The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels

Yizhar Lavner a, Isak Gath a,*, Judith Rosenhouse b

a Department of Biomedical Engineering, Technion, Israel Institute of Technology, 32000 Haifa, Israel
b Department of General Studies, Technion, Israel Institute of Technology, Haifa, Israel

Received 14 October 1998; received in revised form 9 April 1999; accepted 24 June 1999

Abstract

The aim of the present study was to examine the relative importance of various acoustic features as cues to familiar speaker identification for the vowel /a/. To this aim, a group of 20 speakers was recorded. The speakers’ voices were modified using an analysis–synthesis system, which enabled analysis and modification of the glottal waveform, of the fundamental frequency, and of the vocal tract formants. Thirty listeners, very familiar with the speakers’ voices, had to identify the speakers in an open-set, multiple-choice experiment. The results suggest that on average, the contribution of the vocal tract features to the identification process is more important than that of the glottal source features. The exact shape of the glottal waveform was found to be of minor importance. Examination of individual speakers reveals that changes of identical features affect the identification rate of various speakers differently. This finding suggests that for each speaker a different group of acoustic features serves as the cue to the vocal identity. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Familiar voices; Speaker identification; Acoustic features; Analysis/synthesis

1. 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, little is known about the link between the acoustic features of the speakers’ voice and higher processes of speaker identification by the listener.

A promising approach for finding this link is by analyzing and resynthesizing the speech signal, while controlling one or several different acoustic features. The performance of the listeners in identifying the speaker’s voice serves as a criterion to evaluate the importance of the modified parameters. In this type of experiment the number of subjects is of crucial importance: The number of speakers has to be large enough to simulate the natural situation of speaker recognition and to prevent artificial identification by guess or by elimination. The number of listeners is also important, for statistical validity and generality.

Several studies on speaker recognition by human listeners have been carried out in the past 50 years, but most of them treat small groups of both...
speakers and listeners. In the few cases where relatively large groups of subjects were investigated (Van Lancker et al., 1985a,b; Papcun et al., 1989; Ladefoged and Ladefoged, 1980; Schmidt-Nielsen and Stern, 1985), the speakers' voices were presented to the listeners either without changes, or with gross modifications such as reversing the speech signal, or transferring the speech signal through an LPC vocoder. These changes affect many acoustic features simultaneously, so that their value for understanding the acoustic bases for speaker individuality is limited.

Very few studies attempt to investigate directly the relationship between individual vocal features and speaker identification. Kuwabara and Takagi (1991) investigated the effects of modifying acoustic features on speaker identification scores. The results showed that shifting the formant frequencies affected speaker identification significantly, while formant bandwidth and pitch modification caused relatively smaller effects. In another study (Itoh, 1992), the vocal tract spectral envelope was found to have the largest effect on speaker identification. However, in both these studies only very small numbers of speakers and listeners were investigated, a fact that limits their generality.

The objectives of the present study were to investigate extensively, the contribution to speaker identification of each of the various acoustic features, on a large group of speakers and listeners very familiar with one another.

The main questions that motivated this study were:

1. What are the most important acoustic features that convey information on speaker individuality? Do the same acoustic features code all the speakers, or does each speaker possess a unique set of features?
2. What is the strategy of the listener in speaker identification?
3. Does the exact shape of the glottal excitation wave have substantial effect on the identification of speakers?

In addition to their theoretical importance, the answers to these questions can contribute to areas such as automatic voice recognition, voice transformation systems, and natural voice synthesis.

Section 2 deals with the methodological aspects of the study. The psychoacoustic experiments and the speech analysis/synthesis system are described in detail. Results of the effects of modification of the various speech features on the identification process are given in Section 3 and discussed in Section 4.

2. Methods

The study includes three main stages: (1) recording the speakers, (2) processing speech material and modifying various acoustic features, (3) examining and analyzing the listeners' responses to both natural and modified voices.

2.1. The psychoacoustic experiments

2.1.1. Speakers, listeners and speech material

Twenty male speakers, all native speakers of Hebrew (age range: 26–59) were recorded. The speech material contained two short sentences in Hebrew, short words, and isolated vowels. Speech was recorded using a condenser microphone (ACO 7040), and digitized with a 16 bit high quality sound card ('Multisound Monterey' of Turtle Beach) at 11 kHz sampling rate.

A total of 30 listeners, 22 females and 8 males aged 15–58, participated in the psychoacoustic experiments. Because of time limitations different numbers of listeners participated in each type of experiment (see Table 1). Both speakers and listeners were members of the same kibbutz and had been living in the same kibbutz for at least 5 years, a fact that ensured a high level of familiarity between them. 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. Only the results of the vowel /a/ will be reported in this paper. The isolated /a/ vowel was recorded in comfortable F0 and comfortable loudness, and its average duration was 0.5 s.

2.1.2. Experimental procedure

The psychoacoustic experiments were carried out in a quiet room. In each session, one listener
had to identify all 20 speakers from their voices in one type of experiment (see below). The speakers’ voices were played using a Multisound sound card. The sound card output was connected to a Harman–Kardon audio amplifier, with high quality loudspeakers (Boston Acoustics HD-5). The listener was instructed to respond to an interactive computer program. In each trial, 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. The purpose of these instructions was to inhibit the listener from using elimination strategies as part of the identification process. 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 by the listener in the first part. The logic behind this choice was that the chance of a listener identifying a modified voice while not identifying the speaker’s natural voice was negligible. The voices were played in a random order in both parts, so that each listener attended to a totally different sequence of voices.

