=.0715, p=.93$]. A 3 x 2 (tone, delay) analysis of variance was performed which compared confidence ratings given immediately and those given a week later. No significant results were obtained [$F(2,18)=.48, p=.625$].

Paired t-tests were conducted for each delayed recall subcondition, and no significant results were obtained in any subcondition: Positive Tone, $t(5)=1.94, p=.110$; Neutral Tone, $t(5)=.21, p=.842$; Hostile Tone, $t(8)=1.26, p=.244$.

3.7 Emotional intensity results

The mean EIS (Emotional Intensity Survey) rating (see Figure 9) was calculated for each condition, with little difference across the Positive, Neutral, and Hostile Tone conditions: 106.4 (SD=13.53), 102.2 (SD=14.96), 114.3 (SD=17.78), respectively.

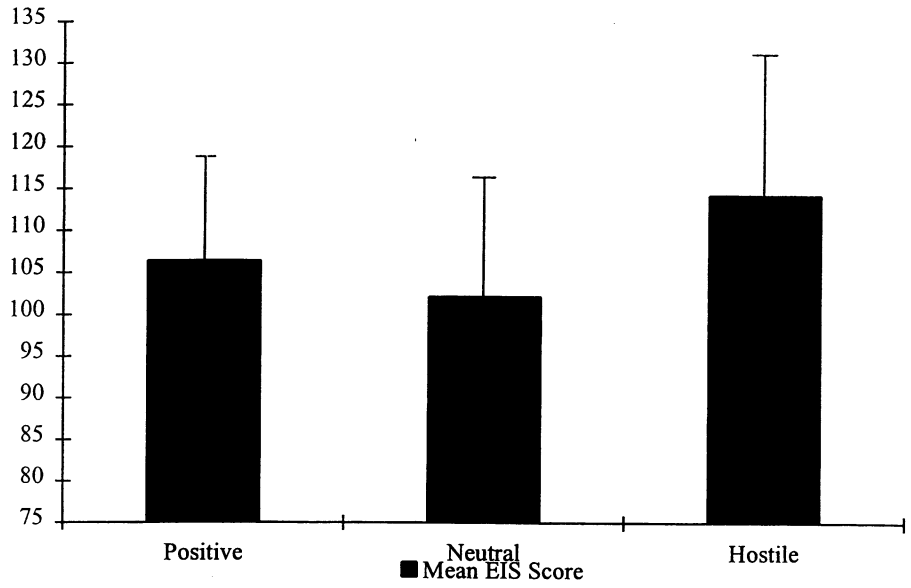

Figure 9. Mean EIS score per condition.

A Pearson's Correlations Coefficient was calculated for each Tone condition for the arousal, confidence, and EIS variables. No significant effects were found for any correlation.

As stated above, the subjects were asked to mark adjectives which described their emotional state when they heard the voice through the speaker. The first ten adjectives were classified by Zimmerman (1960) as anxiety adjectives, and the last ten as non-anxiety adjectives. Table 8 presents the results.

Adjective	Neutral	Positive	Hostile
afraid	0	1	2
fearful	1	0	0
frightened	1	0	3
nervous	3	3	4
tense	5	2	5
terrified	0	1	0
upset	0	1	0
worried	2	0	1
excited	5	3	5
provoked	2	3	3
calm	4	4	1
secure	3	2	4
steady	4	2	3
undisturbed	2	1	1
tranquil	1	1	0
serene	0	1	0
quiet	2	2	0
untroubled	2	3	5
relaxed	2	3	3
unexcited	5	2	3

Table 8. Adjective descriptions per condition.

The subjects listed 56 anxiety adjectives, and 66 non-anxiety adjectives, and the number or kind of adjectives chosen (anxiety or non-anxiety) did not differ much across tone conditions. Subjects in the Positive Tone condition listed 14 anxiety adjectives and 21 non-anxiety adjectives. Subjects in the Neutral Tone condition listed 19 anxiety adjectives and 25 non-anxiety adjectives. Finally, those in the Hostile Tone condition listed 23 anxiety adjectives and 20 non-anxiety adjectives. In the free response portion of

this task, subjects from all groups listed "curious," "surprised," and "startled" the most. For those subjects in the delayed recall subcondition there was little difference in the adjectives they circled or listed immediately after the test and those they listed one week later.

