ON THE IDENTIFICATION OF FAMILIAR VOICES

Judith Rosenhouse
Dept. of General Studies

Yizhar Lavner and Isak Gath
Dept. of Biomedical Engineering

The Technion, I.I.T., Haifa, Israel

ABSTRACT

This study aimed at examining voice identification levels in and by a relatively large group of subjects. These subjects had a high degree of familiarity as they lived in the same kibbutz. We analyzed 5 modern Hebrew vowels, 5 voiced consonants, and the Hebrew phrase "good morning". The results suggest differential perception and identification processes among subjects and among speech elements. We suggest an explanation to the results in the framework of the prototype model.

PROBLEM

Voices of different people sound different, and a few seconds of speech suffice to identify a speaker without being in direct visual contact with him/her. Speaker identification is possible under various detrimental circumstances, after long time periods, in various speech contexts, when the speaker expresses different attitudes, etc. These facts seem to imply that there are some acoustic features that do not change under different conditions. Despite these observations, the topic is still not well understood as to the parameters that affect speaker identification or how much they are related to phonetics.

Human speaker recognition has been defined as any process of decision about the identity of a speaking person by certain features of the speech signal. A presupposition for such a decision is previous acquaintance with the speaker.

Acoustic characteristics of voices have thus been described, inter alia, in long-term- and short term-, inherent- and learned-, glottal source- and vocal tract-dependent features.

Listeners' role in voice identification has been studied so far from numerous angles, e.g., voice recognition learnability, time effect on recognition (memory), number of voices recognized, test type, utterance type/length effect on recognition, language dependence, masking effects, the effect of various acoustic parameters on recognition, etc. Experiments often used small numbers of subjects and/or voices.

This paper has the following goals: 1. testing voice identification of a large group of speakers by a large group of listeners well acquainted with the speakers; 2. Phoneme-dependency of voice identification. Reported here are results of our psycho-acoustic tests, performed as part of a systematic study of acoustic cues important for voice identification as described.

METHOD AND SUBJECTS

The method includes a few stages for recording the test material and testing the subjects.

The subjects were from a kibbutz in the north of Israel, all native speakers of Hebrew without speech or hearing impairments or foreign features.

In Stage I, 20 men of this kibbutz (age range: 26-59) were recorded saying the same test materials (see below). They were recorded (mono-channel) on a 486 PC computer using a voice card at 22 kHz sampling rate.

In Stage II the listeners were men and women from the same kibbutz (age range: 25-55). All know the speakers well or very well. Each listener was asked to fill in a form grading in a 5 grade scale his/her acquaintance with the speaker. There were 28 people's voices, 20 of which were later used in the tests. Subjects had to grade in a 5 grade scale the "uniqueness" of each speaker's voice, and describe this feature in words.

Speaker identification tests included recognition by: 1. /a, e, i, o, u/, the 5 vowels of Hebrew uttered in isolation; 2. /aCa/ syllable sequences, C being the nasals /m,n/, based on the literature which described them as good predictors of individual features and /l, r, z/ which tend to have numerous allophones; and 3. the 2-word Hebrew utterance /'boker 'tov/, i.e., 'good morning'.

Each session lasted up to an hour and a half, in which each listener heard a 100 vowels of 20 different speakers in random order. The listeners were asked to identify the speakers from a list of 28 people's names, i.e., more speakers than actually used in the test. After hearing the vowels, they had to fill in another questionnaire in which they wrote down the speaker's name (as they identified him), their confidence level in it, their evaluation of this voice's uniqueness and (optionally) a verbal description of the voice. They were allowed up to 8 times listening to each stimulus. On the same session they were also asked to identify speakers by the utterance "good morning". Speaker identification by /aCa/ syllables was tested in a separate session.

The tests were aimed to give answers to the following questions: 1. What is the average speaker identification level by individual listeners? 2. Are there inter-listener differences in speaker identification? 3. Are there inter-listener differences in speaker identification by different phonemes? 4. Does successful idenification of a voice by a certain phoneme imply successful speaker identification by the same phoneme by other listeners?

FINDINGS

1. The identification test of the utterance /'boker 'tov/ yielded an average correct identification rate of 60% (216 correct identifications for 360 stimuli). The test was an open test, in the sense that the listeners did not know which speaker out of 28 possible people they were about to hear. This proportion of successful identification is much higher than reported in previous studies (see e.g., Ladefoged & Ladefoged, 1980, van Lancker et a., 1985). The listeners identified speakers, except for three cases where the voices of certain speakers were erroneously considered those of others. Thus, most errors were of the type "cannot identify". In addition, the successful identification range by individual listeners (45% - 85%) is smaller than the successful identification range of the speakers' voices (11% - 100%).

2. Speaker identification by voice recognition of vowels: The results of 20 listeners (men and women) were included in the data analysis and are summarized in Table 1 and demonstrated as an example in Table 3. The results show that there are vowel-dependent significant differences in listeners' identification abilities. The best identification was yielded for /a/ - 37.6%. Next come /i/ and /e/ without any difference between them - 29%. These three vowels are better identified than /o/ (25%) and /u/ (17%) (See Table 1). The average identification rate for the total number of vowels was 29% (range: 16%-51%), which is much lower than for the words (60%). This result may be expected owing to the little information in isolated vowels as compared to two-word-utterances.