Each experimental session lasted approximately 1–1.5 hours. The listeners were told that they were going to hear voices from a list of 29 speakers, and that some of the voices were modified or synthesized. The interactive screen of the experiment program consisted of graphical push-buttons for playing the voices, and for selecting the chosen speaker. The listener was free to play each stimulus up to 6 times. The reason for this limitation was that in preliminary experiments it was found that after 4–6 times of repeated hearings there was no improvement in the identification.

After the selection was made, the listener had to indicate his confidence in the selection by using a scale of 1–5, and to indicate the 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/she had the possibility to select the non-identified push-button. All the stimuli and the listener selections were automatically recorded and analyzed by the computer.

### 2.2. Types of experiments

Five types of experiments were conducted in the present study and in each of them a different acoustic feature was investigated (Table 1).

In the first experiment, the contribution of individual formants to speaker identification was examined by shifting the frequency of each of the first four formants separately upward or downward on a logarithmic scale. Four different modifications were made for each formant in each of the speakers, so that a total of $20 \times 4 \times 4 = 320$ stimuli were synthesized. Due to time limitations, only speakers that were identified in their natural

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Modification</th>
<th>Type of modification</th>
<th>No. of speakers</th>
<th>No. of listeners</th>
<th>Total stimuli</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Individual formants</td>
<td>Logarithmic scale</td>
<td>20</td>
<td>23</td>
<td>320</td>
</tr>
<tr>
<td>2</td>
<td>Combinations of formants:</td>
<td>Logarithmic scale</td>
<td>20</td>
<td>10</td>
<td>260</td>
</tr>
<tr>
<td></td>
<td>The whole spectral envelope</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average vocal tract</td>
<td>The speakers average value</td>
<td></td>
<td></td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>One natural formant</td>
<td>three formants fixed average</td>
<td></td>
<td></td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>One fixed formant</td>
<td>To the speakers average value</td>
<td></td>
<td></td>
<td>80</td>
</tr>
<tr>
<td>3</td>
<td>Glottal excitation wave</td>
<td>Closing phase</td>
<td>20</td>
<td>20</td>
<td>160</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Opening phase</td>
<td></td>
<td></td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>Fundamental frequency</td>
<td>70–130% of the original, in steps of 5%</td>
<td>20</td>
<td>17</td>
<td>240</td>
</tr>
<tr>
<td>5</td>
<td>Hybrid voices</td>
<td>Original vocal tract and synthetic glottal pulse or vice versa</td>
<td>20</td>
<td>4</td>
<td>60</td>
</tr>
</tbody>
</table>
voice were played to the listener in their modified form.

The second experiment consisted of voices in which various combinations of formant modifications were carried out. In this experiment the following stimuli were synthesized and played back:
1. vowels in which the whole spectral envelope of the vocal tract filter was shifted upward and downward on a logarithmic scale;
2. vowels in which the original vocal tract transfer function was substituted by a synthetic version, based on the formants’ average speaker population values;
3. voices in which one of the formants was fixed to that of the average speaker population value;
4. voices in which three out of the first four formants were fixed to their average population values, and the forth formant left intact.

In the third experiment the effect of changing the glottal waveshape was investigated. The original glottal wave of each speaker was modified using a linear mapping of segments within each pitch period. The modifications included changes to various new values of the opening or closing quotient, without changing the fundamental frequency or the vocal tract filter. The new resulting glottal excitation wave was passed through the original vocal tract filter. Eight different values for the closing quotient and three for the opening quotient for each speaker yielded 220 stimuli. It was found during an informal listening test that the changes in the closing quotient had perceptively greater effect than changes in the opening quotient, hence the different number of closing and opening values.

In another experiment, the effect of modifications of the fundamental frequency was investigated. The average fundamental frequency of each speaker was changed at the rates of 70–130% of its original value in steps of 5%. A total of 240 stimuli were synthesized for this test.

In the last experiment of this series, the relative importance of the 3 components of the voice production system (the vocal tract filter, the glottal excitation wave and the fundamental frequency) was investigated. In this experiment the speakers’ voices were played in 4 different versions:
1. the natural voice;
2. a synthetic voice generated by the original vocal tract model, but with artificial synthetic glottal excitation;
3. the same as in 2 but in addition, the fundamental period was fixed to the speakers’ average value;
4. a synthetic voice generated from the original glottal excitation wave (computed by an inverse filtering technique) passing through a vocal tract of another speaker who did not participate in the experiments.

2.3. The analysis/synthesis system

A block diagram of the analysis/synthesis system is depicted in Fig. 1. The system was based on a linear model of the speech production (Fant, 1960). The system enabled analysis and synthesis of voices while controlling the various acoustic features: the first four 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 (Alku, 1992). Both the vocal tract transfer function and the glottal excitation wave were estimated by this method. In contrast to the original technique, the covariance method was employed in the LPC analysis instead of the autocorrelation method. In addition, better results in reconstruction of the glottal wave from synthetic vowels were obtained (comparing the reconstructed signal with the original parametric model) by marking each pitch period as the time interval between two adjacent peaks in the glottal wave, instead of between two adjacent minima.