3.8 Confusion matrix

A confusion matrix was constructed for each of the eight voices across all three conditions (Positive, Neutral, and Hostile). Table 9 presents the results. As can be seen from the table, the target voice (S1) was most often confused with S6. (D.R. = distinctiveness rating.)

Lineup	S6	S3	S5	S7	S8	S4	S2
A	0	0	1	1	4	2	0
A1	0	0	1	0	1	1	0
B	1	0	0	3	1	0	0
B1	1	0	2	0	1	0	0
C	5	0	3	0	0	0	0
C1	3	0	1	0	0	0	0
D	1	2	0	1	0	0	0
D1	0	2	0	2	0	0	0
E	1	5	1	0	0	0	0
E1	0	1	0	0	0	0	0
Total	12	10	9	7	7	3	0
D.R.	5	4	6	4	4	10	7

Table 9. Confusion of target voice with lineup voices.

It can be seen from Table 9 that voice S6, which was most confused with the target voice, was not rated as the most distinctive voice, while the voice with the highest distinction rating (S4) was almost least confused with the target voice. The target voice

(S1) has the second highest distinction rating, and if subjects were choosing voices from the lineup based on how distinctive the voice sounded, it would seem more likely that they would more often confuse S4 with S1 than S6 with S1. That was not the case here.

3.9 Similarity measurements

Table 10 reports the averages of the voice similarity measurements. (vd=vowel duration; nd=nasal duration; fd=fricative duration; wd=word duration (all measured in milliseconds). F1, F2, F3, and F0 are all measured in Hertz.) Data from speaker S5 is not available.

	Speaker S2	Speaker S3	Speaker S4	Speaker S6	Speaker S7	Target S1	Speaker S8
vd	283	270	362	247	238	238	215
nd	178	205	169	147	173	303	210
fd	156	170	130	116	137	217	157
wd	634	684	680	611	584	810	640
F1	355	408	394	327	341	303	341
F2	2193	2078	2232	2188	1995	2217	2207
F3	2842	2622	3020	3102	2799	2669	3006
F0	112	120	111	102	99	137	142

Table 10. Acoustic similarity measurements for speakers in lineup.

Although voice S6 was confused most often with the target voice (S1), the above acoustical analysis shows that they do not share many distinct similarities. A Hierarchical Cluster Analysis reveals that voice S3 was most acoustically similar to the

target voice, despite the fact that voice S3 was the second most confused with the target and in fact was rated as least distinctive among the voices in the lineup.

Furthermore, perceptual judgments of the voice pairs do not clarify the confusion. In none of the pairs rated eight (out of ten) or above (indicating high similarity) did the target voice appear. In fact, in only one of the pairs (rated six) did the target voice appear (when paired with voice S8—even though voice S8 was rarely confused with the target voice, and rated as only third most distinctive of all the voices). Voices S4 and S6 were rated as highly similar to each other, even though it was voice S6 which was most often confused with the target voice S1, and voice S4 which was rated as most distinctive. Table 11 presents the similarity matrix. Again, data from voice S5 is not available.

	S1	S2	S3	S4	S6	S7	S8
S1	*	2	3	4	3	1	6
S2	2	*	4	8	4	4	3
S3	3	4	*	8	3	3	6
S4	4	8	8	*	5	3	3
S6	3	4	3	5	*	4	3
S7	1	4	3	3	4	*	2
S8	6	3	6	3	3	2	*

Table 11. Similarity judgments for paired voices.

