Acoustic Features and Perceptive Processes in the Identification of Familiar Voices

Yizhar Lavner, Isak Gath* and Judith Rosenhouse***
* Dept. of Biomedical Engineering and ** Dept. of General studies,
Technion, Israel Institute of Technology, Haifa, Israel

Abstract
The present study aims at examining the relative importance of various acoustic features as cues to familiar speaker identification. The study also attempts to examine the validity of the prototype model, as the key to human speaker recognition. To this aim 20 speakers were recorded. Their voices were modified using an analysis-synthesis system, which enabled analysis and modification of the glottal waveform, of the pitch, and of the formants. A group of 30 listeners had to identify the speakers in an open-set experiment. The results suggest that on average, the contribution of the vocal tract features is more important than that of the glottal source features. Examination of individual speakers reveals that changes of identical features affect differently the identification of various speakers. This finding suggests that for each speaker a different group of acoustic features serves as cue to the vocal identity, and along with other predictions that were found to be valid, supports the adequacy of the prototype model.

Introduction
Human listeners have an extraordinary ability to identify numerous familiar voices, under varying conditions and contexts, in a manner that algorithms for automatic speaker recognition can hardly achieve. Still, not much is known about the link between the acoustic features of the speakers' voice and higher processes of speaker identification by the listener.

The common attitude for finding the important acoustic features for speaker recognition is by analyzing and resynthesizing the speech signal, while controlling different acoustic parameters. Very few studies attempt to investigate directly the relationship between individual vocal features and speaker identification ([1],[2]). Unfortunately, the small number of both speakers and listeners which were used in these studies, limits the possibility of drawing general conclusions.

The present study investigates extensively the contribution of each of various acoustic features to speaker identification. The speakers' voices have been modified using an analysis-synthesis system, in which the speech was separated into the glottal excitation source and the vocal tract transfer function. The influence of the glottal wave signal, of the first four formants and of the fundamental frequency on speaker identification was evaluated. In addition to evaluating the relative importance of the acoustic features, the experiment can lead to a better insight into the mechanisms and strategies of listeners in speaker recognition.

The prototype model ([3],[4],[5],[6]) suggests that in each listener's memory there is a representation of a prototype voice. Identification of a person by his/her voice is achieved by extracting features that deviate significantly from this prototype. Thus, another goal of this research was to examine the validity of the prototype model to human speaker recognition.

The main research questions that motivated this study were: 1. What are the most important acoustic features that convey information of speaker individuality? Are all the speakers coded by the same acoustic features, or does each speaker possess a unique set of features? 2. What is the strategy of the listener in speaker identification? 3. Could these findings concerning human perceptual strategy be used for automatic speaker recognition?

Methods
The work involved 3 main stages: Recording the speakers, processing the speech material and modifying various acoustic features, and examining listeners' responses to both natural and modified voices.

Isolated vowels from 20 male speakers were recorded using a condenser microphone (ACO 7040).

Thirty listeners participated in the psychoacoustic experiments. Both speakers and listeners had been living for at least 5 years in the same kibbutz, a fact that ensured a high level of familiarity between them.

In an interactive computer program, each listener was instructed to select his choice from a list of 29 speakers, including 9 who were not recorded. In the first part of each experiment only natural voices were heard, while in the second part mainly modified voices were presented.