The performance of the analysis section of the system was tested with synthetic speech, consisting of the four vowels: /a/, /e/, /i/ and /o/. The synthetic glottal excitation wave in these vowels was simulated using a parametric model (Hedelin, 1984), and the vocal tract was represented by six formants. The synthetic vowels were analyzed using the system, and the estimated values were compared to the original (real) values. The results showed that formant frequencies were estimated quite accurately, with errors of less than 1%, and the syn-
thetic glottal pulse was fully reconstructed except for a small ripple in the closing phase (Fig. 2).

It is important to note that transferring the original speech signal through the analysis/synthesis system produced a reconstructed output speech that was 100% identical to the original natural speech.

2.3.1. Formant estimation and modification

The first stage of formant estimation is carried out by estimating the vocal tract transfer function, pitch synchronously. The next stage is to ascribe a single pole pair to each formant. The following empirical rules were applied (Childers and Lee, 1991):

1. Poles with frequencies lower than 250 Hz, or pure imaginary poles were not candidates for formant representation.
2. Poles with radii below 0.9 were discarded.

The relation between the formants and the poles are given by the following equations:

\[ f_i = \frac{F_s}{2\pi} \theta_i \]  

where \( f_i \) is the formant center frequency, \( F_s \) the sampling frequency and \( \theta_i \) is the pole angle, and

\[ B_i = -\frac{F_s}{\pi} \ln(r_i), \]
where $B_i$ is the formant bandwidth (Deller et al., 1993).

Modification of the vocal tract characteristics was carried out by shifting the center frequency of each formant separately. Shifting was upward or downward, on a logarithmic scale and was accomplished by changing the poles' angles, i.e., by rotating the pole around the unit circle with fixed radius. For example, if the desired change in frequency is of $\Delta f$ Hz, the corresponding change in the angle of the pole should be

$$\Delta \theta = \frac{\Delta f}{F_s} \cdot 2\pi,$$

where $F_s$ is the sampling frequency. The new angle and the new pole are therefore

$$\tilde{\theta}_i = \theta_i + \Delta \theta,$$

$$\tilde{p}_i = r_i \cdot e^{i\tilde{\theta}_i}.$$

The same change in the opposite direction is carried out on the complex conjugate pole.

For each modification a new all-pole system is created:

$$\tilde{V}(z) = \frac{1}{\prod_{i=1}^{N}(1 - \tilde{p}_i \cdot z^{-1})} = \frac{1}{1 - \sum_{i=1}^{N} \tilde{a}_i \cdot z^{-i}},$$

where $\tilde{a}_i (i = 1, 2, \ldots, N)$ are the new LP coefficients.

Although modification of one formant is carried out by changing the location of the relevant pole, undesirable changes in other formants might occur. The reason is that the whole spectrum is affected by the pole locations and the relative distances between them. Nevertheless, it could be verified that when the modifications were not too large, the undesirable changes were negligible (Fig. 3).

### 2.3.2. Glottal wave modification

The glottal waveform was reshaped by changing the duration of various segments within each pitch period. A block diagram of the process for creating synthetic vowels with glottal wave modification is given in Fig. 4. The glottal wave is estimated in the first stage (1), using the PSIAIF algorithm. Pitch synchronous estimation of the vocal tract is carried out simultaneously in this stage. The glottal wave is then smoothed, using a median filter (2), to remove any high frequency noise, which could exist due to incomplete removal of the formants. Subsequently (3), any trend in the...
The resulting waveforms are segmented manually into three phases (4). Utilizing this segmentation, the glottal wave is modified pitch synchronously (5), applying linear mapping to the various segments. The final modified wave is employed as the excitation to the original vocal tract filter (6), estimated earlier.

The resulting waveforms were hand-marked into 3 phases: closed phase, opening phase and closing phase (Fig. 5). The closed phase is the time period during which the glottis is completely or partially closed and the air flow through the glottis is minimal. In the glottal wave this period is between the end of one pulse (t₁) and the beginning of the consecutive pulse (t₂). The opening phase is the period in which the subglottal pressure causes gradual opening of the glottis to its maximal area (t₂ to t₃ in Fig. 5). The closing phase is the period of time in which the vocal folds are adducted from maximal to minimal glottal area (t₀ to t₁ in Fig. 5).

Two instances were actually labeled by the user: the opening point (t₂), which is the point between the closed phase and the opening phase, and the closing point (t₁), which is the demarcation point between the closing phase and the closed phase. The point of maximum of the glottal waveform, which is the point of separation between the opening and closing phases, was found automatically by a simple peak-picking algorithm. Both the glottal pulse and its derivative were used for marking t₁ and t₂. The opening points were defined as the first sample of the positive peak in the glottal wave derivative. The closing points were chosen as the first zero crossing of the return phase from the main excitation of the glottal derivative.

