1985

Intelligibility of one's own speech relative to the speech of others

John W. Hawks

Intelligibility of One's Own Speech Relative to the Speech of Others

John W. Hawks
Central Institute for the Deaf

May, 1985

INTRODUCTION

A number of parallels appear to be present between avian vocal development and the development and perception of human speech (Marler, 1970). Many young birds experience a phase of vocal rambling known as subsong that has been equated to the babbling stage of infants (Nottebohm, 1970). For normal song to develop, young birds require exposure to the conspecific adult song during a 'critical period' early in life. The concept of a 'critical period' in human language development has long been exemplified in studies of traumatic aphasia and second language acquisition (Lenneberg, 1967). Nottebohm has demonstrated the lateralization of song learning in the chaffinch to one side of the hypoglossus nerve. Should this lateralization extend to higher regions as well, it may parallel the lateralization of human speech processing in the brain (Nottebohm, 1970).

Please do not remove from library -

The role of auditory feedback has also been demonstrated as a critical factor in song learning of certain species of birds. That is, some birds must hear their own voice for normal development of their songs to occur (Mulligan, 1966). From observations of speech development in the hearing-impaired, we may assume auditory feedback is also a critical factor for normal speech development in humans.

In studying White-Crowned Sparrows, Margoliash has recently found auditory neurons selective to the bird's own song in the Hyperstriatum Ventrale pars caudale (HVc), a forebrain nucleus in the motor pathway for control of song (Margoliash, 1985). There was no specific topographic organization to the neurons, but maximal excitation was observed in response to autogenous song when compared with responses to the tutor song, similar songs of the same dialect, and synthetic song. The vocal feedback experience seemed to explicitly modify the neuronal response, in that even a song which is a poor copy of the tutor song elicits the more maximal response.

Considering these findings and the previously stated parallels between avian and human vocal development, Margoliash has theorized that perhaps the human perceptual mechanism for speech is most sensitive for our own individual speech patterns. It is possible that the acoustic content of our own spoken phonemes is more closely correlated to our perceptual ideal of how that phoneme should sound to us than those spoken by others.

To test this theory, a speech recognition experiment was designed using the same subjects as both talkers and listeners. Subjects recorded word lists which were then randomized, mixed with a competing noise and played back to the subjects as a word identification test.

I. EXPERIMENTAL DESIGN

1. Subjects

Six females, ages 21-47, volunteered as subjects for the experiment. All reported having normal hearing at the time of a hearing test within the last year. None reported having any speech impediment and only one reported any noticeable dialectic influence on her speech. This subject, Talker #1, used some particular pronunciations associated with a New England dialect, however they were not noticeable in the words used for the stimuli recordings.

2. Stimuli

Using the CID Auditory Test W-22 (Hirsh, et al., 1952) word lists (lists C&D), 1ØØ monosyllables were recorded in a randomized order by each of the six subjects. The recordings were made in a sound-isolated booth using a Bruel & Kjaer 4179/266Ø condensor microphone and preamplifier. This signal was fed directly to a Sony PCM-F1 digital audio recorder with the digitally-encoded signal stored on video tape via a Fisher 72Ø video cassette recorder. Subjects spoke at conversational levels at a distance of 1/2 meter from the microphone and monitored their own voice levels with a sound level meter (7Ø dB

SPL = -9 VU). A carrier phrase and a number of practice trial words were also recorded at this time by a male talker.

These recordings were then digitized using the CID RAP computer system (Engebretson, 1977) and the individual words edited and stored on disks. Using Parapet (Hakkinen & Engebretson, 1979), a computer program for experiment design, the 600 words were balanced for equal intensity by visually observing a Ballantine Model 321 RMS-reading voltmeter and digitally attenuating each accordingly such that the peak energy of each word fell within a 1 dB range. The carrier phrase "Write the word" was inserted preceding each stimulus word at approximately the same volume level with a .5 second period of silence between carrier phrase and word. A silent interval of four seconds was used between trials with three short tones placed before each fifth trial to serve as a cue for maintaining trial order numerically.

Four test sessions of 150 words each were configured such that each of the monosyllables was presented once. To balance the test sessions for an equal number of stimuli from each talker and to include all 100 words per session, the initial randomized list of 100 words were ordered sequentially and divided into four cells of 25 words each (e.g. words 1-25 = List A; 26-50 = List B; 51-75 = List C; 76-100 = List D). A method of Latin squares (Fisher and Yates, 1949) was then used (Table 1) such that each session contained 25 words from each talker; all 100 words were used per session; and all 600 words were used only once by the end of four sessions. After preparing the stimuli in this manner, the stimuli for each session was randomized again by computer before making the final test tapes, via a Sony TC-645 reel-to-reel recorder.

