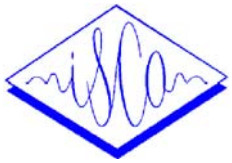

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

The prototype model in speaker identification

Yizhar Lavner (1) (2), Judith Rosenhouse (3) and Isak Gath (1)
(1) Dept. of Biomedical Engineering, Technion, Haifa, ISRAEL
(2) Dept. of Computer Science, Tel-Hai Academic College, ISRAEL
(3) Dept. of General Studies, Technion, Haifa, ISRAEL

ABSTRACT

Little is known on the perceptual processes of speaker identification and its relations to the acoustic features of the speaker's voice. A study of speaker perception and identification by psycho-acoustic experiments was carried out. Statistical analysis of the results suggests that the prototype model is appropriate for explaining the process of speaker identification. It has been also found that the most important features for speaker identification were the fundamental frequency, the third and fourth formants and the closing phase of the glottal wave. For different listeners, different sets of features were found to be significant for coding speaker identity.

1. INTRODUCTION

The extraordinary ability of humans to recognize many familiar people by their voices is exceptional both in its accuracy and adaptivity: The identification can be carried out within different conditions, (e.g. through noisy channels), and can be age and context independent.

In the present study are reported findings of unique acoustic features that convey information pertinent to speaker identity. Subsequently, a psychophysical model of familiar speaker identification is suggested and the experimental predictions of this model examined. The question whether each listener has a different group of perceptive acoustic features, or different listeners use the same acoustic features for the identification, is also discussed.

In addition to the contribution of the study to basic research of speaker perception, the results are relevant to several practical issues, such as automatic speaker identification, voice conversion systems, and natural voice synthesis.

In a series of preliminary studies (1,2) it was found that speaker identification rates varied on a wide scale. In addition, listeners tended to choose incorrectly the same speakers for other speakers utterances (false positive), these speakers denoted "Default voices". Regarding these findings and some other studies reported in the literature (3,4,5), it was examined whether the 'Prototype Model' (3,4) could explain the results and produce experimental predictions.

The prototype model assumes the presence of a prototype ('average') voice in the memory of each listener, comprised of an ensemble of acoustic features, related to the language, the accent, the phonemes and allophones and to the voice production system. For each new voice, only those features that significantly deviate from the prototype are stored (memorized), and identification of familiar voices is based on searching and locating the deviated features. As a consequence, because different speakers have somewhat different voice production systems, it is reasonable to assume that the identity of different speakers is cued by different sets of acoustic features. Indeed, in previous studies, evidence for this assumption (2) has been found.

2. METHODS

2.1 Speakers, Listeners, Recording session

Part of the methods used in this study are reported elsewhere (1,2). Both speakers and listeners who participated in this study were members of an Israeli kibbutz for at least 5 years, ensuring a high level of familiarity between them. Twenty male speakers (age range 26-59) were recorded, all native speakers of Hebrew without any known speech defects. Five isolated Hebrew vowels (/a/, /e/, /i/, /o/ and /u/) were investigated, but in this paper the results of the psychoacoustic experiments with /a/ only will be reported.

Speech was recorded using a condenser microphone (ACO 7040), and digitized with a 16 bit high quality sound card at 11 kHz sampling rate.

Thirty listeners, 22 females and 8 males aged 15-58, participated in the psychoacoustic experiments. A short questionnaire was given to each listener before the experiment. The listener had to rate the level of familiarity with the speaker and the uniqueness of the speaker's voice on a 1-5 scale.

The psychoacoustic experiments were carried out in a quiet room. In each session, the listener had to select his/her choice from a list of 29 people, including 9 who were not actually recorded. The additional names were added to the list to make the test more realistic and open-set-like. Listeners were told that on each occasion any speaker could be heard once, more than once, in succession, or even not at all. Each experiment consisted of two parts. In the first part only natural voices were heard, while the second part contained mainly modified voices, of those speakers who had been identified in the first part. The voices were played at a random order in both parts, so that each listener attended to a totally different sequence of voices.

After each selection , the listener had to indicate his confidence in the selection, by using a scale of 1-5 and the subjective naturalness of the voice, on the same scale, where 5 meant natural voice, and 1 meant a completely artificial synthetic voice. In case of the listener being unable to recognize the voice, he had the possibility to select the "non identified" push-button.

10.21437/Eurospeech.1999-187

All the stimuli and the listener selections were automatically recorded and analyzed by the computer.