**2.3.2.1. Linear mapping of the glottal waveform.** Modification of the glottal waveform was performed utilizing linear mapping technique, by which segments of the glottal wave were expanded or constricted to various durations, within each pitch period. An example of the mapping procedure is demonstrated in Fig. 6, using a lattice, in which the spacing between two consecutive lines denotes the sampling interval. The figure demonstrates an expansion of the closing phase. The closed phase is therefore constricted by the same factor, to keep the original fundamental frequency, and the opening phase is mapped unchanged.

By using this procedure in discrete-time signals two problems may arise:
1. There could be samples in the modified wave that actually did not exist in the current sampling rate. For example the points 3’ and 5’ (Fig. 6).

![Fig. 5. Segmentation of one period of the glottal excitation signal.](image)

![Fig. 6. An example of glottal pulse signal mapping using lattice. The original glottal wave is depicted under the lattice, whereas the modified wave can be seen on the left side. The main diagonal (45°) in the lattice denotes mapping without change, while the xx and yy lines serves for expansion (more than 45°) and constriction (less than 45°), respectively.](image)
2. The mapping may cause a change in the sampling rate. The expansion causes a lower sampling rate, and the constriction a higher sampling rate. In both cases the change is usually by a rational factor.

Both problems can be solved by using interpolation with a rational factor. However, this procedure is accompanied by using an anti-aliasing filter, which might contain a large number of coefficients. Since the segments are very short, a large error could result, due to the use of zero initial conditions. To avoid this error, a finer lattice is created by resampling the whole signal with a much higher sampling rate (110 250 Hz). For the new desired time samples, the signal values are determined by the finer lattice samples that are closest to the target samples.

2.3.3. Modification of the fundamental frequency

Various studies (Monsen and Engebretson, 1977; Sundberg and Gauflin, 1979; Karlsson, 1985) have shown that the open quotient of the glottal pulse wave remains relatively constant during natural changes of the fundamental frequency. This finding and the assumptions of linearity and uncoupling between the glottal source and vocal tract filter, justify the hypothesis that fundamental frequency modification can be obtained by linear mapping of the glottal wave, i.e., by its constriction for raising $F_0$ and by expansion for lowering $F_0$.

The modified glottal wave was transferred through the original vocal tract transfer function and the lip radiation filter. Modifications were made by changing the sampling rate by a rational factor (Proakis and Manolakis, 1992). The average fundamental frequency of each speaker was changed within the range 70–130% of the natural average $F_0$, since such changes occur in natural conversation. For the glottal wave, the modifications spanned the range of closing and opening duration found in the analysis stage for the speaker population. Formants modifications were on a logarithmic scale, due to the nature of the hearing system (at least for $F_2$, $F_3$, and $F_4$).

2.3.4. Motivation for the specific modifications

The modifications in the various parameters were based on the following considerations:

1. The modification should be within the physiological range. For example, the fundamental frequency was changed within the range 70–130% of the natural average $F_0$ since such changes occur in natural conversation. For the glottal wave, the modifications spanned the range of closing and opening found in the analysis stage for the speaker population. Formants modifications were on a logarithmic scale, due to the nature of the hearing system (at least for $F_2$, $F_3$, and $F_4$).

2. The modified speech signal had to sound as natural as possible. All speech signals were tested for naturalness in an informal test, and examples with a lack of this feature (where the sound seemed too artificial) were discarded from the psychoacoustic experiments. After the experi-

![Fig. 7. Lowering the fundamental frequency of speaker IDG to 70% of its original: (a) glottal excitation signal before (top) and after (bottom) the modification; (b) the speech signal before (top) and after (bottom) the modification.](image-url)
ments, average naturalness was calculated for each voice based on the listeners’ grades. Voices in which this parameter was lower than a threshold of 2.5 were not used in the analysis.

3. The steps within each range of change were selected so that there would be a perceptible difference between two consecutive modifications.

3. Results

In all the experiments, the tests were carried out in two stages: In the first stage the listeners heard the natural voices of the speakers without any modification. In the second stage, the listeners heard only voices which were identified correctly in the first stage. The overall identification rate of the natural stimuli was 49.6% (852 out of 1719). A stimulus–response confusion matrix for the natural voices is depicted in Table 2. The stimuli are listed in the first column, and the listeners’ responses in the first row. It can be shown that there are great variations in the identification scores between different speakers, both in correct identifications and in the false positives. For example there are speakers with high identification percentage (3, 15 and 16), while others have very low percentages (12 and 19).

In the following figures and tables (except for Table 2) the identification percentages are calculated according to the speakers who were identified by their natural voices in the first stage of the experiments.

3.1. The effect of shifting individual formants

3.1.1. Average for all speakers

In the following psychoacoustic experiments, 23 listeners had to identify 20 speakers out of 29 possibilities. The frequency of the first four formants was shifted separately by one and two tones, upward and downward, hence, for each speaker 16 synthetic voices have been generated.

The identification scores of the listeners for all the speakers for each modification are summarized in Fig. 8. It can be seen that the decrease in identification scores is related to the shift of the formant frequency. Shifting a formant frequency by two tones, resulted in a greater and statistically significant decrease in identification rate than shifting by one tone, logistic regression analysis, \((p > \chi^2 < 0.00025)\). Lowering the formant center frequencies affected identification rate more than raising those frequencies.