3.10 Decision factors

When asked what factors led them to their decision, there were no significant differences between the different tone conditions. Many subjects used the process of elimination to reach their decisions, trying to match the voice they heard over the speaker to one in the lineup by a specific voice quality such as pitch, tone, intonation, intensity,

and rate of speech. There was little change in one week in subject response from those who were tested immediately and those in the delayed recall condition.

Subjects were also asked to list physical and personality characteristics of the voice they chose from the lineup. The most common answers commented on the voice's appearance, such as "attractive," or "average looking." Some subjects listed the race of the voice (all voices were white males). Interestingly enough, some subjects listed characteristics such as "brown hair," "a runner," "rugged and likes to be outdoors," "organized," and "clearheaded."

4 Discussion

The most prominent result of this study was that tone of voice had no effect on identification accuracy. That is, identification accuracy was poor no matter what voice tone (Positive, Neutral, Hostile) subjects heard.

Only eight correct identifications were made on the target voice. Since not all the hits came from the same lineup, it can be assumed that the construction of the lineup did not significantly affect the selection of the target voice by the subjects. It could be, in future experiments of this nature, that the lineup should be altered to include fewer distractor foils. Bull and Clifford (1984) found that accuracy increased when the number of foils decreased. In their study, they found a 48% accuracy rate with six in the lineup. The identification accuracy rate in the present study was well below the results found by Bull and Clifford. Their study also indicated that the position of the target in the lineup did not influence identification accuracy, and the results of the present study mimic, at least in this regard, those found by Bull and Clifford.

Furthermore, the most hits occurred in the Neutral Tone condition, which suggests that although voice tones did not have a significant effect on identification, voice tones with a heightened sense of emotion conveyed less relevant information for identification than the Neutral Tone condition. Of course with such a small hit rate, no definite claims can be made, but further experimentation is warranted.

In two out of the eight hits, the subjects who identified the voice correctly immediately after hearing it could not in fact identify the voice a week later. Further examination of their records indicated that they chose the same number in the lineup, regardless of which voice corresponded to that number. Only two of the subjects could re-identify the target voice a week later, and one subject could not identify the target voice initially, but could a week later (after no further exposure than that initial time). The low levels of identification accuracy after a one week delay are not unexpected, considering those obtained by others who have studied the effects of delay on identification accuracy (see McGehee, 1937, 1944; Clifford, Rathborn, and Bull, 1981; Clifford and Denot, 1982).

The construction of the experiment, however, might have inadvertently facilitated the lack of hits of the target voice, especially from those subjects in the delayed recall subcondition. Since the subjects were tested both immediately and one week later, they were actually participating in a multiple confrontation experiment. Multiple confrontation refers to repeatedly exposing subjects to lineups (see Broeders and Rietveld, 1995). If the subject chooses the target voice in all the lineups, then it can be assumed that the subject can truly identify the target voice. Yet the danger from multiple confrontation testing is that, after exposure to the first lineup, the subject will try and identify from future lineups a particular voice that stood out from the first lineup. In effect, the subject is identifying a voice heard earlier, but that voice is not necessarily the target's voice.

The testing of the delayed recall subconditions for the different voice tones (Positive, Neutral, and Hostile) can be considered multiple confrontation testing, and in effect the subjects might have been trying to match a voice they heard from the first lineup with a voice in the second lineup, even if that voice was not the target voice. In fact, out of the 21 subjects tested both immediately and one week later, three chose the same number in each lineup, and eight chose the same voice in each lineup, regardless of the fact that only once was the voice chosen actually the target voice.

Since this study also tested the effects of incidental recognition, the low level of identification accuracy was not an unexpected finding, though identification accuracy was higher in other studies. Saslove and Yarmey (1980) found a 62% identification accuracy rate, and Geiselman and Bellezza (1977) also found speaker recognition to be better than chance.