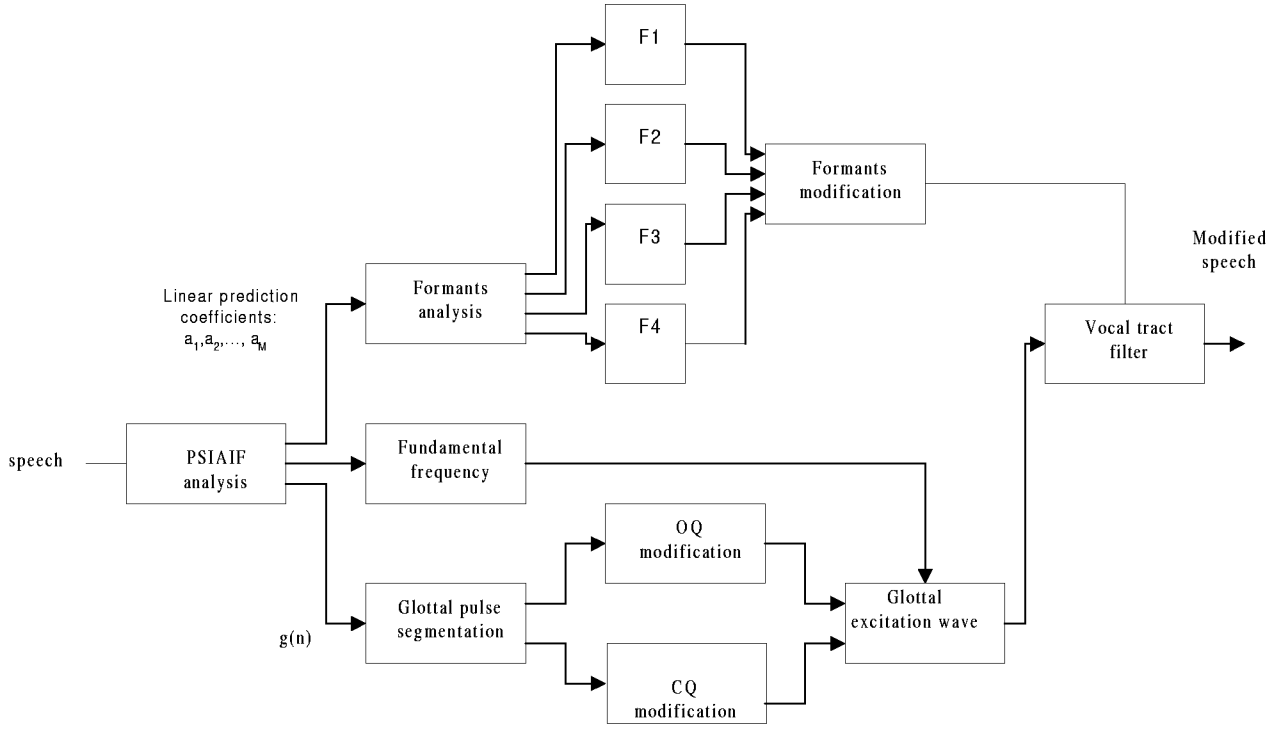

Figure 1: Block diagram of the analysis/synthesis system for modification of voice parameters.

A block diagram of the analysis/synthesis system is depicted in Fig. 1. The system was based on a linear model of speech production. The system enabled analysis and synthesis of voices while controlling 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 part was an iterative pitch synchronous inverse filtering algorithm, PSIAIF ([7]). Both the vocal tract transfer function and the glottal wave excitation were estimated by this method. Modifications included: 1. Shifts of single formant frequencies upward and downward at one and 2 tones; 2. Shifts of the whole spectral envelope upward and downward in a logarithmic scale; 3. Replacement of the spectral envelope by the speakers' average vocal tract model; 4. Fixation of the closing quotients of each speaker's glottal pulse at values ranging between 0.08-0.28 and the opening quotients at 0.2, 0.3, 0.4; 5. The fundamental frequency was changed at rates of 70%-130% the original value at 5% steps.

The statistical validity of the results was tested using logistic regression analysis.

Results

Formant changes caused the greatest decrease in identification percentage. The results for the modifications of single formants are summarized in Table 1. It is shown that the decrease in identification scores is related to the rate of shift of the formant's frequency. Lowering frequencies affected identification rate more than raising those frequencies. In addition, shifting higher formants (e.g. F3 and F4) affected identification percentages more than shifting lower formants. Shifting the whole spectral envelope by one tone caused a decrease in identification rate to 40%, significantly more than a parallel change in the glottal wave (65%) and in F0 (~58%). Exchanging the natural vocal tract filter by an artificial one, based on average speaker formants values, lowered the identification rate to 27%.

Similar results have been achieved by experiments in which voices were synthesized using hybrid voice production models. In these experiments one component of the speech production model was substituted by an artificial parametric model, or by that of other speakers. Identification rate dropped from 68% for natural voices, to only 14% when the speakers' voices consisted of their original glottal source in combination with other speakers vocal tract filter. Substitution of the glottal waveform by a parametric glottal model, while leaving other features intact (i.e. F0 and the vocal tract filter) resulted in identification rate of 44%.

All these findings suggest that on average, the contribution of the vocal tract features to human speaker identification is more important than that of the glottal source features, at least for vowels.

The identification scores of the listeners for each speaker separately are summarized in Fig. 2. It is shown that each speaker possesses a unique identification profile, i.e. the modifications affected differently the identification rate of different speakers. Modification of a certain formant in one speaker drastically reduced the identification rate for this speaker, whereas the same change had hardly any effect on other speakers. Other phenomena can be noticed from Fig. 2.