2.2 Analysis: Glottal pulse, Pitch and Formants

The purpose of the analysis stage was to estimate the major acoustic parameters of the speakers voices, that are considered to determine the unique voice quality of each speaker. The range and the distribution of the parameters serve as references for the modification stage. The analysis/synthesis system was based on a linear model of the speech production. The system enabled analysis and synthesis of voices while controlling the various acoustic features: the first 4 formants, each separately or in various combinations, the glottal excitation waveform, and the fundamental frequency. The main component of the analysis section consisted of an iterative pitch synchronous inverse filtering algorithm, a variation of the PSIAIF algorithm (6). Both the vocal tract transfer function and the glottal excitation wave were estimated by this method.

3. RESULTS

3.1 Identification of natural voices

In these experiments 20 speakers were identified by 30 listeners. The average percentage of all the listeners for all the stimuli was 49.6%. Table 1 describes the scores of individual listeners. High Variability in both speakers' and listeners' percentage rates was found (speakers' range 11.6%-93.0%, listeners' range 15.0%-79.0%)

Listener	Correct	N	%
1	79	100	79.0
2	71	100	71.0
3	65	100	65.0
4	25	40	62.5
5	61	99	61.6
6	46	80	57.5
7	57	100	57.0
8	56	100	56.0
9	44	80	55.0
10	10	20	50.0
11	29	60	48.3
12	29	60	48.3
13	38	80	47.5
14	46	100	46.0
15	35	80	43.8
16	43	100	43.0
17	17	40	42.5
18	17	40	42.5
19	32	80	40.0
20	15	40	37.5
21	14	40	35.0
22	7	20	35.0
23	6	20	30.0
24	22	80	27.5
25	5	20	25.0
26	5	20	25.0
27	5	20	25.0
28	4	20	20.0
29	3	20	15.0
30	3	20	15.0

Table 1: /a/, identification scores, individual listeners.

3.2 The relation between acoustic features and speaker identification

The relation between identification percentages of individual speakers and two perceptive features, the subjective familiarity and the subjective uniqueness of the speakers' voices, as ranked by the listeners, were examined. No significant correlation was found between the identification rate and any of the features. It could be, therefore, concluded that the difference in identification rate between speakers was due to differences in their acoustic features, and not because of different familiarity.

According to the prototype model, the prototype pattern is assumed to be represented by averages of perceptually prominent acoustic features in the speaker population. The rate of correct identification for a given speaker depends on the deviations of the speakers' acoustic parameters from the average, speakers with many significant deviate features are expected to have high identification scores, and vice versa. To validate these predictions qualitatively, profiles of the deviations of the features from the population average for each speaker were drawn (Fig.1), along with the identification scores. Voices with many deviates from the average were identified by most listeners (Fig.1, upper four traces, from above: 89%, 93%, 80% and 69% respectively), whereas the voices of ITC and TSF, whose features are close to the average, scored low percentages (Fig.1, lower two traces, 12% and 29%, respectively). It can be concluded, that at least qualitatively, the identifiability of a given voice can be determined by his deviations' profile.

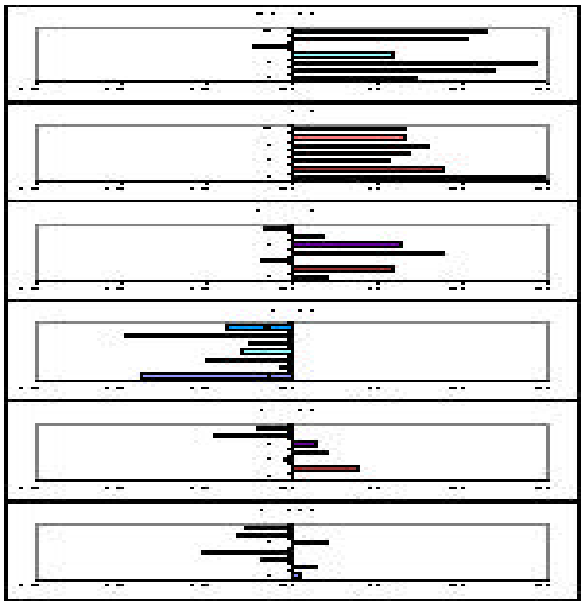

Figure 1: Deviation maps of 6 speakers. The horizontal axis represents the normalized deviation from the average. In each map, the bars represents the deviation of the following features: F_0, F_1, F_2, F_3, F_4, OQ, CQ.