It can be seen also that the tone of voice did not affect the selection of the target voice by subjects. This could be due to several reasons. One reason is the tone of the voice itself. Only subjective judgments were used to classify a range of utterances into three different groups. The Positive and Neutral Tone voices were rated very similar, and could have easily been switched with the same results. The Hostile Tone, while rated the most hostile from the choices, might not have been perceived as hostile by the subjects. Furthermore, Saslove and Yarmey (1980) found that a change in tone resulted in reduced identification. Read and Craik (1995) found no suggestion that voices were differently memorable depending on the tone of voice heard or the emotionality of the statement.

Furthermore, the subjects heard the voice over a loudspeaker, and not in person. They were engaged in a cover task (the anagrams) and the voice was supposed to induce some level of arousal, but it could be the case that hearing the voice over a loudspeaker did not induce any sufficient level of arousal in them which would affect their identification. As can be seen from both the timing test and the adjective listing, the subjects were not in any way significantly aroused by hearing the voice. Further studies might focus on using an actor's voice, changing the methods of arousal, or changing the method by which arousal is measured. The additional challenge then is to find an effective method which is also forensically relevant.

As stated in the results, the target voice was rated as second most distinctive of all the voices in the lineup, yet the hit rate seems to be at odds with this rating. Furthermore, the voice which was rated most distinctive (S4) was not chosen as often as other voices, despite the fact that it was included in four of the five lineups (all except Lineup B--which was presented least often). In fact, that voice was the third least chosen. The voice similarity measurements yielded no further illumination on this subject either. S6 (who

was most often confused with target voice S1) was not very similar to S1 in acoustic measurements, although not one of the voices stands out as being similar to S1.

As to how they made their decision of which voice to select from the lineup, most subjects listed a sort of process of elimination, whereby they latched on to some particular feature of the target voice and then tried to match that feature to a voice in the lineup. Most often that feature was a particular pronunciation the target voice made, such as a vowel sound. Often, the subjects tried to match the intensity of the voices, the tone of the voice, or the rate of speech. While the answers were diverse, apparently they did not help the subjects enough to correctly identify the target voice.

Finally, when the subjects were asked to ascribe personality characteristics to the voice they heard, they listed many surprising qualities, such as "brown hair" (which was correct--all the males participating had brown or dark hair). Appearance was most often noted, with positive aspects of appearance listed most often. This could be a desire to be polite, since the subjects were overwhelmingly female. Some characteristics, however, pertained to occupations or hobbies. Female subjects listed more occupational features than male subjects. This, too, could be due to sociological factors.

Height estimates were significant only for comparison between the Neutral and Hostile and Neutral and Positive Tone conditions, yet not between the Positive and Hostile Tone conditions. This significance did not pertain to the accuracy of the height estimates, but to the Tone conditions. The Neutral Tone of voice produced the closest estimate to the actual height, and the Hostile and Positive Tone conditions both produced a slightly higher estimate of height. No significant results were found between the delayed recall subconditions but in the Neutral Tone condition the height estimate in the delayed recall test was barely larger than in the no-delay test. This could be due to the fact that any perceived emotion the subjects felt decreased after a week and allowed them to more accurately estimate the height. In the case of the Positive and Neutral Tone conditions, the subjects could merely be remembering their response from the first time.

For the weight estimates, those in the Hostile Tone condition were closest in their estimate of the target's actual weight, with the Positive and Neutral Tone estimates

somewhat lower. This was an expected condition, because it was hypothesized that the Hostile Tone would produce a larger weight estimate. Whether the Hostile Tone actually produced a more actual estimate is unclear, however, since for those in the delayed recall condition the weight estimates went down. The weight differences for each group, however, were very large, much larger than what others in the field have reported, especially Lass and his colleagues (see Lass et al., 1978; Lass and Brown, 1978; Lass et al., 1980).

Confidence ratings for the three groups were not significantly different, nor was there any significant difference between the confidence ratings for the delayed recall subconditions. The confidence ratings also did not correspond with the identification accuracy rate, since only eight hits occurred, and those who hit as well as missed the target voice reported high confidence levels. This is not an unusual occurrence (see Clifford, 1980; Bothwell, Deffenbacher, and Brigham, 1987; Cutler, Penrod, and Stuve, 1988). What is interesting to note is that the mean confidence levels for the Hostile and Neutral Tone conditions are close, while the Positive Tone condition confidence level is lower. In the delayed recall subconditions, the confidence level lowered after one week in all groups, especially the Hostile Tone. That was to be expected, however, since one week had elapsed between testing.

