Distortions in the Memory of the Pitch of Speech

Steven E. Stern[1], John W. Mullennix[1], Olivier Corneille[2], and Johanne Huart[2]

[1]University of Pittsburgh at Johnstown, USA, [2]Catholic University of Louvain, Louvain la Neuve, Belgium

Abstract. Corneille, Huart, Becquart, & Brédart (2004) found that people remember ambiguous race faces as closer to a race prototype than they actually are. In three studies, we examined whether this memory bias generalizes to voice memory. In Studies 1 and 2, participants listened to synthesized male and female speech samples (high, moderate, or low pitch) and were asked to identify a voice target when paired against distracters higher or lower in pitch. The results showed that pitch distortions occurred, with the pattern consistent with assimilation toward low and high ends of the pitch continuum. Study 3 replicated this result with a wider voice pitch range. The results parallel those of Corneille et al. (2004). The implications of this work are discussed.

Keywords: categorization, memory distortions, voice memory, earwitness testimony

As long noted, categorization helps people deal with the complexity of their environment, but also entails a host of cognitive and perceptual biases (e.g., Allport, 1954). Among these biases are the accentuation of perceived between-categories differences and within-categories resemblances (e.g., Harnad, 1987; Tajfel & Wilkes, 1963). Recently, the implications of these biases were considered in the context of face recollection. Corneille, Huart, Becquart, and Brédart (2004) reported evidence that people recollect ambiguous race faces as more typical of their race category than they are. Huart, Corneille, and Becquart (2005) recently extended this memory distortion finding to the recollection of faces that are gender-ambiguous. To illustrate, faces moderately typical of the Asian category are remembered as more Asian-like than they are (Corneille et al., 2004) and faces moderately typical of the female category are remembered as more female-like than they are (Huart et al., 2005).

One of the practical implications of the work on category-based distortions of face memory is eyewitness testimony (e.g., Corneille et al., 2004; Huart et al., 2005). It is important to understand the manner in which face memories are distorted and lead to erroneous face identifications. In the same vein, the area of "earwitness testimony" is also becoming important. Earwitness testimony has received much less attention than eyewitness testimony (Olsson, Juslin, & Winman, 1998). Nevertheless, earwitness testimony has important legal implications for a variety of crimes for which witnesses might be asked to later recognize a voice, including bomb threats, ransom demands, obscene phone calls, and hooded rapes (Bull & Clifford, 1984). Earwitness testimony has been used in courts of law since at least 1660 (Deffenbacher, Cross, Handkins, Chance, Goldstein, Hammersly et al., 1989). It continues to be used to prosecute criminals in modern law enforcement. Unfortunately, as with eyewitness testimony, it can be fallible and lead to wrongful prosecution (Olsson et al., 1998; Yarmey, Yarmey, Yarmey, & Parliament, 2001). Given the potential importance of earwitness testimony, it is worthwhile to examine whether memory distortions occur when remembering and recognizing voices. In the present study, our goal is to assess whether these memory distortions occur when considering auditory stimuli. More specifically, we examined memory distortions occurring in the recollection of voice pitch.

Classic work and more recent research in cognitive psychology helps us understand why the memory of an ambiguous face is likely to be distorted toward a face more typical of a given category. For example, Huttenlocher and colleagues (Huttenlocher, Hedges, & Duncan, 1991; Huttenlocher, Hedges, & Vevea, 2000; see also Bartlett, 1932; Brewer & Nakamura, 1984) proposed that people encode fine-grained (i.e., perceptual) information along with categorical information self-generated about stimuli at encoding. Because fine-grained representations are generally poor, people rely on the self-generated categorical information when asked for a report from memory. Huttenlocher et al. (1991, 2000) explain this in terms of categorical information at recollection shifting memory reports toward more prototypical category values, a process that they call "weighting with a prototype." In this model, recollections are assimilated toward more prototypical values of the category, which are spontaneously computed as participants process discrete category exemplars (Posner & Keele, 1968; in the context of face memory models, see also Valentine, 1991).

In the area of speech perception, research on the "perceptual magnet effect" suggests that speech sounds are categorized into memory via phonetic prototypes (Kuhl, 1991). Basically, the perceptual magnet effect refers to a situation where perceptual distinctions near phonetic pro-

DOI 10.1027/1618-3169.54.2.148

totypes are reduced and perceptual distinctions between phonetic categories are enhanced. However, recent work has failed to provide further evidence for perceptual magnet effects in speech (Walley & Sloane, 2001), suggesting that other research is needed to support a prototype view.

In terms of memory for voice, there is some evidence that voices are stored in long-term memory in a highly organized categorical format. Some researchers have suggested that categorization of voices is related to voice prototypes stored in long-term memory (Kreiman & Papcun, 1991; Spisak, Mullennix, Moro, Will, & Farnsworth, 2002; Papcun, Kreiman, & Davis, 1989). For example, Papcun et al. (1989) suggested that voices rated as "hard to remember" tended to be average sounding voices that constitute voice prototypes. They found that, over time, memory for specific target voices that listeners were exposed to converged upon the "hard to remember" voice prototypes. Kreiman and Papcun (1991) suggested further that, over time, the distinctions that initially separate a voice that one just heard from other voices stored in memory are lost over time, with voice identification converging on the most "typical" sounding voices. This work suggests that memory distortions for voice can occur under certain circumstances, with the distortions defaulting toward prototypes of voices stored in long-term memory categories.

Although Huttenlocher's model has been applied to specific instances as in the retrieval of relatively simple visual stimuli (i.e., the localization of dots in a circle) and complex visual stimuli (i.e., the identification of faces), it is intended as a general model of reports from episodic memory that can be applied to a variety of situations, including, as examined in the present study, auditory percepts. It should be noted, however, that recent research on category-based representations has shown that the nature of the category representation in long-term memory is still up for debate. A number of researchers have proposed that memory categories are organized around exemplar-based representations, not prototypes (e.g., Estes, 1994; Medin & Schaffer, 1978; Nosofsky, 1992). Much debate about the nature of category representations has ensued, and the issue is still unresolved (see Lamberts, 1996; Minda & Smith, 2002; Nosofsky & Zaki, 2002; Smith, 2002, 2005; Storms, Boeck, & Ruts, 2000; Zaki, Nosofsky, Stanton, & Cohen, 2003).

One particular discussion of the representation issue comes from Palmeri and Nosofsky (2001), who distinguish between what are called "central-tendency" prototypes and "extreme-point" prototypes. Based on data from a series of perceptual categorization experiments, they conclude that categorization is best explained in terms of psychological extreme-point representations compatible with an exemplar model. The distinction between central-tendency and extreme-point representations appears useful in describing the situation in the present study, hence we will return to this distinction later on.

In the studies presented below, we examined the problem of voice memory distortions by focusing on one aspect of human voice, voice pitch, which is a major factor when it comes to describing the acoustic attributes than distinguish voices from one another (e.g., Klatt & Klatt, 1990). We chose to examine memory distortions for voice in a manner parallel to the work on face memory (Corneille et al., 2004; Huart et al., 2005) discussed above. Specifically, we created sets of voice stimuli that varied in voice pitch and set up a procedure by which we could assess recognition memory for pitch. Our specific predictions concerned whether voices of ambiguous voice pitch would be remembered as more typical of their pitch category than they actually are. If the results for voice pitch recognition parallel the findings of Corneille et al. (2004) and Huart et al. (2005) for facial recognition, we would expect to see an increase in the selection of higher pitched distracters as a target voice's pitch increases and an increase in the selection of lower pitched distracters as a target voice's pitch decreases. There are two ways that such a pattern of distortions could be interpreted, one being assimilation toward central-tendency representations of voice pitch categories, and the other being assimilation toward extreme-point representations located on each end of the pitch continuum.