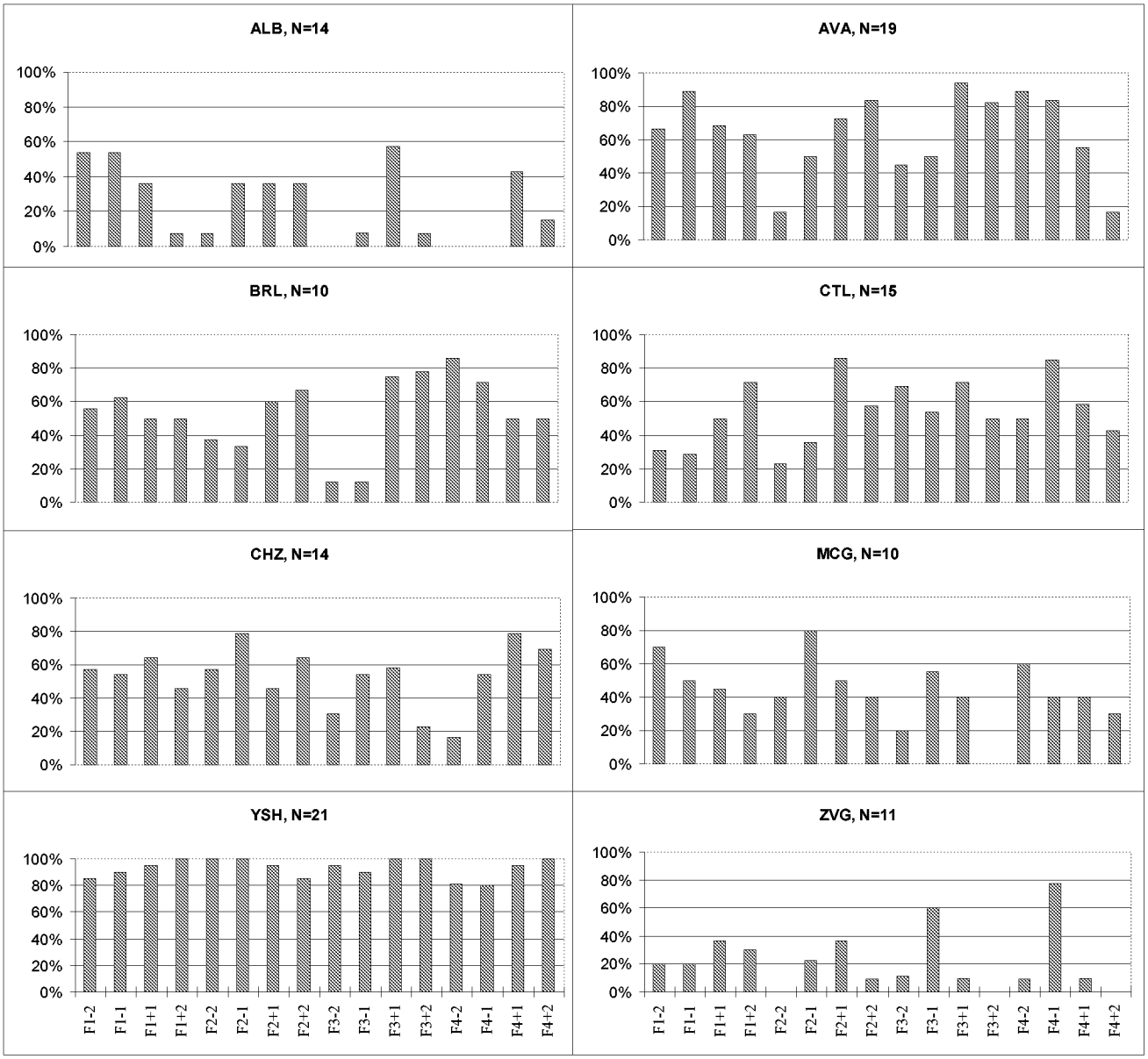

Figure 2: Identification scores of speakers for modifications of individual formants. The ordinate of each histogram describes the correct identification percentage. Each bar in the histogram represents one modification. The (+) and (-) signs represent raising and lowering of the formant frequency, respectively.

For some speakers, the modifications hardly affected the listeners' score. This result can be explained by the hypothesis that either the information for voice individuality is not coded in the vocal tract characteristics, or that additional features beside those of the vocal tract reveal speaker identity in spite of the modification. For other speakers, identification degraded by most of the modifications.