		Talker					
		1	2	3	4	5	6
	I	D	C	B	A	D	C
Session	II	C	D	A	B	A	D
	III	A	B	D	C	B	A
	IV	B	A	C	D	C	B

Table 1. Method of latin squares used to order sessions.

3. Apparatus

Each subject was tested individually in the same sound-isolated booth where the recordings had been made. The stimuli and a speech noise were presented under a TDH-49 headphone to the right ear. A dummy phone was used on the left ear. A Grason-Stadler 901B Noise Generator was used as a speech noise source and as a mixer via an auxiliary input. The output of the noise generator/mixer was fed to a power amplifier as the final drive stage for the headphone (See Figure 1).

A signal-to-noise ratio of approximately -8 dB was found adequate to provide word identification scores of approximately 50% correct responses based on pilot tests with other listeners. The overall output level of the noise was maintained at 68 dB SPL as measured in a 6cc coupler with a B&K 2205 sound level meter.

4. Procedure

Subjects were provided with an alphabetical list of the 100 words and were instructed as follows:

1) Write the word you hear that follows the carrier phrase.
2) A word may be used more than once.
3) Use only words from the list provided.
4) Leave no blank spaces, guess if neccesary.

Subjects were allowed to view the VU meter of the recorder and were instructed in the use of the recorder's 'pause' switch to provide the option of additional time between stimuli.

5. Subject Conditioning

Prior to the first listening session, subjects were initiated to the test protocol with a short sample session using different monosyllables spoken by a male talker, but the same carrier phrase and stimulus conditions as the actual tests. Subjects were asked first to listen to the session without the noise masker and to concentrate on the structure and timing of the stimulus trials. The noise masker was then added and subjects were instructed to take the test as though it was an actual session. This sample session was repeated until the subjects felt comfortable with the testing structure.

II. RESULTS

Table 2 indicates the number of correct responses for all four listening sessions by listener and talker and the sums and means for the rows and columns. The average number of words correct per subject for the entire test was 138 out of 600 possible, or 23%. The author is unsure why this percentage is considerably lower than the 50% correct figure predicted from the pilot tests, as well as what effect it may have on the overall results of the experiment. The scores in bold-face underscored on the diagonal represent the overall scores for the listeners correctly identifying their own words. It is these scores relative to the off-diagonal scores which are of greatest interest.

		Talker #							
		1	2	3	4	5	6	E	X
Listener #	1	**35**	8	22	33	30	28	156	26.00
	2	34	**18**	19	32	24	30	157	26.17
	3	19	6	**20**	23	22	24	114	19.00
	4	27	6	16	**26**	20	23	118	19.67
	5	29	19	24	36	**29**	24	161	26.83
	6	22	10	16	28	24	**22**	122	20.33
	E	166	67	117	178	149	151	828	
	X	27.67	11.17	19.50	29.67	24.83	25.17		23.00

Table 2. Total correct responses by listener for all sessions. (E= sum; X= mean.)

The mean of the diagonal scores equals 25 (150/6), while the mean of the off-diagonal scores equals 22.6 (678/30), with a difference between the two mean scores of 2.4. Since each cell of the table represents the number correct out of 100 stimulus presentations, the scores can be thought of as percentages. This would indicate that the listeners performed 2.4% better at identifying their own words as opposed to the words of the others. Before considering the significance of these findings, an equation was formulated for predicting scores that would reflect a weighting for subjects' abilities as listeners and talkers:

$$P_{T+L} = (XC_T+XR_L)/2$$

Where the predicted score (P)for any combination of talker (T) and listener (L) is equal to the mean of the column mean of the talker (C_T) plus the row mean of the listener (R_L) from the obtained scores (Table 2). These predicted scores were calculated for each of the 36 possible combinations and then subtracted from the obtained scores (O) in Table 2. The results are shown in Table 3 and represent the difference between a listener's actual score for a given talker and what was predicted by the averaged abilities of the listener and talker for that combination. A positive number therefore, indicates the number of correctly identified words greater than was expected and a negative number, less than was expected.