Study 1

Method

Participants and Design

A total of 58 undergraduates at the Catholic University of Louvain participated in this study[1]. The design was a 3 (voice pitch: low, moderate, and high) by 2 (distracter pitch: higher, lower) design, with voice pitch as a between-subjects factor and distracter pitch as a within-subjects factor.

Materials

The first study was conducted on a Belgian Francophone sample using French male synthetic speech. A text-to-speech synthesis program was used to create the target and distracter voice samples, which avoided the inevitable problems occurring if we attempted to record speech samples from different speakers in order to vary pitch of speech or had a single speaker distort his/her own voice to create different pitched samples. To choose target stimuli, we created speech samples (.wav files) using the Elan Sayso Speech Engine v. 4.302 presented through Scube Sample v 1.0 running on an IBM-PC compatible computer using a Windows XP platform. Elan Sayso is a realistic sounding text to speech (TTS) synthesizer. Unlike the previous gen-

1 In this initial study we did not collect data on participants' gender.

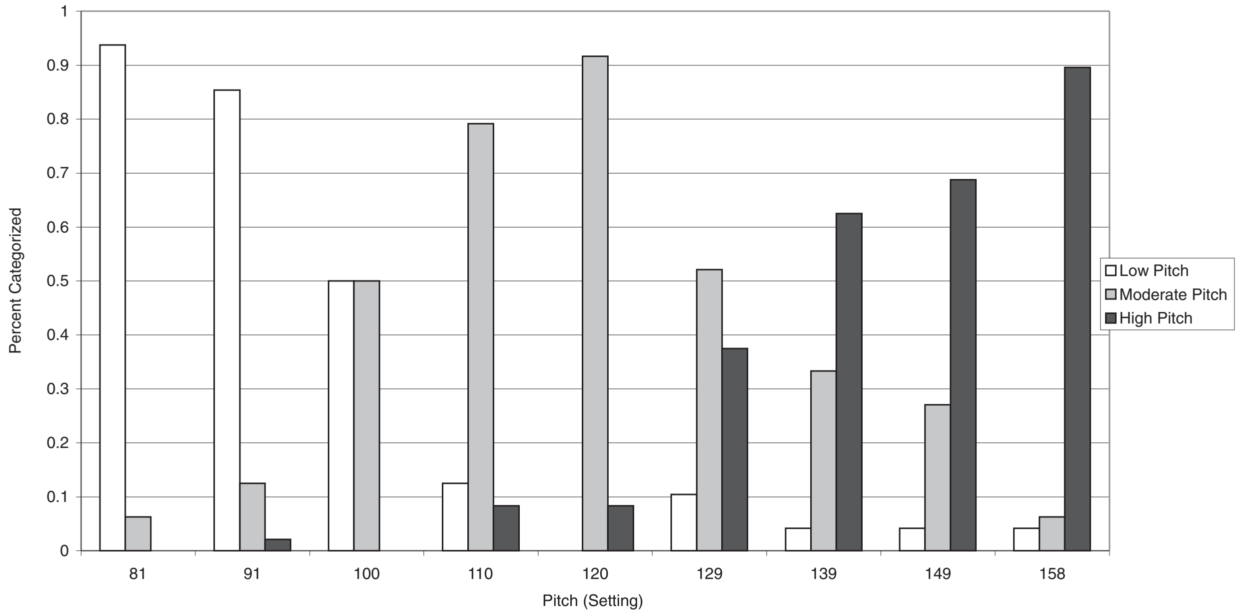

Figure 1. Categorization of speech samples for data selection in Study 1.

eration of TTS programs, it produces speech samples from concatenated pieces of real human speech. These files were then converted into a Sound Edit format in order to be read by the PsyScope programs that we used to collect data at the pretest and experimental sessions.

Each sample was a speech segment consisting of a sequence of five French words: chapeau (hat), enfance (childhood), vacances (holidays), immeuble (building), and marteau (hammer). The default male voice "Robert" was used. Nine samples were created. The voice fundamental frequency of each utterance was analyzed using a pitch extraction software package (TFPR) from Avaaz Inc. The utterances were analyzed using a waveform matching algorithm based on 0 dB silence threshold and 1500 Hz zero-crossing frequency threshold parameters. To analyze voice pitch, the midpoint of the second vowel in the word "Française" was located and the voice fundamental pitch extracted from the steady-state portion of the vowel. The results of these measurements indicated that the 9 pilot samples ranged from a low of 94 Hz to a high of 186 Hz.

Table 1. Synthesizer pitch settings for stimuli used in Study 1 on pitch of male speech

Condition	Distracter	Setting
Low Pitch (Target = 91 [105 Hz])	Lowest	81
	Lower	85
	Higher	97
	Highest	100
Moderate Pitch (Target = 120 [141 Hz])	Lowest	110
	Lower	113
	Higher	126
	Highest	129
High Pitch (Target = 149 [174 Hz])	Lowest	139
	Lower	145
	Higher	155
	Highest	158

Note: The intervals in between the targets and the distracters are not perfectly equal or symmetrical because of rounding errors. Pitch settings from the synthesis program are shown for convenience, as the pitch settings were correlated with measured fundamental frequency values

We pretested these nine samples on 12 participants by randomly presenting 36 trials to each participant. After each sample was presented, the participant was prompted by the computer to press one of three keys to categorize the sample as low, moderate, or high pitched. Figure 1 illustrates the percentage of times that each stimulus was categorized as high, moderate, and low pitched. From these data, we chose three voices that were reliably categorized as low pitched, moderate pitched, and high pitched. These three voices were operationally defined as our voice targets for this experiment. For the high and low pitched targets, we purposely chose stimuli that were to some degree ambiguously high or low (selected less than 90% as high or low). Needless to say, choosing totally unambiguous target voice stimuli should not be expected to result in recollection distortions toward any voice pitch prototype.

For each of the three selected target stimuli, we selected four distracters, two lower in pitch and two higher in pitch. The percent difference in Hz between the target and each distracter fell well over the 0.3–0.5% fundamental frequency limen for male voice (see Flanagan, 1972,

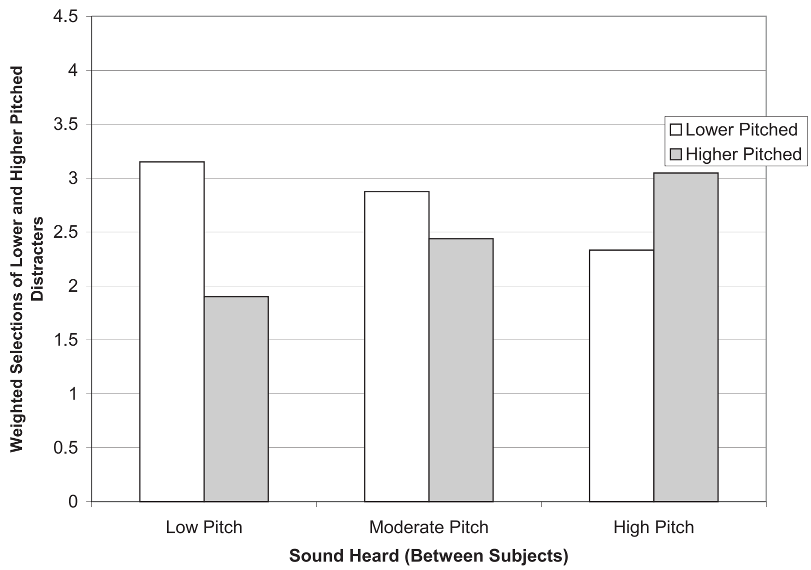

Figure 2. Weighted selection of lower and higher distracters by experimental condition for Study 1 (male voice).

p. 281), indicating that the voice pitch differences between the targets and all distracters were audible to the listener. In Table 1, a list of the stimuli used for the three experimental conditions (low pitch, moderate pitch, and high pitch) are shown along with the synthesizer pitch settings corresponding to these stimuli. The distracter stimuli involved the same speech segment (i.e., the same sequence of words) as the target stimuli, but differed in frequency.