3.3 Scatter diagrams of speaker identification as a function of various acoustic features

The relation between the speaker identification percentage in natural /a/ vowels and some of the acoustic features (F0, F1, F2, F3, F4, OQ, CQ, SQ, CORR) were examined. Weak relations were found to exist between the identification rate and two features:

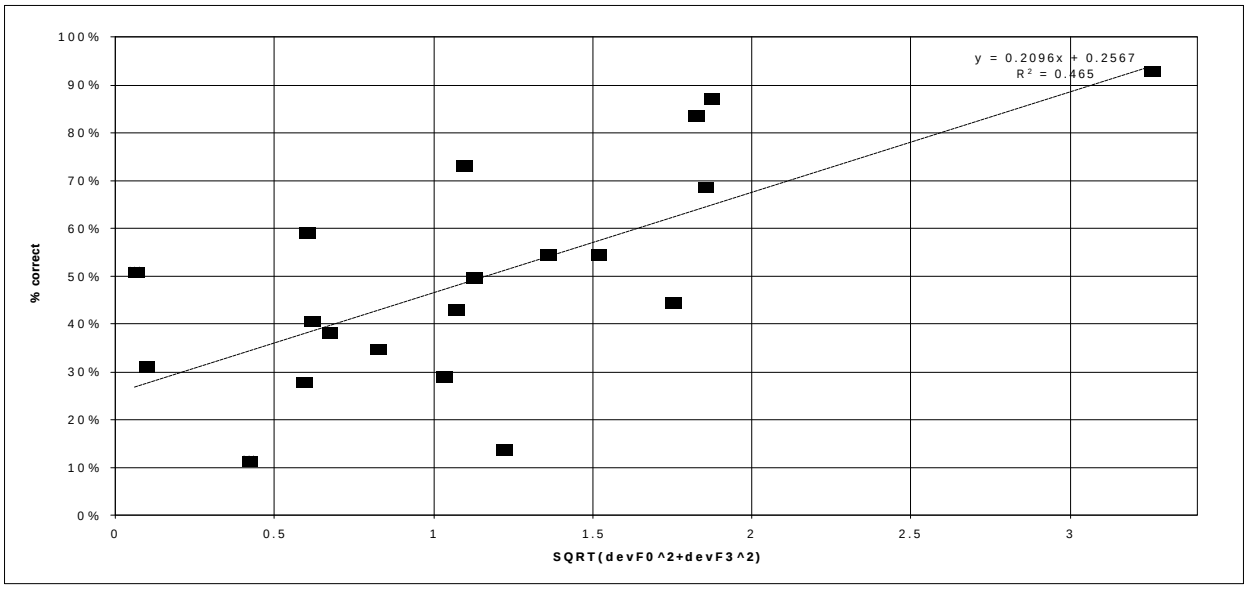

Figure 2: Scatter diagram of the identification rate of each speaker versus the Euclidean distance from the origin in two dimensional space comprised of the deviations of two features: F_0 and F_3.

F0 and F3. Identification of speakers with the most deviated features was clearly higher than others, especially for the upward deviation. However, a significant correlation between the identification percentage and a combination of more than one feature (for example F_0 and F_3, $r = 0.70$, Fig. 2) could be verified. A more conspicuous connection between the identification rate and combinations of various features was found by plotting a scatter diagram of the identification rate as a function of algebraic sum of deviations of features (Fig. 3).

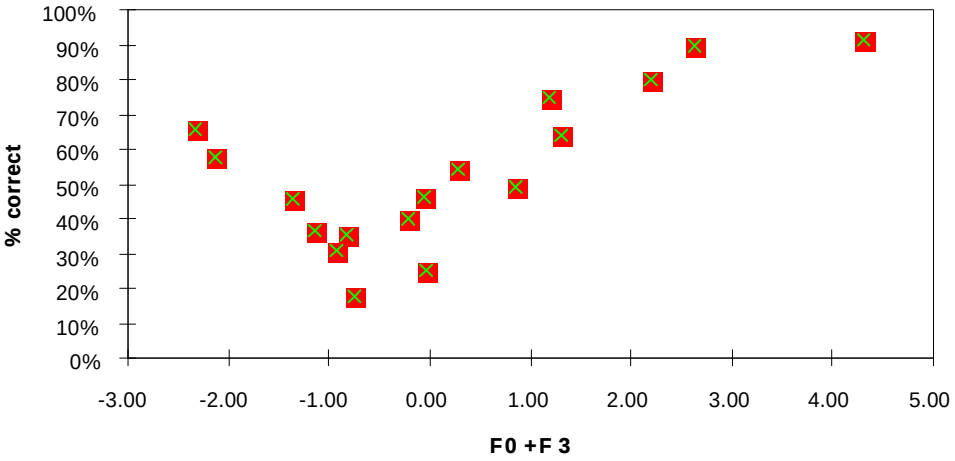

Figure 3: Scatter diagram, identification rate of each speaker versus the sum of deviations of two features: F_0 and F_3.