Table 1. Percentage of speaker idetification by vowels

vowel	N	correct	percent
/a/	330	124	37.6%
/e/	295	89	30.2%
/i/	300	87	29.1%
/o/	285	72	25.3%
/u/	189	32	16.9%

3. Speaker identification by voice recognition of voiced consonants in /aCa/ syllables: Differences were also found for identification of speakers in this environment. The best identified consonant in this environment was /z/ (63%), followed by /n/ (62%) and /m/ (58%). The speakers of syllables with /l/ were correctly identified in 53% of the cases, and for /r/ - in 50% of the stimuli (see Table 2).

Table 2. Percentage of speaker idetification by consonants

consonant	N	correct	percent
/r/	198	99	50.0%
/l/	198	105	53.0%
/m/	198	114	57.6%
/n/	197	123	62.4%
/z/	198	124	62.2%

4. Confusion matrix results revealed that some of the speakers were more successfully identified for certain vowels than other speakers (See for example, Table 3).

5. Speakers whose voices were correctly identified by all the listeners in all the phonemes had special vocal features.

6.Most speakers had considerable pitch variations even for relatively short utterances (<300 ms.). But as F0 ranges were very similar for most speakers, it may be assumed that F0 is not the most important cue for speaker identification, at least in isolated vowels.

DISCUSSION

This paper presents results of a study of speaker identification by human listeners. The test lanaguage was Hebrew, which to the best of our knowledge, has so far not been studied from this respect. For the purpose of this study both speakers and listeners were native speakers of this language.We are dealing here with Modern Hebrew,a Semitic language with a phonetic system comprisng 5 vowels and 20 consonants (traditionally 22 consonantal and 10 vowel phonemes). This is apparently the first report on this issue based on such a new source of database.

Another issue that this study tackles is the number of subjects used in the tests. Most previous experiments used very small numbers (e.g., 3,5,7) of listeners and/or recordings. The present experiments were performed with a much larger group of subjects, both speakers and listeners, and thus results are probably more valid.

The tests can be considered of the open-test type, in the sense that the listeners did not know which persons of the list were going to be heard. In this sense, this test type is closer to the real-world situation of speaker identification.

At least two basic models for speaker identification by listening can be suggested:

1. All listeners use one and the same voice identification strategy using the same features.

2. Different listeners use different strategies to identify speakers' voices.

The results of our tests suggest a third model which combines the above two to some extent:

3. Listeners use the same strategy for speaker identification but different acoustic features of the speakers' voices. This model is based on the protype model (Rosch 1973, Rosch, 1976). According to this model, learning a new voice is achieved by comparing it to a prototype voice (e.g., mens' vs. women's or children's voices) and extracting from this comparison those features which deviate from the prototype pattern. Thus, voices which are less similar to the prototype will be easier to learn and memorize than voices similar to it. Thus, the more a voice deviates from the prototype, it will also be easier to identify it when presented as a stimulus for identification, and vice versa: the more similar it is to the prototype the harder it will be to identify it. The results of our experiments and of other experiments reported in the literature make this a likely hypothesis. This hypothesis is also useful, for it allows predicting results of other experiments.

Further research is required to prove whether this model is correct. Current research in speech sciences often applies acoustic analysis of speech signals and systematic resynthesis while controlling individual features and observing listeners' responses. We intend to use this method to examine the prototype model in the next stage of our study of voice identification.

REFERENCES

1. Ladefoged, P. and J. Ladefoged (1980) "The Ability of listeners to identify voices" UCLA Working Papers in Phonetics, 49, 43-51.

2. van Lancker D., J. Kreiman and K. Emmorey (1985) "Familiar voice recognition: patterns and parameters, Part 1: Recognition of backward voices" Journal of Phonetics, 13, 19-38.

3. Rosch, E. (1973) "On the internal structure of perceptual and semantic categories" in: T. E. Moore, (ed.) Cognitive Development in the Acquisition of Language, New York: Academic Press.

4. Rosch, E., C.B. Mervis, W.D. Gray, D.M. Johnson and P.Boyes-Braem,(1976) "Basic objects in natural cetegories", Cognitive Psychology, 8, 382-439.

Table 3. Stimulus-Response Confusion matrix for the vowel /a/

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	other	unrec	total
	1				7				2											1	1					2	3	16
	2	4	4										1					1			1					2	4	17
	3			12														1				2						16
S	4				12		1								2											1	1	16
	5					8								1									3			3		15
T	6				1		5							1												4	5	16
	7							4									4			3						4	1	16
I	8	3			3				5																		5	16
	9									6						1										2	7	16
M	10		1					1			7									1						1	5	16
	11								1			15															1	17
U	12				1			1					8								1					3	3	17
	13	1						2	6	1				1	1						1					1	3	17
L	14				2										4	1					1					3	6	17
	15		1		1			1								4										1	8	16
U	16				2				1								4			1						3	3	14
	17				1				1			2	1					3							7	1	1	17
S	18													1					13							1	1	16
	19				4				1															4		4	3	17
	20		1						1												8					4	3	17
	total	8	7	12	34	8	6	9	18	7	7	18	9	4	7	6	8	5	13	6	13							325
	false	8	3	0	22	0	1	5	13	1	0	3	1	3	3	2	4	2	0	6	5	2	3	4	7	40	138	

1÷24 - SUBJECTS' I.D. NUMBERS; OTHER: VOICES NOT USED AS STIMULI; UNREC: UNRECOGNIZED VOICES