Using this method, the difference between the mean of the diagonal difference scores (1.9983) and the mean of the off-diagonal scores (-.403) remains the same as obtained with the raw score data (1.9983-(-.403)= 2.4013), but now the inter-score differences become more apparent. As can be seen from Table 3, a large range of variance (8.16 to -10.59) exists among all the difference scores as well as within the two groups of interest, the diagonal and off-diagonal scores.

		Talker #					
		1	2	3	4	5	6
Listener #	1	8.16	-10.59	- .75	5.16	4.58	2.41
	2	7.08	- .67	-3.84	4.08	-1.50	4.33
	3	-4.34	-9.09	.75	-1.34	.08	1.91
	4	3.33	-9.42	-3.59	1.33	-2.25	.58
	5	1.75	0.00	.83	7.75	3.17	-2.00
	6	-2.00	-5.75	-3.92	3.00	1.42	- .75

Table 3. Differences for obtained (O) minus predicted (P) scores.

To convincingly demonstrate an advantage for self-recognition, we need to show that the diagonal scores of listeners hearing themselves are larger positive numbers and the off-diagonal scores of listeners hearing all talkers except themselves are small positive, or preferably negative numbers. This is not the case, however, as we find negative numbers in the diagonal scores, as well as relatively large positive numbers in the off-diagonal scores.

III. STATISTICAL ANALYSIS

Statistical analysis was performed on the weighted data using a within subjects t-test. The test compared identification performance when listening to oneself versus the mean identification performance for all other speakers, for individual subjects. The test results indicated no significant differences between the means (t(5) = 1.653, p>.Ø5). Thus, while a small percentage advantage in the self-recognition scores can be demonstrated, a statistically significant difference cannot.

IV. DISCUSSION

It should be acknowledged here that obvious differences exist between how our own speech sounds to us as opposed to the speech of others. This phenomenon is easily recognized by anyone having heard a tape recording of their own voice. These differences in perception are predominantly caused by the additional acoustic input contributed during autophonic hearing to the normal air-conducted route by bone and tissue conduction of one's own speech through internal pathways. In constrast, the speech signal from others is generally received only via the air-conducted route. The intensity, frequency, phase, and temporal parameters of internally-conducted sound have been studied (Bekesy, 1948, 1949; Stromsta, 1959, 1962), but their actual acoustic contribution during speech production has not been fully demonstrated.

For the purposes of this investigation, it was assumed that the additional acoustic information contributed during autophonic hearing produces a tilt in the speech spectrum, but does not appreciably change the spectral envelope of individual speech sounds. Therefore, the acoustical properties of the speech signal used for perception of self and others is assumed to remain essentially the same.

The concept of an auditory template for the learning of song by certain sparrows has been established (Marler,1970). The results from studies by Eimas and colleagues (Eimas et al., 1971; 1975) suggest that infants may also possess auditory templates for at least certain speech sounds that may exist to focus the infant's attention and provide a reference for listening to the speech of others. This template might also, as Marler has suggested (Marler, 1975), provide a model for the development of speech production. As a perceptual auditory pattern, the innate or inherited template may serve as a triggering mechanism for attentiveness to the conspecific song in song birds (Marler & Tamura, 1964), and attentiveness to speech vs. non-speech sounds in human infants (Marler & Peters, 1981).

If we are born with innate auditory templates, the modification of these templates for speech perception may well be heavily based on auditory feedback from our own speech. It is known that this feedback mechanism is critical to normal speech and language development, and it is probably safe to assume that after the first year or so of life, the perception of our own voice occurs more frequently to us than the perception of any other voice. It seems reasonable then, that the central reference for our individual perception of speech sounds might be uniquely correlated to the perceptual patterns of our own voice, despite our ability to perceptually categorize acoustically varying phonemes as the same.

While the results of this experiment do not statistically show an advantage for what might be termed 'autophonetic' perception, neither do they negate the possibility of its existence. Many variations in the experimental design can be envisioned which might prove capable of demonstrating a larger effect. Practical applications of autophonetic advantages could well influence current thinking on automated speech recognition and speech encoding for devices such as digital hearing aids and vibrotactile vocoders.

Dependent on the conditions under which a significant effect might be found, several implications relevant to current speech perception theories could be considered. As applied to the theories of Liberman and associates (Liberman, et al., 1967), the autophonetic effect could be tested for possible correlation with self-specific variations in speech production, which could serve as further support for the relation between the mechanisms of speech production and speech perception. Investigators of the pattern matching of spectral envelopes proposed by Klatt (Klatt, 1979) may wish to consider the possible relevance of autophonetic spectra as the individualistic ideal pattern match. In considering the theory of Miller (Miller, 1982) and speech perception within a phonetically relevant auditory-perceptual space, one might test the possible relation between the autophonetic correlates and the centers of the "learned target zones."