For such speakers, it is assumed that most of the information crucial to the vocal identity is coded within the vocal tract spectral envelope, such that even a moderate change of it impairs the possibility to recognize the speaker.

Changes in the glottal excitation waveform lowered identification rate to an average of 65%. Different rates of modification of the glottal pulse hardly affected this identification percentage. The modifications include changes in the closing and opening quotient of the glottal pulse.

F0 modification yielded on average an asymmetrical gradual lowering of speaker identification rate, with lowered F0 affecting identification rate more than increased F0. Again, as previously, examination of single speakers' results reveals different responses for similar stimuli.

The prototype model suggests that differences between various speakers can be explained by the level of deviation of each voice from the prototype pattern. For natural vowels identification test, it was found qualitatively, that the higher a speaker's voice deviated from the average, the larger was the identification rate. There was no direct and clear relation between the

Percent	# Correct	N	Formant modification
52%	99	192	F_1-2
58%	111	193	F_1-1
100%			F_1
60%	116	194	F_1+1
52%	101	195	F_1+2
47%	89	190	F_2-1
56%	108	194	F_2-2
100%			F_2
61%	121	197	F_2+1
53%	105	197	F_2+2
45%	88	195	F_3-1
53%	102	194	F_3-2
100%			F_3
64%	123	192	F_3+1
44%	84	190	F_3+2
39%	77	195	F_4-2
56%	108	193	F_4-1
100%			F_4
59%	102	174	F_4+1
44%	85	192	F_4+2

Table 1: A summary of the identification scores for modifications of individual formants, for all listeners and speakers.

identification rate and any of the features estimated (i.e., F_0, F_1, F_2, F_3, F_4, Opening Quotient, Closing Quotient and others). Nevertheless, a significant correlation of 0.7 was found between identification rate and a combination of two features, namely F_0 and F_3. These two features were selected due to the finding of a weak relation between the identification rate and each of these features in the scatter diagrams.

Conclusions

The present study investigates extensively some important features of human speaker recognition. A large group of listeners had to identify 20 speakers out of 29 possibilities, in an open-set psychoacoustic experiment. We studied modifications of individual and combined features of the vocal tract, the glottal waveform and F0. These experiments were limited to vowels, and therefore our conclusions should be considered with caution.

Numerous findings show that vocal tract characteristics are the most important for human speaker identification. Shifting the spectral envelope of the vocal tract reduced the identification rate significantly more than similar changes in the glottal waveform or of the fundamental frequency. These results were also found in hybrid experiments in which various components of the voice production system were exchanged by parametric models or by other features. Exchanging the vocal tract filter caused the greatest decrease in identification rate. The exact shape of the glottal waveform has a minor contribution for speaker identification. Although small changes in the opening and closing quotients reduced identification, different rates of modification yielded almost the same results.

A number of experimental predictions derived from the prototype hypothesis were found to be valid. The results support the prediction that each speaker has a different set of features characterizing his voice for identification purposes. It was found that the higher the departure of the voice from the average pattern, the easier it was identified. In addition, modifying an acoustic feature such as F0 by shifting it towards the average will reduce speaker identification significantly more than shifting it away from the average.

These findings suggest that the prototype model is an adequate perceptual model for human speaker recognition. It would be of value to investigate the applicability of this model to automatic speaker recognition.

References

[1] Kuwabara, H., and Takagi, T., (1991). "Acoustic parameters of voice individuality and voice quality control by analysis-synthesis method", *Speech Communication*, 10, 491-495.

[2] Itoh, K., (1992). "Perceptual analysis of speaker identity", in Speech science and technology, Saito, S. editor, IOS Press, 133-145.

[3] Rosenhouse J., Lavner, Y., and Gath I. (1995). "On the identification of familiar voices", *Proc. ICPhS 95*, Stockholm, 1, 190-194.

[4] Van Lancker, D., Kreiman, J., and Emmorey K., (1985). "Familiar voice recognition: patterns and parameters, Part I: Recognition of backward voices", *J. of Phonetics*, 13, 19-38.

[5] Papcun, G., Kreiman, J., and Davis, A. (1989). "Long-term memory for unfamiliar voices", *J. Acoust. Soc. Am.*, 85 (2), 913-925.

[6] Rosch, E.H., (1973). "On the internal structure of perceptual and semantic categories", In T.M. Moore (Ed.), Cognitive development and the acquisition of language, New York: Academic Press.

[7] Alku, P., (1992). "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering", *Speech Communication*, 11, 109-118.