Arousal ratings for the three groups did fall into a predictable pattern. Although not statistically significant, the Hostile Tone group reported the highest arousal rate, followed by the Positive Tone group, and the Neutral Tone group last. What was not predictable was that in the delayed recall subconditions a higher arousal rate would be reported. This was not, however, born out in the adjective listing, since the subjects in the delayed recall condition listed less anxiety adjectives after a week.

Emotional Intensity Survey scores for subjects in each condition followed an unexpected pattern. Those placed in the Hostile Tone condition scored the highest overall in their EIS scores, while those in the Positive Tone condition scored in the middle, and those in the Neutral Tone condition scored the lowest overall on the EIS. Subjects were randomly placed in their respective groups. This division of scores and

conditions, however, did not significantly correlate with either arousal or confidence ratings.

First and foremost, this study showed that the tone of the target voice had no effect on the accuracy of identification. Height and weight estimates did not vary much between the different voice tones. In this case, the tone of voice did not significantly influence the ability of the subjects to make height and weight estimates which differed little across conditions. Arousal levels did not differ significantly, regardless of the voice tone heard. Confidence levels did not differ significantly regardless of voice tone heard, nor did confidence correlate with identification accuracy. Though the current study yielded few statistically significant results, it was forensically relevant, and further studies in this area can only shed more light on this field of forensic study.

5 References

Bachorowski, J. and E.B. Braaten. (1994) Emotional intensity: Measurement and theoretical implications. *Personal and Individual Differences* 17, 191-199.

Bothwell, R.K., Deffenbacher, K.A., and J.C. Brigham. (1987) Correlation of eyewitness accuracy and confidence: Optimality hypothesis revisited. *Journal of Applied Psychology* 72, 691-695.

Bricker, P.D. and S. Pruzansky. (1966) Effects of stimulus content and duration on talker and identification. *Journal of the Acoustical Society of America* 42, 1441-1254.

Broeders, A.P.A. and A.C.M. Rietveld. (1995) Speaker identification by earwitnesses. *Studies in Forensic Phonetics* 64, 24-40.

Bull, R. and B.R. Clifford. (1984) Earwitness voice recognition accuracy. In G.L. Wells and E.F. Loftus (Eds.), *Eyewitness Testimony* (pp.92-123). New York: Cambridge University Press.

Clifford, B.R. (1980) Voice identification by human listeners: On earwitness reliability. *Law and Human Behavior* 4, 373-394.

Clifford, B.R. and H. Denot. (1982) Visual and verbal testimony and identification under conditions of stress. Unpublished manuscript, North East London Polytechnic.

Clifford, B.R. and C. Hollin. (1981) Effects of the type of incident and the number of perpetrators on eyewitness memory. *Journal of Applied Psychology* 66, 364-370.

Clifford, B.R. and J. Scott. (1978) Individual and situational factors in eyewitness testimony. *Journal of Applied Psychology* 63, 352-359.

Cohen, J.R., Crystal, T.H., House, A.S., and E.P. Neuburg. (1980) Weighty voices and shaky evidence: A critique. *Journal of the Acoustical Society of America* 68, 1884-1886.

Cutler, B.L., Penrod, S.D., and Stuve, T.E. (1988) Juror decisionmaking in eyewitness identification cases. *Law and Human Behavior* 12, 41-55.

Deffenbacher, K.A. (1983) Eyewitness accuracy and confidence: Can we infer anything about their relationship? *Law and Human Behavior* 4, 243-260.

Deffenbacher, K.A. (1991) A maturing of research on the behavior of eyewitnesses. *Applied Cognitive Psychology* 5, 377-402.