Acknowledgements

The author would like to acknowledge and thank the following for the parts they played in this study. Dr. Daniel Margoliash, for suggesting the hypothesis; Dr. James D. Miller, for his aid and guidance throughout the project; Dr. Janet Weisenberger, for her assistance with the statistical analysis; and the six friends and colleagues who found time in their busy schedules to serve as subjects.

Bekesy, G.v. (1962). "The gap between the hearing of external and internal sounds," Symp. Soc. Exp. Biology XVI, 267-288.

Bekesy, G.v. (1949). "The structure of the middle ear and the hearing of one's own voice by bone conduction," J. Acoust. Soc. Am. 21, 217-232.

Bekesy, G.v. (1948). "Vibration of the head in a sound field and its role in hearing by bone conduction," J. Acoust. Soc. Am. 20, 749-760.

Eimas, P.D. (1975). "Speech perception in early infancy," in Perception, edited by L.B. Cohen and P. Salapatek (Academic Press, New York), pp. 193-231.

Eimas, P.D. (1985). "The perception of speech in early infancy," Sci. Am. 252(1), 46-52.

Engebretson, A.M. (1977). "Computer system for auditory research," J. Acoust. Soc. Am. 62 (Suppl. 1), p. 12.

Fisher, R.A. and Yates, F. (1949). Statistical Tables for Biological, Agricultural and Medical Research. London: Oliver and Boyd.

Hakkinen, M., & Engebretson, A.M. (1979). "PARAPET: A general utility program for RAP-III," Central Institute for the Deaf Periodic Progress Report (22, p. 30). St. Louis, MO: Central Institute for the Deaf.

Hirsh, I.J., Davis, H., Silverman, S.R., Reynolds, E.G., Eldert, E., Benson, R.W. (1952). "Development of materials for speech audiometry," J. Speech Hearing Dis. 17, p. 321-337.

Klatt, D.H. (1979). "Speech perception: A model of acoustic-phonetic analysis and lexical access," J. Phonetics 7, 279-312.

Konishi, M. (1965). "The role of auditory feedback in the control of vocalization in the White-Crowned Sparrow," Z. Tierpsychol. 22, 770-783.

Lenneberg, E.H. (1967). Biological Foundations of Language. New York: Wiley.

Liberman, A.M., Cooper, F.S., Shankweiler, D.S., and Studdert-Kennedy, M. (1967). "Perception of the speech code," Psychol. Rev. 74, 431-461.

Margoliash, D.(1984). "An auditory representation of autogenous song in the White-Crowned Sparrow,"

Marler, P. (1981). "Birdsong: the acquisition of a learned motor skill," TINS, Apr., 88-94.

Marler, P. (1975). "On the origin of speech from animal sounds," in The Role of Speech in Language, edited by J.F. Kavanagh and J.E. Cutting (MIT Press, Cambridge, MA), pp. 11-37.

Marler, P. (1970). "Birdsong and speech development: could there be parallels?," Am. Sci. 58, p. 669-673.

Marler, P. & Peters, S. (1981). "Birdsong and speech: Evidence for special processing," in Perspectives on the Study of Speech, edited by Peter D. Eimas and Joanne L. Miller, (Lawrence Erlbaum Associates, Hillsdale, New Jersey), pp. 75-112.

Marler, P. & Tumara, M. (1964). "Culturally transmitted patterns of vocal behavior in a sparrow," Science 146, 1483-1486.

Miller, J.D. (1982). "Implications of the auditory-perceptual theory of phonetic perception for speech recognition by the hearing impaired." Read before the Workshop on Speech Recognition by the Hearing Impaired sponsored by the National Institute of Neurological and Communicative Disorders and Stroke, Bethesda, MD, Sept. 24, 1982.

Mulligan, J.A. (1966). "Singing behavior and its development in the song sparrow Melospiza melodia," Univ. of California Publications in Zoology 81, p. 1-73.

Nottebohm, F. (1970). "Ontogeny of bird song," Science 167, 950-956.

Stromsta, C. (1962). "Delays associated with certain sidetone pathways," J. Acoust. Soc. Am. 34, 392-396.

Stromsta, C. (1959). "Experimental blockage of phonation by distorted sidetone," J. Speech and Hearing Res. 2, 286-301.