Although the profiles for the identification scores for the four first formants are quite similar to each other, it can be seen that shifting higher formants (e.g., \(F_3\) and \(F_4\)) affected identification percentages to a larger extent than shifting lower formants (\(F_1\) and \(F_2\)). For example shifting \(F_3\) and \(F_4\) upward in two tones, reduced the identification score significantly more than the same modification in \(F_2\) (logistic regression analysis \((p > \chi^2 < 0.05)\). The same modification affected \(F_3\) and \(F_4\) more than \(F_1\), but not significantly \((p > \chi^2 < 0.1)\). The same result was found for the upward direction: The modification affected identification percentage in \(F_4\) more than in \(F_1\) or \(F_2\).

3.1.2. Results for individual speakers

The identification scores of individual speakers were analyzed. Statistically, significantly different scores were obtained for different modifications (Cochran \(Q\)-test, \(z = 0.001\)). For example, 16 listeners out of 18 identified the speaker AVA (Fig. 9) when \(F_4\) was lowered by two tones, whereas only three identified him when the same modification was carried out on \(F_2\). Similar results were found for other speakers. These results imply that for each speaker, different formants contribute differently to speaker individuality.

3.1.3. Differences between speakers

One of the goals of these experiments was to investigate whether the information regarding the speaker identity is being coded for all speakers in a similar way, i.e., that the same set of parameters or acoustic features serve for speaker individuality for all the speakers, or alternatively, that each speaker possesses his own unique code – i.e., the set of features is different for every speaker.

It was hypothesized that for a constant set of features for all speakers, each modification would cause similar identification results in the various speakers. But if the clues for speaker identification are variable, it is expected that the same
Table 2
Stimulus–response confusion matrix for the natural voices. S – stimuli, R – responses, N – non-identified, O – others (other names which were used as distractors), false – false positive

| S  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | N | O | Total |
|----|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|---|---|-------|
| 1  | 27| 2 | 2 |   | 6 |   | 14|   | 11|   | 9  | 15 | 86 |
| 2  | 5 | 47| 1 | 2 | 1 |   | 4 | 3 | 1 | 17 | 86 |
| 3  | 75|   |   |   | 1 | 1 |   | 1 |   | 8  | 86 |
| 4  | 37| 3 | 4 | 7 | 1 | 1 | 8 | 1 | 3 | 20 | 86 |
| 5  | 63| 1 |   |   | 3 | 3 | 3 | 4 | 9 | 86 |
| 6  | 75| 1 | 35| 2 | 2 | 1 | 6 | 23| 10| 86 |
| 7  | 33| 4 | 30| 1 | 6 | 2 | 5 | 1 | 21| 11| 86 |
| 8  | 1 | 5 | 1 | 1 |   | 43| 1 | 14|   | 4  | 12| 86 |
| 9  | 2 | 3 | 2 | 1 | 2 | 1 | 1 | 47| 6 | 9  | 12| 86 |
| 10 | 3 | 6 |   | 2 | 1 | 24| 1 | 1 | 5 | 4  | 18| 15| 86 |
| 11 | 1 | 1 | 2 | 1 | 1 | 47| 6 | 9 | 12| 6 | 86 |
| 12 | 1 | 2 |   | 1 | 1 | 12|   | 21|   | 34| 12| 86 |
| 13 | 3 | 1 | 8 | 2 |   | 51| 1 | 4 | 1 | 10| 5 | 86 |
| 14 | 3 | 1 | 6 | 1 | 3 |   | 38| 1 | 6 | 26| 85 |
| 15 | 1 | 2 |   |   | 2 | 80|   |   | 1 | 86 |
| 16 | 1 |   |   |   |   |   | 72| 1 | 1 | 12| 86 |
| 17 | 9 | 1 |   | 3 | 4 |   | 44| 1 | 10| 23| 86 |
| 18 | 6 | 1 | 1 | 4 | 8 |   | 1 | 1 | 59| 2 | 9 | 86 |
| 19 | 4 | 1 | 1 | 5 | 3 | 8 | 7 | 10| 4 | 32| 86 |
| 20 | 5 | 1 | 2 | 2 |   | 2 |   | 1 | 25| 27| 86 | 867 |

Total hits

| False | 23 | 25 | 16 | 21 | 6 | 27 | 11 | 4 | 13 | 6 | 14 | 15 | 13 | 49 | 0 | 19 | 76 | 0 | 10 | 16 | 219 | 284 | 867 |
modification would affect identification rate differently in different speakers.

To test this hypothesis, identification results of all the listeners for each speaker were collected and summarized (cf. Fig. 9). Each histogram in this figure summarizes the identification scores of all the modifications for one speaker. It can be easily seen, that the modifications affected identification rate of various speakers differently (logistic regression analysis $p > \chi^2 < 0.05$). Modification of a certain formant in one speaker reduced the identification rate for this speaker drastically, whereas the same change had hardly any effect in other speakers. For example, lowering the third formant in the voice of BRL, for both one and two tone shifts (Fig. 9), reduced identification rate to 13% (average was 53% for 10 listeners who identified this speaker by his natural voice), whereas the same modification had only a moderate effect on the identification percentage of CTL (69% for two tones, average = 54%, in 14 listeners). Raising $F_3$ center frequency of MCG by one tone totally degraded the identification (0%, average = 43%, 10 listeners) while hardly affecting the identification of AVA (more than 80%).