Procedure

On their arrival to the laboratory, participants were greeted by the experimenter and seated at an Apple computer. The instructions for the experiment and the presentation of all experimental stimuli were handled automatically via the computer using PsyScope.

Depending upon the experimental condition (low, moderate, high), each participant was presented with the target speech sample, which consisted of five words. After hearing the target, each participant completed eight trials during which they listened to two speech samples and were asked to hit the "1" key if the first sample was the same as the original (target) sample or the "2" key if the second sample was the same as the original sample. The eight trials consisted of the target sample paired with each of the four distracter samples. Half the time, the target was presented first, while half the time the target was presented second. For each participant, the order of presentation of the eight trials was randomized by the experimental software.

Results and Discussion

In Table 2, the mean number of errors for each distracter type are listed across the three pitch conditions. Because larger errors represented greater amounts of memory distortion, stronger weightings were assigned when participants chose the distracters that were relatively farther from the target. A weight of " + 1" was assigned when participants mistakenly selected a distracter that was closer to the target while a weight of " + 2" was assigned when participants mistakenly selected a distracter that was farther from the target[2]. Two scores, one representing the weighted sum of incorrect selections of higher pitched distracters, and one representing the weighted sum of incorrect selections of lower pitched distracters were then calculated for each participant. These scores were entered in a mixed ANOVA where they were examined as a function of voice pitch.

Table 2. Mean number of errors by type of error and experimental condition for Study 1

		Experimental condition			
		Low pitch	Moderate pitch	High pitch	Total
Error type	Lowest	.95	.88	.71	.84
	Low	1.25	1.12	.90	1.09
	High	.60	1.19	1.05	.93
	Highest	.65	.63	1.00	.77
	Total	3.45	3.81	3.67	3.63

Note: Because there are eight trials per participant, participants could make a maximum of two errors per type of error, and eight errors total.

2 In all three studies we used this weighting scheme that we had designed a priori. Post hoc analyses of errors do not demonstrate a significant difference between participants' choice of distracters that are close versus far from the target. Nonetheless, we have retained this weighting scheme inasmuch as it reflects the magnitude of the errors that participants are making.

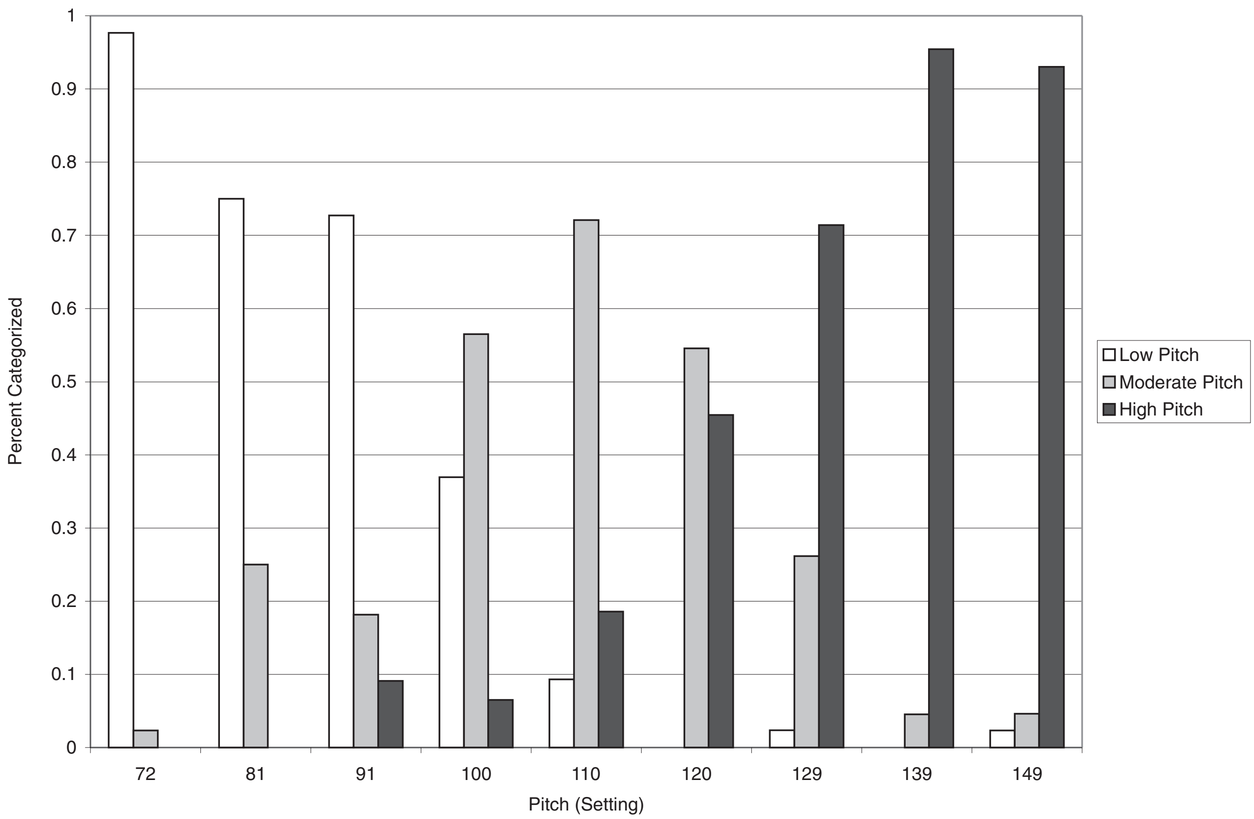

Figure 3. Categorization of speech samples for data selection in Study 2.

Across all conditions, there was no significant effect of distracter pitch, $F(1, 54) = 1.13$, $p = .29$. There was also no significant main effect of voice pitch, $F(2, 54) = .09$, $p = .92$. Most importantly, however, there was a significant interaction between voice pitch and distracter pitch, $F(2, 54) = 3.83$, $p = .03$ (see Figure 2). To examine this interaction more closely, we calculated a score for each participant by subtracting their weighted high pitch direction errors from their weighted low pitch direction errors. By performing a linear contrast (weights of + 1, 0, –1) on the mean difference scores for the three groups of participants, we found that this linear trend was significant, $t(54) = 2.75$, $p = .008$. Thus, the higher the pitch of the target sample, the stronger the tendency of selecting a higher pitched over a lower pitched distracter.

