Perceptual and Motor Skills, 2003, 97, 665-670.

INCREASED PITCH INCREASES ACCURACY OF VOICE IDENTIFICATION[1]

MARY-ALICE RODSTROM AND JOHN G. NEUHOFF

The College of Wooster

Summary.—Previous work suggested that greater accuracy rates in identifying voices that have been increased in frequency over those that have been decreased in frequency may be due to complex vocal characteristics and specific memory for familiar voices. Here we asked 17 men and 21 women between the ages of 18 and 21 to learn a simple vowel exemplar produced by an unfamiliar target speaker and measured the proportion of times the frequency-shifted exemplar was identified as the originally encoded target speaker. Analysis showed that exemplars when increased in frequency were perceived as belonging to the target speaker significantly more often than exemplars which were decreased in frequency. These findings suggest that the greater accuracy in identifying speakers with increased frequency voice samples does not require previous familiarity with the vocalizations of a particular speaker or complex memory schemata for familiar voices.

Pitch is widely recognized as an important cue in the discrimination and recognition of vocal identity. Previous work has shown that changing the fundamental frequency (f_0) of a voice produces a dramatic decrease in identification and discrimination rates (Brown, 1981; Van Dommelen, 1987; Kuwabara & Takagi, 1991; Lavner, Gath, & Rosenhouse, 2000). Typically, listeners are presented with a target voice sample and asked whether subsequent pitch-shifted samples are instances of the same speaker. This work has shown a perceptual bias such that listeners tend to identify voices with increased pitch as the target speaker more often than voices with equivalently decreased pitch.

Brown (1981) recorded his own voice and manipulated f_0 both 20% higher and 20% lower than the original recording. Listeners judged similarity of stimulus pairs wherein the original voice sample was followed by samples with altered fundamental frequency. Voice samples with increased f_0 were judged similar more often than those with decreased f_0. Similarly, when listeners are asked to identify pitch-shifted exemplars after hearing live speakers, identification rates are higher for exemplars which are increased in pitch than for those decreased (Van Dommelen, 1987).

Evidence of a perceptual priority for increased vocal pitch is also found in memory tests for spoken words (Church & Schacter, 1994). For example, recognition of previously presented words is faster and more accurate when

[1]Please send correspondence to John G. Neuhoff, The College of Wooster, Department of Psychology, Wooster, OH 44691 or e-mail (jneuhoff@wooster.edu).

the words are repeated in the same voice than when they are repeated in a different voice (Church & Schacter, 1994). The significance of f_0 in this effect was tested by increasing and decreasing f_0 by 10%. Decreasing f_0 greatly reduced the priming effect compared to keeping the same frequency for the repeated word. However, increasing f_0 had no effect.

Some researchers have suggested that the listener's familiarity with a known speaker is an important element in the higher identification rates for voices that are increased in pitch. In an examination of recognition rates of familiar speakers, Lavner, *et al.* (2000) reported a higher identification rate for voices that were increased in pitch when both f_0 and formant frequency were manipulated. The study of familiar speakers was based on the assumption that listeners use cues from complex memory schemata for familiar voices. Indeed, results of word recognition studies and episodic memory suggest that many characteristics of familiar voices are retained and play an important role in strategies used to identify speakers (Pisoni, 1997). Some work has shown that increases in frequency occur during emotional vocalizations (Utsuki & Okamura, 1976; Frick, 1985). Thus, it may be that familiarity with these more salient emotional vocalizations of familiar voices contributes to the higher identification rates for voices that are increased in pitch.

However, this "speaker familiarity" hypothesis cannot account for the results of other studies that show higher identification rates for unfamiliar voices that are increased in pitch. It may be, of course, that spoken instances of single words provide listeners with enough context to extrapolate about more emotional or more salient high-pitched vocalizations. Most studies have provided listeners with at least word-length utterances. However, perhaps a better explanation is that listeners tend to remember voices as higher in pitch because more general vocal characteristics occur across speakers. For example, some work has shown that vocal frequency increases under stressful or urgent conditions (Utsuki & Okamura, 1976; Frick, 1985) and that listeners attend to vocal frequency when assessing emotion (Breitenstein, Van Lancker, & Daum, 2001). Other work has shown that higher frequency values in spontaneous speech are used to signal salient syntactic and semantic characteristics of the speech stream (Deutsch, North, & Ray, 1990). Thus, increased attention to emotional and salient speech may predispose listeners to remember voices as higher in pitch overall, regardless of familiarity with a speaker. Given these findings, we hypothesized that listeners would identify voice samples of increased frequency as the original target voice more often than voice samples of decreased frequency. We also hypothesized that higher target identification rates for voices that were increased in pitch would not require exposure to idiosyncratic vocal characteristics or even to complete words from a single speaker. We tested these hypotheses by having listeners encode a single vowel utterance from an unfamiliar voice and then

assessed whether samples of this utterance, which were altered only in frequency, were produced by the same speaker or a different one.

Method

Participants

Seventeen male and 21 female undergraduate students between the ages of 18 and 21 years were participants. None received training prior to the experiment, and all reported normal hearing.