Other phenomena can be found in Fig. 9. For some speakers, modifications hardly affected the listeners’ score (for example, the histogram of YSH). 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 convey information on the speaker’s identity in spite of the modification. For other speakers, identification degraded in most of the modifications (see for example the histogram of ZVG). For such speakers, it is assumed that most of the information crucial for tracing the vocal identity is coded within the vocal tract spectral envelope, such that even a moderate change impairs the possibility of recognizing the speaker.

It can be concluded from this experiment that there is no closed fixed set of variables or features, by which the speaker is identified and that each speaker possesses his own unique set of features, that characterizes him.

3.2. Effects of shifting the spectral envelope

The effects on speaker identification of shifting all the formants simultaneously were examined in this psychoacoustic experiment. The vocal tract was modified by shifting the whole spectral envelope upward or downward (i.e., to higher or lower frequencies) by a half tone (5.9%) or by one tone (12.25%). Shifting by two tones or more, seriously degraded the naturalness of the synthetic voice, and was therefore not presented to the listeners.

Ten listeners participated in this psychoacoustic test and the results are presented in Table 3. As in the case of modification of individual formants, a slight but statistically significant asymmetry between lowering and raising the formant frequencies can be noticed (logistic regression analysis, $p > \chi^2 < 0.05$). As expected, shifting the whole spectral envelope impaired the possibility of identifying the speaker more than modification of individual formants. A downward shift of the spectral envelope by one tone caused a decrease in the identification rate to 39%. The same shift yielded 58% identification rate for modification of $F_1$, and 56%, 53% and 56% for $F_2$, $F_3$ and $F_4$, respectively. Similar results were found for modification in the upward direction: 41% identification rate.

---

1 Results of similar psychoacoustic tests for other features reveal the same conclusion, see below.
rate for the spectral envelope modification, and 60%, 61%, 64% and 59% for modification of each of the first four formants.

### Table 3
Identification results for modification of the whole spectral envelope

<table>
<thead>
<tr>
<th>Modification</th>
<th>N</th>
<th>Correct</th>
<th>Percent (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 tone downward</td>
<td>95</td>
<td>37</td>
<td>39</td>
</tr>
<tr>
<td>1/2 tone downward</td>
<td>96</td>
<td>49</td>
<td>51</td>
</tr>
<tr>
<td>1/2 tone upward</td>
<td>96</td>
<td>55</td>
<td>57</td>
</tr>
<tr>
<td>1 tone upward</td>
<td>96</td>
<td>39</td>
<td>41</td>
</tr>
</tbody>
</table>

Fig. 9. Identification scores of speakers for modifications of individual formants. The ordinate of each histogram describes the correct identification percentage. The (+) and (−) signs represent raising and lowering of the formant frequency, respectively.

#### 3.3. Average vocal tract

In this experiment the speakers’ voices were synthesized by replacing their original vocal tract transfer function with a model derived from the average formants of all the speakers. Synthetic vowels were constructed by exciting this common vocal tract with the natural derivative of the glottal pulse waveform of each speaker.

The average identification percentage for all the listeners was 27%. In comparison with the former
experiments and also with experiments examining other features, it is probable that most of the information pertinent to speaker identity is stored in the vocal tract features.

Large variabilities in the identification scores were found for different speakers. Some of the speakers were not identified at all (for example, AVA or ALB, Fig. 10), while others were identified by most of the listeners (CHZ and YSH, Fig. 10).

3.4. The effects of glottal waveform modifications on speaker identification

Modifications of the glottal waveform were carried out using the glottal source – vocal tract analysis/synthesis system, and included fixation of the closing quotients of each speaker’s glottal pulse at values ranging between 0.08 and 0.28 in steps of 0.04 and the opening quotients at 0.2, 0.3 and 0.4. Twenty listeners participated in these psychoacoustic experiments. Results of psychoacoustic experiments for modifications of the closing quotient are given in Fig. 11. From Fig. 11 it can be seen that changes in the closing quotient lowered the identification rate to an average of 65%. Different rates of modification hardly changed this identification percentage. This finding was found to be true both for the absolute values of the modifications, as well as for deviations from the original values.

Results for modifications of the opening quotient are shown in Tables 4 and 5. The best identification score was achieved for an OQ of 0.3.

3.5. Modifications of the fundamental frequency

Results of the psychoacoustic experiments in which the average fundamental frequency was modified are given in Figs. 12 and 13. Fig. 12

![Fig. 10. Speaker identification for synthetic voices with fixed vocal tract. The ordinate describes the identification percentage. Only speakers identified by at least six listeners are included.](image)

![Fig. 11. Diagram showing speaker identification as a function of the closing quotient of the glottal pulse waveform. The ordinate describes the identification percentage.](image)
shows average identification rates for all speakers for a given change in $F_0$. $F_0$ modification yielded an asymmetrical gradual lowering of speaker identification rate, with lowered $F_0$ affecting identification rate more than increased $F_0$. As previously, examining the results of single speakers reveals different responses for similar stimuli. The results for three individual speakers are depicted in Fig. 13.