Eyesenck, M. (1975) Arousal and speed of recall. *Journal of Social and Clinical Psychology* 14, 269-277.

Fay, P.J. and W. C. Middleton. (1940) Judgment of Krestschmerian body types from the voice as transmitted over a public address system. *Journal of Social Psychology* 12, 151-162.

Geiselman, R.E. and F.S. Bellezza. (1976) Long-term memory for speaker's voice and source location. *Memory and Cognition* 4(5), 483-489.

Geiselman, R.E. and F.S. Bellezza. (1977) Incidental retention of speaker's voice. *Memory and Cognition* 5(6), 658-665.

Gunter, C.D. and W.H. Manning. (1982) Listener estimations of speaker height and weight in unfiltered and filtered conditions. *Journal of Phonetics* 10, 251-257.

Hammersley, R. and J.D. Read. (1985) The effect of participation in a conversation on recognition and identification of the speakers' voices. *Law and Human Behavior* 9(1), 71-81.

Hollien, H. (1990) *The Acoustics of Crime.* New York: Plenum.

Komulainen, E.K. (1988) Subjective voice identification: The literal meaning of talking to yourself behind bars. *Alberta Law Review* 26, 521-547.

Kuehn, L. (1974) Looking down a gun barrel: Person perception and violent crime. *Perceptual and Motor Skills* 39, 1159-1164.

Kunzel, H.J. (1994) On the problem of speaker identification by victims and witnesses. *Forensic Linguistics* 1, 45-57.

Lass, N. J., Beverly, A.S., Nicosia, D.K., and L.A. Simpson. (1978) An investigation of speaker height and weight identification by means of direct estimations *Journal of Phonetics* 6, 69-76.

Lass, N.J., and W.S. Brown. (1978) Correlational study of speakers' heights, weights, body surface areas, and speaking fundamental frequencies. *Journal of the Acoustical Society of America* 63, 1218-1220.

Lass N.J., Philips, J.K., and C.A. Bruchey. (1980) The effect of filtered speech on speaker height and weight identification. *Journal of Phonetics* 8, 91-100.

Legge, G.E., Grossman, C., and C.M. Pieper. (1984) Learning unfamiliar voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 10, 298-303.

Leippe, M.R., Wells, G.L., and T.M. Ostrom. (1978) Crime seriousness as a determinant of accuracy in eyewitness identification. *Journal of Applied Psychology* 63, 345-351.

McGehee, F. (1937) The reliability of the identification of the human voice. *Journal of General Psychology* 17, 249-271.

McGehee, F. (1944) An experimental investigation of voice recognition. *Journal of General Psychology* 31, 53-65.

McLaughlin, R.J. and J.J. Eysenck. (1967) Extroversion, neuroticism, and paired associate learning. *Journal of Experimental Research: Personality* 2, 128-132.

Pollack, I., Pickett, J.M., and W.H. Sumby. (1954) On the identification of speakers by voice. *The Journal of the Acoustical Society of America* 26, 403-406.

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

Saslove, H. and A.D. Yarmey. (1980) Long-term auditory memory: Speaker identification. *Journal of Applied Psychology* 65, 111-116.

Thompson, C.P. (1985) Voice identification: Speaker identifiability and a correction of the record regarding sex effects. *Human Learning* 4, 19-27.

Walden, B.E., Montgomery, A.A., Gibeily, G.J., Prosek, R.A., and D.M. Schwartz. (1978) Correlates of psychological dimensions in talker similarity. *Journal of Speech and Hearing Research* 21, 265-275.

Uehling, B.S. and R. Sprinkle. (1968) Recall of a serial list as a function of arousal and retention interval. *Journal of Experimental Psychology* 78, 103-106.

Van Dommelen, W.A. (1993) Speaker height and weight identification: A re-evaluation of some old data. *Journal of Phonetics* 21, 337-341.

Zimmerman, M. (1960) The development of an affect adjective check list for the measurement of anxiety. *Journal of Consulting Psychology* 24, 457-462.