Apparatus

Each participant was tested individually in a sound attenuated booth. Stimuli were produced by a 16-bit sound card in an IBM Pentium computer and presented via Sony MDR-V-600 headphones.

Stimuli

One woman, aged 22, and one man, aged 60, were selected for stimulus voice sampling because they were conveniently recruited, lived in a city distant from that in which the experiment was conducted, and were unfamiliar to all participants. Both voices were within the range of normal fundamental frequency for voices (Baken & Daniloff, 1991). Voices were digitally recorded in a sound attenuating booth at 16-bit resolution and 44.1-kHz sampling rate. One recording from each speaker was used. Both recordings consisted of long vowel sound /a/ (pronounced "ahh"), which was 3.5 sec. in length. A vowel was chosen because of its uniformity (Sundberg, 1999). The entire spectrum of each recording was digitally altered in frequency by one-half semitone intervals using the computer program "CoolEdit Pro" (Syntrillium Software Corporation, Phoenix, AZ), while the temporal integrity of each sample was preserved. This yielded 16 samples of increased frequency and 16 samples of decreased frequency. The digitally altered samples were one-half semitone apart on a tempered scale, so that the difference between the highest and lowest sample was 16 semitones. Recordings presented during the recognition portion of the experiment were 2.5 sec. in duration and were created by truncating the final 1 sec. of the encoding stimulus. The 2.5-sec. samples of the female and male voices had mean fundamental frequencies of 187.3 Hz and 122.5 Hz, respectively. Each of the 33 samples (16 increased frequency, 16 decreased frequency, and the original unaltered recording) was played four times for a total of 132 completely randomized trials.

Procedure

Participants completed the experiment at their own pace. Half of the participants heard the male speaker, and the other half heard the female speaker. Recordings were played at approximately 68 dBa. During the encoding phase of the experiment, participants were told that they would hear

a voice and that they should listen carefully and try to remember what the speaker sounded like. For the listeners who heard the male voice, the target speaker was identified as "Ethan." For the listeners who heard the female voice, the target speaker was identified as "Cara." Participants heard the original 3.5-sec. recording of either "Ethan" or "Cara" three times. Next, they heard the series of transposed recordings in random order. In a two-alternative forced-choice task, participants indicated whether they thought the transposed samples were instances of the target speaker or of a different speaker. Our dependent variable was the proportion of trials on which listeners identified the sample as the target speaker.

Results and Discussion

We defined a correct response as any in which the listener indicated that the vowel was produced by the target speaker. A two-tailed independent measures t test showed there were no significant differences in the proportion of correctly identified samples between the male and female stimulus voices ($t = 0.36$, $p = .72$). Thus, the data from the male and female speakers were pooled for subsequent analysis. The mean proportion of correctly identified samples at each transposed frequency is shown in Fig. 1. A 2 (direction of frequency change) × 16 (interval size) analysis of variance showed that interval size significantly affected the proportion of samples correctly identified ($F_{15,555} = 80.13$, $p < .001$) and that the target speaker was correctly identi-

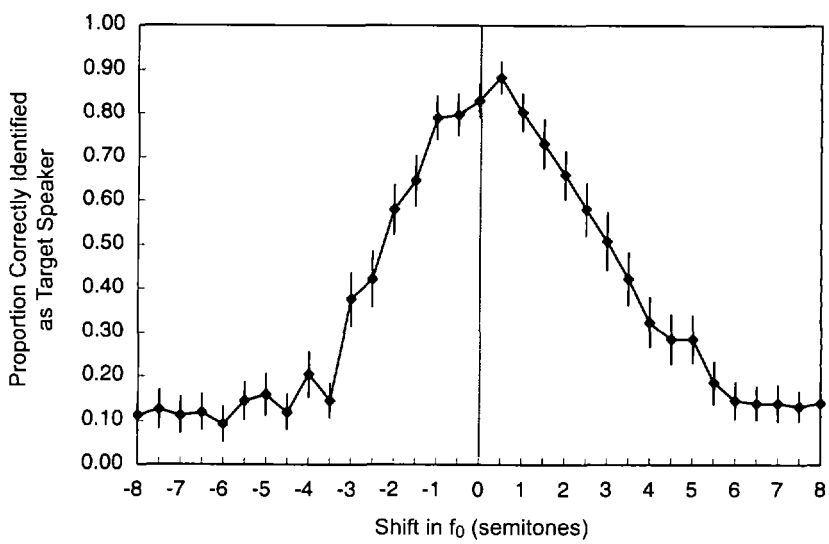

Fig. 1. Mean proportion of trials correctly identified as the original target speaker at each interval. Error bars represent ± 1 *SE* of the mean.

fied on a greater proportion of trials in which the frequency was increased ($M = .40$, $SD = .39$) than for trials on which the frequency was decreased ($M = .31$, $SD = .38$; $F_{1,37} = 5.93$, $p = .02$).