### 3.6. Hybrid voice experiments

The results of the psychoacoustic experiments, designed to evaluate the contribution of the vocal tract characteristics versus the contribution of the glottal characteristics, are depicted in Fig. 14. The histogram describes the average identification percentage of four listeners for four types of stimuli:

1. natural voices;
2. voices in which the original glottal waveform was replaced by a parametric model, while the fundamental frequency and the vocal tract were kept intact;
3. the same configuration as in (2), but in addition, the fundamental frequency was fixed to that of the average population;
4. voices in which the original glottal source excited the vocal tract transfer function of another speaker, who did not participate in the listening test.

Table 4
Results of psychoacoustic experiments with modifications of the opening quotient (for details see text)

<table>
<thead>
<tr>
<th>Modification</th>
<th>N</th>
<th>Correct</th>
<th>Percent (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0</td>
<td>80</td>
<td>37</td>
<td>46</td>
</tr>
<tr>
<td>0–0.1</td>
<td>143</td>
<td>88</td>
<td>62</td>
</tr>
<tr>
<td>0.1–0.2</td>
<td>157</td>
<td>102</td>
<td>65</td>
</tr>
<tr>
<td>&gt;0.2</td>
<td>121</td>
<td>82</td>
<td>68</td>
</tr>
</tbody>
</table>

Table 5
The effect of relative modification of the opening quotient on speaker identification

<table>
<thead>
<tr>
<th>Modification</th>
<th>N</th>
<th>Correct</th>
<th>Percent (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>161</td>
<td>94</td>
<td>58</td>
</tr>
<tr>
<td>0.3</td>
<td>161</td>
<td>118</td>
<td>73</td>
</tr>
<tr>
<td>0.4</td>
<td>161</td>
<td>97</td>
<td>60</td>
</tr>
</tbody>
</table>

Fig. 12. Diagram showing identification rates for modifications of the fundamental frequency. The ordinate describes the identification percentage.

Fig. 13. Identification of speakers AVA, ZGR and TSG as a function of modification of $F_0$. The ordinate describes the identification percentage.
The four listeners selected for this experiment exhibited high scores (above 60%, i.e., at least 12 out of the 20 speakers) in the identification of speakers in previous experiments. Replacing the original vocal tract with that of another speaker caused the largest decrease in the identification rate, from 68% for natural voices, to 14%. Substitution of the glottal waveform by a parametric glottal model, while keeping other features intact (i.e., $F_0$ and the vocal tract filter) resulted in an identification rate of 44%. Additional modification of $F_0$ (to the average value of the speakers) yielded 24% correct identifications. In previous experiments (see Section 3.2), it was found that shifting the whole spectral envelope by one tone caused a decrease in identification rate to 40%. This decrease is significantly larger than that caused by a parallel change in the glottal wave (65%) or in $F_0$ (~58%).

4. Discussion and conclusions

The main questions that motivated the present research were the following:

1. What are the most important acoustic features for speaker recognition by a human listener?
2. Can each speaker be identified by a single unique feature, i.e., is there a single feature that conveys most of the information regarding the speaker’s individuality, or do several features code each speaker’s individuality?
3. Is there any fixed and closed set of features that is utilized by listeners to recognize all the speakers, or is the speaker represented by a unique set of features that is different from that of other speakers?

The psychoacoustic experiments with individual formant modifications (Section 3.1) have shown that the identification percentage is related to the rate of shift of the formant’s frequency. In addition, lowering the formant frequencies affected identification rate significantly more than raising those frequencies. The results also show that shifting the formants’ frequencies yielded greater degradation in the identification rate for the higher formants ($F_3$ and $F_4$) than for the lower formants ($F_1$ and $F_2$). This finding is somewhat surprising, since it is expected that the lower formants would be of a greater perceptual importance, because of their higher energy content. On the other hand, the higher formants are less restricted by phonemic constraints, so it seems reasonable to assume that their interspeaker variability is higher.

This latter finding is supported by other results from the set of experiments in which single formants in each voice were fixed to the average population values, e.g., the identification percentage for $F_2$ modification decreased less than those of other formants.

The same asymmetry between lowering and raising formant frequencies was also found for changing the whole spectral envelope section (Section 3.2). As expected, this change caused a greater decrease in the identification rate than changes of individual formants. An average identification rate of 40% was obtained for shifting the spectral envelope by one tone, whereas the same modification of individual formants resulted in significantly higher identification rates. Thus, it seems reasonable to assume that the information content in the spectral envelope regarding speaker individuality is larger than that of the individual formants separately. Moreover, the decrease in speaker identification caused by shifting the spectral envelope by one tone was more pronounced than the decrease found in experiments with other acoustic features (like the glottal waveform or the fundamental frequency) utilizing changes of similar magnitudes.

The lowest rate of speaker identification was obtained when the original vocal tract model of each speaker was replaced by a synthetic average.

![Diagram showing the identification results for hybrid voices.](image-url)
vocal tract, derived from the formants’ averages of all the speakers (Section 3.3). In this case the identification percentage dropped to 27%.