The present data offer support for the prediction that people distort their memory for voice pitch stimuli, with a substantial number of distracters misidentified as the targets. Of specific interest is the pattern of incorrect distracter selections across pitch target conditions. For moderate pitch targets, both lower and higher pitched distracters were misidentified equally often. For high pitch targets, higher pitched distracters were misidentified more often than the lower pitched distracters. Conversely, for low pitch targets, lower pitched distracters were misidentified more often than higher pitched distracters. Before drawing any tentative conclusions from these results, we decided to conduct a second study. In Study 2, we sought to replicate and extend these findings by using female voice samples. Replicating the findings of Study 1 with a female voice allows us to examine whether the pattern of results we observed holds only for male voice or whether the results generalize across gender (i.e., across a different frequency range of voice pitches).

Study 2

Method

Participants and Design

A total of 41 undergraduates (33 males; 7 females; 1 undisclosed) at the Catholic University of Louvain participated in this study. As in Study 1, the design was a 3 (voice pitch: low, moderate, & high) by 2 (distracter pitch: higher, lower) design, with voice pitch as a between-subjects factor and distracter pitch as a within-subjects factor.

Materials

We chose target stimuli in the same way that we did in the first study, except that we used the default female voice

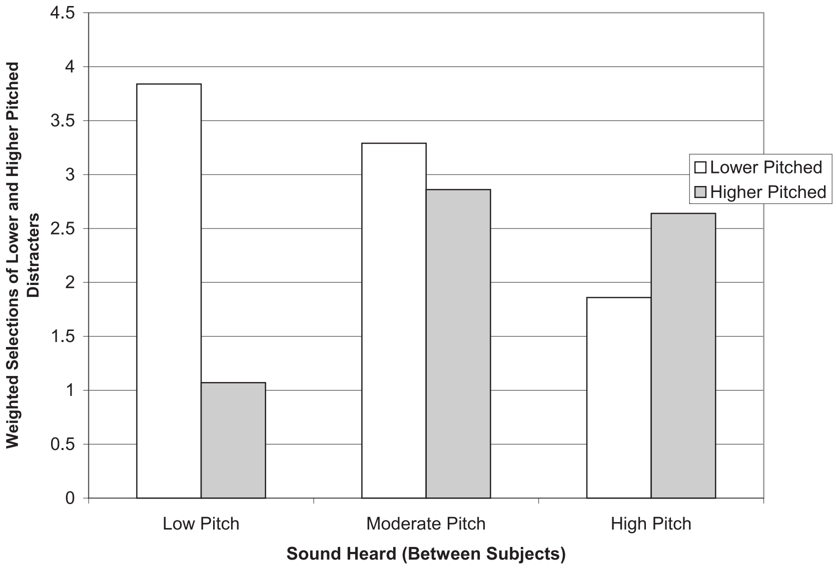

Figure 4. Weighted selection of lower and higher distracters by experimental condition for Study 2 (female voice).

Table 3. Synthesizer pitch settings for stimuli used in Study 2 on pitch of female speech

Condition	Distracter	Setting
Low pitch (Target = 81 [167 Hz])	Lowest	72
	Lower	75
	Higher	88
	Highest	91
Moderate pitch (Target = 110 [231 Hz])	Lowest	101
	Lower	104
	Higher	117
	Highest	120
High pitch (Target = 129 [270 Hz])	Lowest	120
	Lower	123
	Higher	136
	Highest	139

Note: The pitch settings for each study are relative values that pertain to each particular synthesized voice.

Table 4. Mean number of errors by type of error and experimental condition for Study 2

		Experimental condition			
		Low pitch	Moderate pitch	High pitch	Total
Error type	Lowest	1.23	1.00	.57	.93
	Low	1.38	1.29	.71	1.12
	High	.46	1.00	.79	.76
	Highest	.31	.93	.93	.73
	Total	3.38	4.21	3.00	3.53

Note: Because there are eight trials per participant, participants could make a maximum of two errors per type of error, and eight errors total.

"Cathy." We pretested nine speech samples on 10 participants (3 males and 7 females). The fundamental frequencies (Hz) of the pilot samples used in the pilot study ranged from 145 Hz on the low end to 315 Hz on the high end. From these samples, we chose three stimuli that were reliably categorized as low pitched, moderate pitched, and high pitched, which we designated as the voice targets (see Figure 3). For each of the three selected target stimuli, we chose four distracters, two lower in pitch and two higher in pitch. See Table 3 for a list of the stimuli used and the pitch settings from the synthesis program used for these stimuli.

Procedure

The procedure for the second study was identical to the procedure for Study 1 except for the stimuli used.

Results and Discussion

In Table 4, the mean number of errors for each distracter type are listed across the three pitch conditions. In terms of analysis, we used the same weighting and analysis procedures as for Study 1. Across all conditions, there was a significantly greater likelihood of making errors toward the low pitched distracters (M = 2.98) versus the high pitched distracters (M = 2.22), $F(1, 38) = 4.33$, $p = .04$. There was no significant main effect of voice pitch, $F(2, 38) = 1.97$, $p = .15$. As in Study 1, there was a significant interaction between voice pitch and distracter pitch, $F(2, 38) = 7.16$, $p = .002$ (see Figure 4). Again, to examine this interaction, we calculated a score for each participant by subtracting their weighted high pitch direction errors from their weighted low pitch direction errors. We found that the linear trend was significant, $t(38) = 3.73$, $p < .001$. Once again, the higher the pitch of the target sample, the stronger the tendency of selecting a higher pitched distracter over a lower pitched distracter. We did not find any main effects or interactions involving gender.

The second study thus replicates Study 1 with a different set and range of stimuli. The data provide further support for the view that ambiguous voice pitch stimuli are likely to be remembered as more typical of their voice pitch category than they actually are.

In examining the results from Studies 1 and 2, it is important to note that Peterson and Barney (1952) showed that the average voice pitch for males producing English vowels ranged from 124 Hz to 141 Hz, while the average voice pitch for females ranged from 210 Hz to 235 Hz. Other studies show similar values for French (e.g., Boë, Contini, & Rakotofiringa, 1975; Chevrie-Muller & Gremy, 1967). In the first two studies, the moderate pitched targets fell within these ranges (i.e., 141 Hz for male voice and 231 Hz for female voice), therefore reflecting voice pitches typically found within the population. However, the low and high pitch target values were somewhat below and above the range, respectively. In each of the first two studies, pitch distortions for the moderate pitched targets were approximately equal for lower and higher pitched distracters surrounding the target.

However, for low pitch targets and high pitch targets, an asymmetry was observed, where distracters were misidentified more often if they came from the lower end and higher ends of the pitch continuum, respectively. Overall, the results are consistent with both explanations discussed earlier concerning the nature of the representation of voice pitch that is at work. One explanation is that assimilation toward central-tendency representations of the three voice pitch categories occurred. The second explanation is that assimilation toward two extreme-point representations of voice pitch on the lower and higher ends of the pitch continuum occurred.

In Study 3, we sought to investigate whether the effects we found in Studies 1 and 2 were due to assimilation toward extreme-point representations of voice pitch. We thus set up a situation to "push the effects" found previously for low pitch and high pitch targets along a continuum of more extreme voice pitches found outside the normal range for male human voice. Our logic was that, if extreme-point representations are involved, then the same exact pattern of distortions for the high and low ends of the voice pitch continuum found in Studies 1 and 2 should also occur when examining extreme voice pitch ranges that fall far beyond the range of average-pitched speech. In addition, Study 3 also provided us with a gender balanced sample that could permit the examination of any gender differences.