Here we explored the ability of listeners to identify frequency-shifted samples of a previously encoded voice. We hypothesized that listeners would identify voice samples that were increased in frequency as the original target voice more often than voice samples that were decreased in frequency. Our results support this hypothesis. At all 16 intervals voice samples with increased frequency had higher rates of recognition of target speaker than voice samples with equivalently decreased frequency.

These findings are consistent with other studies that have reported differential performance favoring higher frequencies and intensities (Church & Schacter, 1994; Lavner, *et al.*, 2000). Using familiar speakers, Lavner, *et al.* (2000) suggested that higher identification rates for voices with raised frequency occur because of the familiarity that listeners have with complex speaker-specific vocal characteristics. Here we have demonstrated analogous effects using unfamiliar speakers and relatively impoverished vowel utterances. Our results suggest that higher identification rates for voices with raised frequency do not require previous familiarity with the vocalizations of a particular speaker, complex memory schemata for familiar voices, or even complex utterances from which listeners might extrapolate such characteristics. Better identification of voices with raised frequency can occur with simple vowel vocalizations produced by unfamiliar speakers. Thus, if the better identification rates for raised vocal frequency are due to a familiarity with vocalizations at high frequencies, it may be that listeners are picking up more general cues in vocalizations that occur across speakers. Although we did not manipulate emotion in the current experiment, the relationship is clear between emotion and the frequency of vocalizations. For example, some work has shown that vocal frequency increases under stress (Utsuki & Okamura, 1976), that speech pitch can indicate emotional state (Levin & Lord, 1975; Yogo, Ando, Hashi, Tsutsui, & Yamada, 2000), and that listeners attend to vocal frequency when assessing emotion (Breitenstein, *et al.* 2001). Other work has shown that speakers emphasize important aspects of the speech stream by raising vocal pitch (Deutsch, *et al.*, 1990). Thus, it may be that the salience of increases in vocal pitch during important or emotional vocalizations plays a role in the better recognition of voices that are raised in frequency.

Conclusions

The current analysis shows that better recognition of raised frequency voices can occur without extensive familiarity with the vocalizations of a particular speaker and without complex utterances. Recognizing the identity of

a speaker by a vocal sample that has been raised in frequency more often than one that has been decreased in frequency may have adaptive significance. The phenomena may stem from the importance of recognizing emotional vocalizations or from an implicit knowledge of salient characteristics of the speech stream. Both phenomena are general characteristics not dependent upon the idiosyncratic characteristics of a particular voice. Thus, these results suggest that higher identification rates for vocal samples which are raised in frequency are due to more general vocal characteristics that occur across speakers.

REFERENCES

BAKEN, R. J., & DANILOFF, R. G. (1991) *Readings in clinical spectrography of speech*. San Diego, CA: Singular.

BREITENSTEIN, C., VAN LANCKER, D., & DAUM, I. (2001) The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample. *Cognition and Emotion*, 15, 57-79.

BROWN, R. (1981) An experimental study of the relative importance of acoustic parameters for auditory speaker recognition. *Language and Speech*, 24, 295-310.

CHURCH, B. A., & SCHACTER, D. L. (1994) Perceptual specificity of auditory priming: implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 521-533.

DEUTSCH, D., NORTH, T., & RAY, L. (1990) The tritone paradox: correlate with the listener's vocal range for speech. *Music Perception*, 7, 371-384.

FRICK, R. W. (1985) Communicating emotion: the role of prosodic features. *Psychological Bulletin*, 97, 412-429.

KUWABARA, H., & TAKAGI, T. (1991) Acoustic parameters of voice individuality and voice-quality control by analysis^synthesis method. *Speech Communication*, 10, 491-495.

LAVNER, Y., GATH, I., & ROSENHOUSE, J. (2000) The effect of acoustic modifications on the identification of familiar voices speaking isolated words. *Speech Communication*, 30, 9-26.

LEVIN, H., & LORD, W. (1975) Speech pitch frequency as an emotional state indicator. *IEEE Transactions on Systems, Man, and Cybernetics*, 5, 259-273.

PISONI, D. B. (1997) Some thoughts on "Normalization" in speech perception. In K. Johnson & J. Mullennix (Eds.), *Talker variability in speech processing*. San Diego, CA: Academic Press. Pp. 9-32.

SUNDBERG, J. (1999) The perception of singing. In D. Deutsch (Ed.), *The psychology of music*. (2nd ed.) San Diego, CA: Academic Press. Pp. 171-214.

UTSUKI, N., & OKAMURA, N. (1976) Relationship between emotional state and fundamental frequency of speech. *Reports of Aeromedical Laboratory*, 16, 179-188.

VAN DOMMELEN, W. A. (1987) The contribution of speech rhythm and pitch to speaker recognition. *Language and Speech*, 30, 325-338.

YOGO, Y., ANDO, M., HASHI, A., TSUTSUI, S., & YAMADA, N. (2000) Judgments of emotion by nurses and students given double-blind information on a patient's tone of voice and message content. *Perceptual and Motor Skills*, 90, 855-863.

Accepted August 18, 2003.