For further examination of the relative contribution of various components of the voice production system to speaker identification, hybrid voices were synthesized. The greatest decrease in identification rate (14%) was produced by substituting the original vocal tract model of each speaker with a vocal tract transfer function of another speaker, whose voice was not presented in the experiments. Replacing the glottal excitation waveform of each speaker with a parametric model, while keeping the fundamental frequency intact, yielded a significantly higher rate (44%).

All these findings suggest that on average, the vocal tract characteristics are the predominant factor contributing to speaker individuality. It is important to emphasize that this conclusion applies to the experimental setup of the present study, namely, listening to isolated vowels, and that additional speech parameters (e.g., $F_0$ variations) may be more important when listening to regular sentences.

The attempt to find the key features for speaker identification is not a new one. Our results agree with some of the studies, while contradict others. For example, the supremacy of the vocal tract characteristics for speaker recognition was suggested by Miller (1964). However, part of that study was based on matching technique rather than on identification. In addition, for generation of the hybrid voices a vocal tract of only one speaker was utilized with artificial glottal waveforms. Thus, the conclusion of this experiment is related to general sound perception rather than to the issue of speaker identification. Matsumoto et al. (1973) claimed that the fundamental frequency is the major contributor to the perception of personal quality. However, in this study same–different task was employed, and therefore, again, the study deals with the contributions of the components of the speech production system to differences in sound quality, rather than to speaker identification.

Contrary to the findings of the present study, Carrell (1984) concluded that the glottal wave alone was sufficient for the identification of speakers. However, it can be argued that the PILT (‘pseudoinfinite length tube’, Sondhi, 1975) method used in that study does not remove all the information content of the vocal tract, and thus, the stimuli did not comprise only the glottal wave.

The glottal excitation signal was found to be of great importance in different areas of speech research, such as improving the naturalness of synthetic speech (Rosenberg, 1971; Hedelin, 1984, 1986), discriminating between speech styles (Cummings and Clements, 1995) or phonation types (Childers and Lee, 1991), and synthesizing natural female voices (Price, 1989, Klatt and Klatt, 1990). Several reasons support the notion that the exact waveform of the glottal excitation signal has also a significant contribution to speaker identification. On the one hand the glottal excitation signal is relatively independent of phonetics, because of the weak coupling between the voice source and the vocal tract. On the other hand, the glottal waveform depends on the dimensions and shape of the speakers’ larynx, and it is reasonable to assume that it could characterize the speaker uniquely. However, this view is not supported by the results of the present study.

The contribution of the glottal excitation signal to speaker recognition was tested by modifying the waveshape of the signal. On average, modifying the glottal excitation signal reduced the identification rate to 65%, irrespective of whether the changes concerned the opening or the closing phases. On the other hand, different rates of change yielded very close results. These findings suggest that the glottal excitation wave does have a contribution to speaker recognition, but that the exact shape of the excitation signal is of minor importance, in particular if the information of the natural fundamental frequency is preserved.

It is interesting to note that the results of the present study concerning the glottal wave resemble the findings obtained by investigating the effect of the glottal waveform on the quality of synthetic speech (Linden and Skoglund, 1994). It was found that as long as the fundamental frequency was preserved, various parameters of the glottal pulse could be changed considerably without a significant perceptible degradation of quality.
An investigation of the effects of modifications of the fundamental frequency on speaker identification was carried out by expanding and condensing the train of glottal waves, keeping the relation between the various glottal wave segments within each pitch period unchanged. The results showed that the identification rate was related to the rate of change of $F_0$. As was demonstrated for formant modification, here too, lowering $F_0$ reduced identification rate more than raising $F_0$. This result can be explained by the natural tendency of speakers to use mainly their lower part of the vocal registers. Thus, raised frequencies are still perceived as possible representatives for a given speaker, whereas lowered frequencies are out of the possible natural range, and therefore, it is less probable that these voices are perceived as being related to that speaker.

The evaluation of identification results for individual speakers reveals different and more complicated perspectives of the process of speaker recognition. For both formants and fundamental frequency modifications, significant differences between the results of different speakers could be verified. The same modifications induced totally different scores for different speakers. For example, shifting the frequency of a given formant by a certain rate could totally obscure the identity of some speakers, while leaving the identification of others almost intact. A possible explanation for this phenomenon is that modifying a feature which uniquely identifies the speaker severely reduces the identification percentage, while modification of features of less importance does not affect the identification process.

If the identities of all speakers were cued by the same set of features, a similar decrease in identification rates of different speakers for a given modification of their voices, were to be expected. This possibility is refuted by the results. The finding that each speaker exhibited a different and somewhat unique sensitivity for modification of various features suggests that each speaker possesses different features as a key to his identity.

Differences between identification scores of individual speakers were also documented by Van Lancker et al. (1985a,b), who used gross and non-specific voice processing procedures, such as forward–backward and rate-altered presentation of stimuli.

In summary, though it can be shown that on average, among the 3 major speech components, most of the information for speaker identification is contained in the vocal tract characteristics (at least for vowels), large individual differences exist between speakers, suggesting that each speaker has a different personal combination of acoustic features that cues his identity.

References