Study 3

A five group design was used that included a very low pitch group as well as a very high pitch group, in addition to the high-, moderate-, and low-pitch groups as in the original studies. The study also examined whether the original effect observed in French generalizes to English language utterances, and to determine if, using a gender balanced sample, whether the effects we observed previously are equal for male and female participants.

Method

Participants and Design

A total of 120 undergraduates (63 males; 57 females) at the University of Pittsburgh at Johnstown participated in this study. In contrast to the previous two studies, the design was a 5 (voice pitch: very low, low, moderate, high, and very high) by 2 (distracter pitch: higher, lower) design, with voice pitch as a between-subjects factor and distracter pitch as a within-subjects factor.

Materials

We chose target stimuli in the same way that we did in the first two studies, except that we used the default English voice "William." We pretested nine speech samples on 22 participants (9 males and 13 females). The fundamental frequencies (Hz) of the pilot samples ranged from 46 Hz on the low end to 186 Hz on the high end. From these samples, we were able to choose five stimuli that could be categorized as very low pitch, low pitch, moderate pitch, high pitch, and very high pitch, which we designated as the targets (see Figure 5). Each sample was a speech segment consisting of a sequence of five English words: hat, childhood, holidays, building, and hammer. These were a direct translation of the French words used in Study 1 and Study 2.

For each of the five selected target stimuli, we chose four distracters, two lower in pitch and two higher in pitch. See Table 5 for a list of the stimuli used and the pitch settings of these stimuli.

Procedure

The stimuli were presented on a PC-compatible computer running the E-Prime experimental software. Participants were seated individually at the computer. The initial procedure for the third study was identical to the procedure for the first two studies except for the stimuli used and the inclusion of two more groups of participants: those listening to very low pitch samples and those listening to very

Table 5. Synthesizer pitch settings for stimuli used in Study 3 on pitch of male speech (5 groups)

Condition	Distracter	Setting
Very low pitch (Target = 50 [46 Hz])	Lowest	40
	Lower	45
	Higher	55
	Highest	60
Low pitch (Target = 65 [62 Hz])	Lowest	55
	Lower	60
	Higher	65
	Highest	70
Moderate pitch (Target = 115 [117 Hz])	Lowest	105
	Lower	110
	Higher	120
	Highest	125
High pitch (Target = 170 [165 Hz])	Lowest	160
	Lower	165
	Higher	175
	Highest	180
Very high pitch (Target = 190 [193 Hz])	Lowest	180
	Lower	185
	Higher	195
	Highest	200

high pitch samples. As in Study 1 and Study 2, in each condition participants were presented with 8 trials comparing the target to distracters. In addition, after completing the primary task, participants performed two additional tasks. For the first task, they were asked to categorize whether the speech they listened to was low pitch, moderate pitch, or high pitch. This was performed as a manipulation check in order to determine how participants were spontaneously categorizing the stimuli. For the second task, participants completed a similarity rating task for the distracter stimuli. This task was intended as a manipulation check to ascertain that the perceptual distances between stimuli were fairly equivalent across the range used.

Results and Discussion

In Table 6, the mean number of errors for each distracter type are listed across the three pitch conditions. We used the same weighting and analysis procedures as for Study 1 and Study 2. Across all conditions, there was no significant difference in the likelihood of making errors toward the low pitched distracters (M = 2.45) versus the high pitched distracters (M = 2.67), $F(1, 115) = 1.00$, $p = .32$. There was a significant main effect of voice pitch, $F(4, 115) = 3.01$, $p = .02$ and a marginally significant interaction between voice pitch and distracter pitch, $F(4, 115) = 2.34$, $p = .06$.

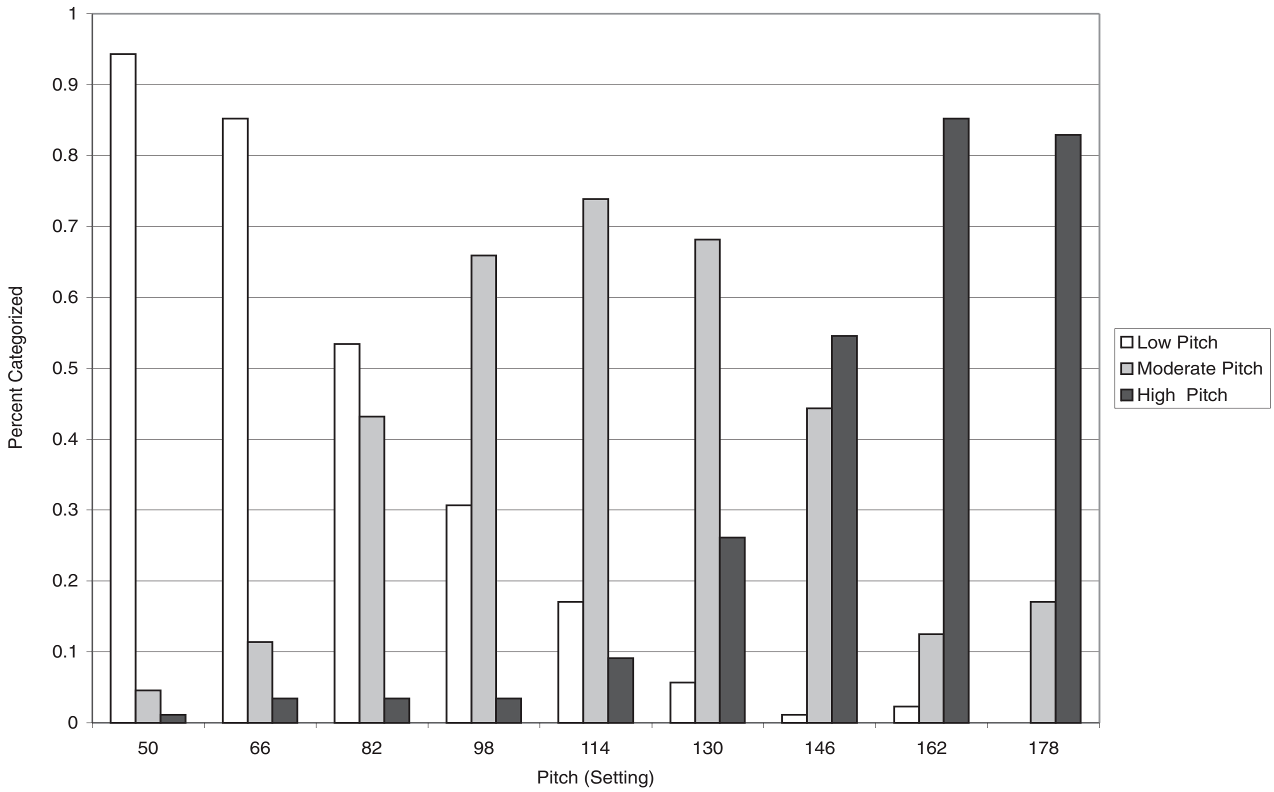

Figure 5. Categorization of speech samples for data selection in Study 3.