3.4 Logistic regression analysis

The purpose of this analysis was: 1) to explain the identification of speakers by various acoustic parameters, and 2) to examine whether different listeners use the same acoustic features for speaker identification. To this aim, logistic regression analysis technique was utilized, in which the dependent variable was the identification correctness and the explanatory variables were the acoustic features of each speaker.

The data of the responses of all the listeners to the voices of all the speakers was submitted to the regression

analysis. In this analysis 1260 stimuli were used (average identification: 48.9%).

Each speaker was represented in the model by his set of average acoustic features: The fundamental frequency (F_0), the first 4 formants (F_1-F_4), the opening and closing quotients of the glottal pulse (OQ and CQ, respectively), the speed quotient (SQ=OQ/CQ) and the similarity coefficient (CORR, the average correlation coefficient between adjacent periods). The purpose of the analysis was to explain the response variable (identification correctness), which takes two possible variables (0 or 1, i.e., correct or incorrect identification), by a set of these acoustic features. A stepwise regression analysis was utilized for finding the statistically significant variables. The listeners effect was taken into account by defining a set of binary variables (N-1, whereas N is the number of listeners, and each binary variable is indicator for each listener), and checking for their significance as a group. The variables were found to enter as a group.

The significant parameters for the model are given in Table 2. The first and second powers of the fundamental frequency and the closing quotient, the third and forth formant, and the similarity coefficient, were found to be significant.

Pr>Chi	Chi square	Std Err	Estimate	Parameter
0.0001	16.7302	0.1204	-0.4926	F_0
0.0001	18.8429	0.0659	0.2860	F_0^2
0.0001	108.2205	0.0945	0.9831	F_3
0.0003	13.1940	0.0959	0.3485	CQ
0.0001	19.3935	0.0591	0.2601	CQ^2
0.0032	8.7112	0.1014	0.2992	F_4^2
0.0029	8.8748	0.0495	-0.1476	$CORR^2$

Table 2: Logistic regression analysis - results for all speakers and listeners. The table contains statistically significant features only.

In addition, logistic regression analysis was performed on the results of each listener separately. The purpose of this analysis was to examine if different listeners use the same set of acoustic features for the identification process, or whether each listener uses different features. This was obtained by investigating which of the features was statistically significant for each listener. The results of a stepwise analysis that was carried out on the identification scores of individual listeners showed that in general, different acoustic features were found to be significant for different listeners. The only common feature was the third formant, which was significant for 5 out of 11 listeners for whom the analysis was performed. Consequently, it is possible to conclude from these results, that each listener has a somewhat different set of features, to which he/she is most sensitive, and uses for the identification process.

4. DISCUSSION AND CONCLUSIONS

It has been shown in this paper that the identification of speakers depends on the deviations of the speaker's acoustic features from an estimated average. The stronger the acoustic features of a certain speaker deviate from the population average, the easier will he be identified, and vice-versa. This result is in accordance with the predictions of the prototype model. In addition, a logistic regression analysis shows that on average, the most important features for identification are F0, F3, F4, and the closing quotient of the glottal pulse. On the other hand, the same analysis on each listener separately, indicates that it is reasonable to assume that each listener has his/her own unique set of characteristics for speaker identification.

REFERENCES

1. Rosenhouse, J., Lavner, Y., and Gath, I. (1995). "On the identification of familiar voices", Proc. ICPhS 95, Stockholm, 1, 190-194.
2. Lavner, Y., Gath I. and Rosenhouse, J., (1997). Acoustic feature and perceptive processes in the identification of familiar voices", Proc. Eurospeech '97, Rhodes, Greece, vol. 5 ,2311-2314.
3. Van Lancker, D., Kreiman, J., and Emmorey K., (1985a). "Familiar voice recognition: patterns and parameters, Part I: Recognition of backward voices," J. of Phonetics13, 19-38.
4. Van Lancker, D., Kreiman, J., and Wickens, T.D., (1985b). "Familiar voice recognition: patterns and parameters, Part II: Recognition of rate altered voices," J. of Phonetics13, 39-52.
5. Papcun, G., Kreiman, J., and Davis, A. (1989). “Long-term memory for unfamiliar voices,” J. Acoust. Soc. Am., 85 (2), 913-925.
6. Alku, P., (1992). “Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering,” Speech Communication, 11, 109-118.