Table 6. Mean number of errors by type of error, gender, and experimental condition for Study 3

		Experimental condition					
		Lowest pitch	Low pitch	Moderate pitch	High pitch	Highest pitch	Total
Error type	Lowest						
	Male	1.15	.92	.77	.77	1.18	.95
	Female	.82	.73	1.09	.73	1.00	.88
	Low						
	Male	.69	1.15	.85	.92	.82	.89
	Female	.73	.55	.73	.91	1.00	.79
	High						
	Male	.23	.54	.54	1.23	1.18	.73
	Female	.36	1.00	.82	1.18	1.00	.88
	Highest						
	Male	.54	.54	1.23	1.08	1.09	.89
	Female	.64	.55	.91	.82	.85	.75
	Total						
	Male	2.62	3.15	3.38	4.00	4.27	3.46
	Female	2.55	2.82	3.55	3.64	3.85	3.29

Note: Because there are eight trials per participant, participants could make a maximum of two errors per type of error, and eight errors total.

Again, to examine this interaction, we calculated a score for each participant by subtracting their weighted high pitch direction errors from their weighted low pitch direction errors. We found that the linear trend was significant, $t(115) = 3.55$, $p = .02$. Once again, the higher the pitch of the target sample, the stronger the tendency of selecting a higher pitched over a lower pitched distracter. This third study thus replicates the effect, but for a wider range of stimuli including very high and very low pitch voices. It also extends the effect to a novel population sample speaking a different language (i.e., English instead of French).

To further probe these effects, inasmuch as this study had a larger and nearly gender balanced sample, we examined the data separately for males and females. In this case, we found that the effect was more pronounced for males than for females. While the linear trend was not significant for females, it was significant for males, $t(58) = 2.69$, $p = .01$ (see Figure 6 for the male data). Although we can only speculate as to why the latter effect was obtained for male participants only, one possibility is that memory distortions are magnified when people are recollecting a memory trace for in-group targets (e.g., male participants exposed to a male voice, as in the present study). Interestingly, Corneille et al. (2004) obtained evidence consistent with this conjecture, with larger distortions in face memory when the

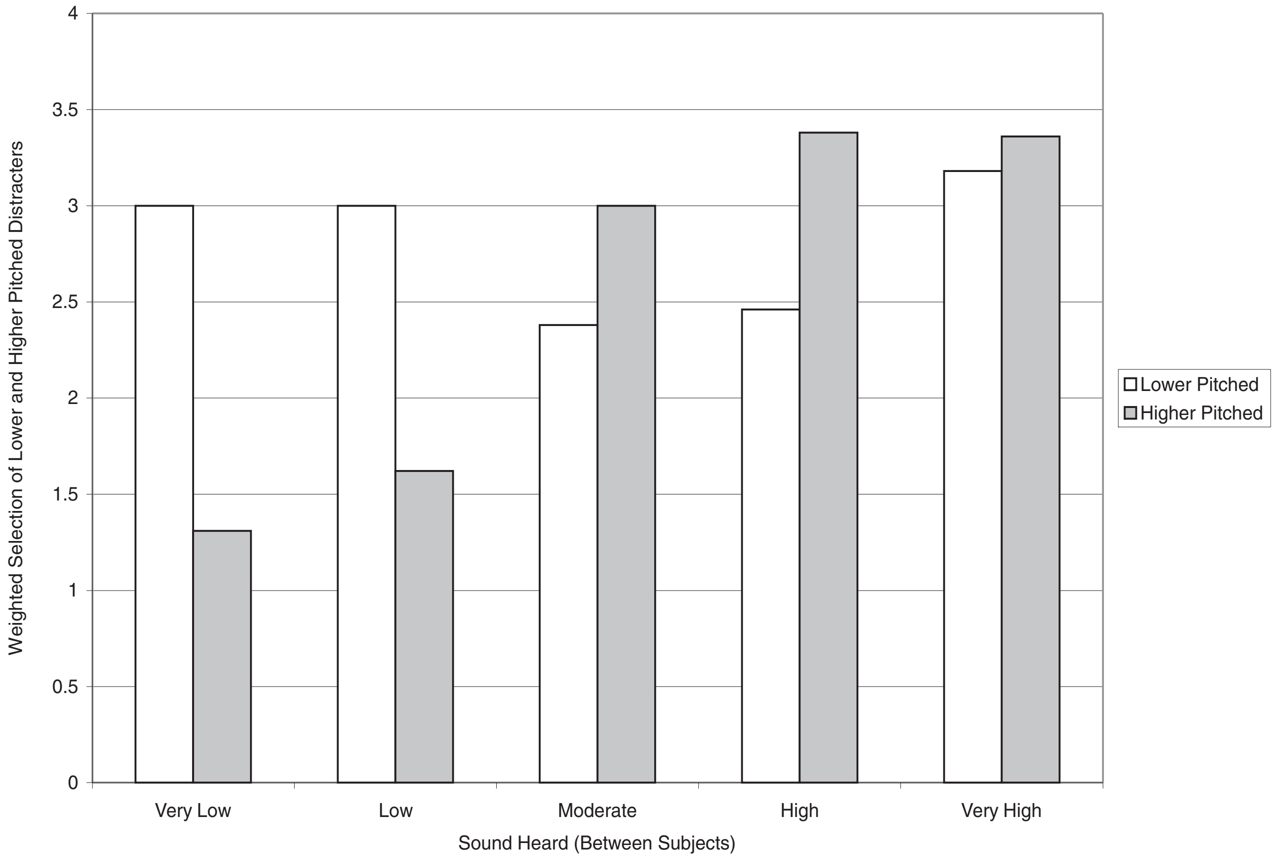

Figure 6. Weighted selection of lower and higher distracters by experimental condition for male participants only in Study 3.

Table 7. Spontaneous categorization data (in terms of counts) for manipulation check for Study 3

	Self-rating		
Condition	Low	Moderate	High
Very low	16	6	1
Low	12	11	1
Moderate	2	21	1
High	3	11	10
Very high	0	17	7

Table 8. Results of similarity judgments between adjacent stimuli across pitch conditions

Condition	Comparisons (pitch settings)	Mean rated difference	*F*	*p*
Very low	40, 45	2.62	4.42	.01
	45, 50	3.04		
	50, 55	2.81		
	55, 60	2.62		
Low	55, 60	2.55	1.97	.13
	60, 65	2.60		
	65, 70	2.88		
	70, 75	2.76		
Moderate	105, 110	2.80	9.18	< .001
	110, 115	2.44		
	115, 120	2.43		
	120, 125	2.23		
High	160, 165	2.94	7.60	< .001
	165, 170	2.53		
	170, 175	2.49		
	175, 180	2.55		
Very high	180, 185	2.13	1.27	.29
	185, 190	2.10		
	190, 195	1.97		
	195, 200	2.11		

(mostly Caucasian) participants recollected moderately Caucasian faces rather than moderately Asian or North African faces. Also, Huart et al. (2005) tended to obtain larger memory distortions in the recollection of ambiguous female faces than ambiguous male faces with their (mostly female) participants.

Also of importance, we note that for the very high pitch condition, the effect dissipated. The results from the spontaneous categorization task manipulation check are shown in Table 7 (results are not shown separately for males and females as the pattern was similar for both). For the very low, low, moderate, and high conditions, participants categorized the stimuli as expected. However, for the very high pitched stimuli, the categorization data resembles that found for the moderate stimuli. There is a correspondence between the memory distortion data and the spontaneous categorization data for the very high condition. For the distortion data, the predicted effect for the very high condition was weak for male participants and absent for female participants. For the spontaneous categorization data, the expected categorization of the stimuli as high pitched was not observed. Given these anomalies, it appears that participants may not have perceived the very high stimuli as high pitched, and this in turn may have affected how they treated these stimuli in the memory distortion task.

A possible explanation for why participants treated the stimuli in this way may lie with pitch registers for human voice. When examining the pitch values for the very high pitch stimuli, the values begin to encroach upon the voice pitch range identified as the loft voice register (Hollien, 1974), which is the speech equivalent of the falsetto register for singing voice. It is possible that participants perceived the very high stimuli as originating from a more moderate-pitched speaker who was temporarily speaking in a higher pitched register.

In terms of the similarity rating task, each participant listened to 80 randomized pairs of adjacent distracters (e.g., setting 65 and setting 70) and performed similarity judgments on a 1 to 7 scale, where 1 was *highly similar* and 7 was *highly dissimilar*. As shown in Table 8, there were significant differences found in three out of the five conditions. However, there was no consistent pattern of results across conditions in terms of differences between adjacent distracters. Furthermore, it should be noted that the mean differences between any two comparisons within any of the pitch conditions were less than 1 point on the 1–7 rating scale, attesting to the high statistical power (120 participants each making approximately 80 judgments) of these analyses to detect subtle differences that may have little practical significance. The results of these analyses suggest that the perceptual distances between the stimuli used in Study 3 were fairly equivalent across the range used.

Overall, although there is some evidence that assimilation toward the very low end of the pitch continuum occurred, the results of Study 3 are inconclusive. The issue of whether the pitch distortions reflect assimilation toward central-tendency voice pitch representations or assimilation toward extreme-point voice pitch representations remains unresolved.

General Discussion

In the last several years, the scientific study of memory distortions has provided social, cognitive, and forensic psychology with a wealth of basic and applied findings (Roediger & McDermott, 2000). This body of research has influenced our field in numerous ways from advancing our understanding of the malleability of memory (Loftus & Palmer, 1974) to informing us about the fragile nature of our memory of people and events when we need to reconstruct them in legal settings where objective truth is of paramount importance to all parties involved (Loftus, 1993; see also Fiedler, Walther, Armbruster, Fay, & Naumann, 1996; Fiedler, Armbruster, Nickel, Walther, & Asbeck, 1996).

As noted earlier, important advances have been made concerning the biases intruding face memory. However, little attention has been paid to memory distortion biases related to voice recognition. In the present studies, we examined this possibility, and through the use of a realistic voice synthesizer that allowed us to precisely control voice pitch, we observed a distinctive pattern of distortions. For moderate pitch targets close to the average person's voice pitch, distortions appear for both low and high pitched distracters that are about equal. For low pitch and high pitch targets, relatively more selections of lower pitched distracters occurred for low pitched target voices and relatively more selections of higher pitched distracters occurred for high pitched targets. Taken together, these results indicate two possibilities of the underlying mechanism responsible for the distortions. One possibility is that the moderate pitch stimuli correspond to the most typical male voices heard by people, and that memory organization for those voices is predicated on a central-tendency representation, while the unusual low and high pitch voices one would encounter less often are organized around assimilation to pitch category extreme points. Supporting evidence for the central tendency idea is found with research on typicality ratings of male voice (Spisak et al., 2002). A second possibility is that assimilation toward the low and high extreme points of the pitch continuum occurred. The pattern observed for moderate targets also fits with this explanation, because one would expect asymmetries in distracter selection for low and high pitch targets, but not for moderate targets. In terms of our perspective, assimilation toward the low and high ends of the pitch continuum appears a more parsimonious explanation, with extreme-point representations of voice pitch deemed more appropriate. However, due to the inconclusive findings of Study 3, we cannot determine for certain whether this is indeed the case.

The findings reported here thus provide empirical evidence that category-based distortions of memories generalize to auditory stimuli. Of importance, these distortions emerged as a function of participants' spontaneous categorization of the voices, as no context information was associated with the voices at encoding. Also, these distortions were obtained despite the inclusion of the target voice in the identification task, a procedure that is known to reduce misidentifications, at least in the context of face memory.

Would this effect hold for other aspects of human speech, such as speaking rate, emotional tone of voice, etc.? At this point we do not have a clear answer to this question. In another experiment (Stern, Corneille, Huart, & Mullennix, 2004), we exposed participants to passages of speech that varied in speaking rate. In that experiment, however, we failed to find the same pattern of results as we have observed for voice pitch. One possible explanation for the differences in findings between pitch of speech and rate of speech may be that within-individual variation is larger on the latter than the former dimension. This raises the issue that different dimensions of speech may be more or less susceptible to category-based memory distortions. Further examination of the limitations of these effects in speech is warranted in order to understand whether the effect we have found with pitch is generalizable to other speech dimensions.

Whether the combination of visual and auditory information makes for more accurate identification is a more complex issue than the benefit of auditory information when that is all that is available. Some research suggests that auditory identification can be a useful supplement to visual identification of assailants. Visual line-up accuracy has been found to improve when it is accompanied by auditory information (Melara, Dewitt-Rickards, & O'Brien, 1989) and witnesses express a preference of voice over other possible cues (e.g., gait) when trying to identify a perpetrator from mugshots (McAllister, Blair, Cerone, & Laurent, 2000). The converse, however, might not be true. Some research has found that auditory identification is degraded when visual information, such as faces, is added (Cook & Wilding, 1997, 2001).

The findings reported here are also in line with Bartlett's (1932) classic contention that "remembering is not the re-excitation of innumerable fixed, lifeless, and fragmentary traces. It is an imaginative reconstruction or construction" (p. 213). More generally, the present research is also consistent with the increasing attention paid over the last decade to memory accuracy issues (for a review, see Koriat, Goldsmith, & Pansky, 2000). According to the memory accuracy framework, memory distortions should not be merely considered as methodological noises, but may inform us on the way memory works (see also Roediger & McDermott, 2000). A simple but powerful effect examined here was the tendency for memory to be distorted toward more prototypical representations. For the first time, we reported evidence for the operation of the latter process in the recollection of auditory stimuli.

As with the growing literature that suggests that visual memory can be distorted to the point to which faces can be mistaken for one another, this data has practical applications, most notably in legal settings. Recently DNA analysis has resulted in the exoneration of at least one person convicted of murder based upon earwitness testimony (Yarmey et al., 2001). The systematic accentuation of memory for pitch of speech should be of interest to those concerned with correct identification based upon voice recognition. It will also be of importance to those who advocate the use of both audio and visual stimuli in criminal identification (Melara, DeWitt-Rickards, & O'Brien, 1989) creating mugshot books and lineups that incorporate both visual and auditory features. Hopefully, a body of future research in this direction will contribute to the development of better identification procedures.

Acknowledgments

This research was conducted while the first author was on sabbatical at the Catholic University of Louvain in Louvain la Neuve, Belgium, and benefited from a grant FNRS 1.5.100.3 awarded to the third author. The authors thank Dominique Bouchat, Sebastien Moros, Timothy Potter, Brian Putnam, Brandon Petricca, Ashley Davis, Donald Horvath, Ashley Mahdavi, Michelle Casey, and Peg Barta for their valuable assistance on this research.

References

Allport, G.W. (1954/1979). *The nature of prejudice*. Cambridge, MA: Perseus Books.

Bartlett, F.C. (1932). *Remembering: An experimental and social study*. Cambridge: Cambridge University Press.

Boë, L.-J., Contini, M., & Rakotofiringa, H. (1975). Étude statistique de la fréquence laryngienne [Statistical study of laryngeal frequencies]. *Phonetica, 32,* 1–23.

Brewer, W.F., & Nakamura, G.V. (1984). The nature and functions of schemas. In R.S. Wyer & T.K. Srull (Eds.), *Handbook of social cognition* (Vol. 1, pp. 119–160). Hillsdale, NJ: Erlbaum.

Bull, R., & Clifford, B.R. (1984). Earwitness voice recognition accuracy. In G.L. Wells & E.F. Loftus (Eds.) *Eyewitness testimony: Psychological perspectives*. Cambridge, UK: Cambridge University Press.

Chevrie-Muller, C., & Gremy, F. (1967). Contribution a l'éstablissement de quelques constantes physiologiques de la voix parlée de l'adulte [Contribution to the establishment of various physiologic parameters of the speaking voice of adults]. *Journal Français d'Oto-Rhino-Laryngologie, 16,* 433–455.

Cook, S., & Wilding, J. (1997). Earwitness testimony 2: Voices, faces and context. *Applied Cognitive Psychology*, *11*, 527–541.

Cook, S., & Wilding, J. (2001). Earwitness testimony: Effects of exposure and attention on the face overshadowing effect. *British Journal of Psychology, 92,*617–629.

Corneille, O., Huart, J., Becquart, E., & Brédart, S. (2004). When memory shifts toward more typical category exemplars: Accentuation effects in the recollection of ethnically ambiguous faces. *Journal of Personality and Social Psychology*, *86*, 236–250.

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

Estes, W.K. (1994). Classification and cognition. *Cognitive Psychology, 18*, 500–549.

Fiedler, K., Armbruster, T., Nickel, S., Walther, E., & Asbeck, J. (1996). Constructive biases in social judgment: Experiments on the self-verification of question contents. *Journal of Personality and Social Psychology, 71*, 861–873.

Fiedler, K., Walther, E., Armbruster, T., Fay, D., & Naumann, U. (1996). Do you really know what you have seen? Intrusion errors and presuppositions effects on constructive memory. *Journal of Experimental Social Psychology*, *32,* 484–511.

Flanagan, J.L. (1972). *Speech analysis, synthesis and perception*. New York: Springer-Verlag.

Harnad, S. (1987). Introduction: psychophysical and cognitive aspects of categorical perception: A critical overview. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 1–25). New York: Cambridge University Press.

Hollien, H. (1974). On vocal registers. *Journal of Phonetics, 2*, 125–143.

Huart, J., Cornielle, O., & Becquart, E. (2005). Face-based categorization, context-based categorization, and distortions in the recollection of gender ambiguous faces. *Journal of Experimental Social Psychology, 41*, 598–608.

Huttenlocher, J., Hedges, L.V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review, 98*, 352–376.

Huttenlocher, J., Hedges, L.V., & Vevea, J.L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General*, *129*, 220–241.

Klatt, D.H., & Klatt, L.C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America, 87*, 820–857.

Koriat, A., Goldsmith, M., & Pansky, A. (2000). Toward a psychology of memory accuracy. *Annual Review of Psychology, 51*, 481–537.

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

Kuhl, P. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics, 50*, 93–107.

Lamberts, K. (1996). Exemplar models and prototype effects in similarity-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1503–1507.

Loftus, E.F. (1993). The reality of repressed memories. *American Psychologist, 48,* 518–537.

Loftus, E.F., & Palmer, J.P. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13,* 585–589.

McAllister, H.A., Blair, M.J., Cerone, L.G., & Laurent, M.J. (2000). Multimedia mug books: How multi should the media be? *Applied Cognitive Psychology, 14,* 277–291.

Medin, D.L., & Schaffer, M.M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207–238.

Melara, R.D., & DeWitt-Rickards, T.S., & O'Brien, T.P. (1989). Enhancing lineup identification accuracy: Two codes are better than one. *Journal of Applied Psychology, 74,* 706–713.

Minda, J.P., & Smith, J.D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 275–292.

Nosofsky, R.M. (1992). Exemplars, prototypes, and similarity rules. In A.F. Healy, S.M. Kosslyn, & R.M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honor of William K. Estes* (pp. 149–167). Hillsdale, NJ: Erlbaum.

Nosofsky, R.M., & Zaki, S.R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 924–940.

Olsson, N., Juslin, P., & Winman, A. (1998). Realism of confidence in earwitness versus eyewitness identification. *Journal of Experimental Psychology: Applied, 4,* 101–118.

Palmeri, T.J., & Nosofsky, R.M. (2001). Central tendencies, extreme points, and prototype enhancement effects in ill-defined perceptual categorization. *The Quarterly Journal of Experimental Psychology, 54*, 197–235.

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

Peterson, G.E., & Barney, H.L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24*, 175–184.

Posner, M.I., & Keele, S.W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77,* 353–363.

Roediger, H.L., & McDermott, K.B. (2000). Distortions in memory. In E.E. Tulving & F.I.M. Craik (Eds.). *The Oxford handbook of memory* (pp. 149–162). New York: Oxford University Press.

Smith, J.D. (2002). Exemplar theory's predicted typicality gradient can be tested and disconfirmed. *Psychological Science, 13*, 437–442.

Smith, J.D. (2005). Wanted: A new psychology of exemplars. *Canadian Journal of Experimental Psychology, 59*, 47–53.

Spisak, B., Mullennix, J.W., Moro, K., Will, J., & Farnsworth, L.M. (2002, June). *Typicality ratings of male and female voices.* Paper presented at the 143rd Meeting of the Acoustical Society of America, Pittsburgh, PA.

Stern, S.E., Corneille, O., Huart, J., & Mullennix, J.W. (2004, May). *Memory accentuation effects for the pitch and speed of speech.* Paper presented at the annual meeting of the Belgian Psychological Society, Free University of Brussels, Brussels, Belgium.

Storms, G., Boeck, P.D., and Ruts, W. (2000). Prototype and exemplar-based information in natural language categories. *Journal of Memory and Language, 42*, 51–73.

Tajfel, H., & Wilkes, A.L. (1963). Classification and quantitative judgment. *British Journal of Psychology, 54*, 101–114.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion and race in face recognition. *Quarterly Journal of Experimental Psychology*, *43*, 161–204.

Walley, A.C., & Sloane, M.E. (2001). The perceptual magnet effect: A review of empirical findings and theoretical implications. In F. Columbus (Ed.), *Advances in psychology research, Vol., 4* (pp. 65–92). Hauppauge, NY: Nova.

Yarmey, A.D., Yarmey, A.L., Yarmey, M.J., & Parliament, L. (2001). Commonsense beliefs and the identification of familiar voices. *Applied Cognitive Psychology, 15,* 283–299.

Zaki, S.R., Nosofsky, R.M., Stanton, R.D., & Cohen, A.L. (2003). Prototype and exemplar accounts of category learning and attentional allocation: A reassessment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 2*, 1160–1173.

Received January 11, 2006
Revision received April 21, 2006
Accepted May 18, 2006

Steven E. Stern

Department of Psychology
University of Pittsburgh at Johnstown
Johnstown, PA 15904
USA
Tel. +1 814 269-2954
Fax +1 814 269-2022
E-mail sstern@pitt.edu

