ATTRIBUTION OF AGE FROM PERCEIVED SPEECH

Leo W.A. van Herpt and Ton Hoebe

1. INTRODUCTION

When we hear someone's voice we form an impression of his appearance. How we trust these judgments shows in our confusion when we find that 'the man's voice' belongs to a woman or when a calm, deep, sonorous sound is produced by a five feet tall man with a baby face. But in general these judgements are reliable and rather accurate. Much is still unknown about the cues which are used in this attribution process, but since they have perceptual importance speech must have certain acoustic characteristics which are responsible for eliciting those perceptual processes. Over forty years ago already Sanford (1942) wondered whether this connection is of a sort thatcan only be sensed by judges or whether it could ever be analyzed and stated in a qualitative manner. To find out which acoustic characteristics are related systematically to the subjective impressions more research on the identification of relevant parameters on both levels must be done.
On the perceptual level the long-term voice features that result from anatomical speaker-characterizing differences are the most reliable.
So primarily the biologically determined voice-indices of age, sex and physique will be used in our investigations to define criteria of speaker-characteristics that can be used for automatic speaker recognition and for the evaluation of coded speech.
We started with a series of experiments concerning factors that influence the perception of aging in the human voice. The first study, the one that is reported on in the present paper, is aimed at methodological aspects for subsequent research.
The experiment consisted of an evaluative rating by twenty-nine listeners based on two minutes oral reading by fifty Dutch speakers aged from eighteen to eighty-seven. The aim was to explore 1) factors that influence the accuracy of age estimation on the basis of vocal cues alone and 2) the potency of the ten bipolar semantic scales and twelve speech therapeutic terms to differentiate between younger and older speakers.
The results will show that listeners are able to estimate the speakers' age with reasonable accuracy although several interfering and interacting variables have been identified, but first we'll give a short survey of some aspects considered relevant in literature.

2. LITERATURE

It is not a new assertion that vocal functioning reflects the on-going physical changes with age and that these changing vocal parameters can be perceptually observed.
Publications in the thirties already showed that listeners are able to estimate the age of speakers solely by listening. Pear (1931) investigated with

assistence of the BBC the problem of judging the personality of nine speakers of different ages, sex and interests by their radio voice. On the whole the (4000) listeners estimated the ages with fair success. Stern (1932) tried to relate changes of the aging voice as perceived by listeners to the physiological changes of the larynx caused by age. In a study of Herzog (1933) too, age was judged more accurately than one would expect by chance. Allport and Cantril (1934) conducted a series of experiments to determine to what extent the natural voice is a valid indicator of various features of personality. Speakers' ages were judged in two experiments. In the first the voices were matched by 50 listeners to the ages given; in the second the ages of three speakers were merely estimated by the 190 judges. Both experiments yielded positive results. In the last one there was a tendency to center the ages around the median, corroborating Pear's discovery of a central tendency in judgments of age from voice.
The general observation that age estimation can be made from voice stimuli has an explanation among other things in physiological changes caused by alterations of the anatomic elements and by changes in the endocrine functions, especially of sex hormone production.

The vocal mutation of puberty (12-15 years) occurring in the course of endocrinological development is closely bound up with the extraordinary growth rate of the larynx. According to Damstee (1973) the vocal cords of boys can increase in lenght by 1.5 times.
Beginning shortly after the age of 20 the hyaline laryngeal cartilages become progressively ossified (Luchsinger and Arnold, 1965).
According to Kressner and to Luchsinger (cited by Endres et al., 1971, p. 1848)'the so-called "Waldeyer's tonsilar ring" - i.e. the palatine tonsils, the gray columns and the lymph glands - shows a strong involution after the 35th to 40th year.' This reduction in size results in an enlargement of the supraglottic tract.
'In the female the aging of the voice is closely correlated with the menopause around age 50.' (Luchsinger and Arnold, 1965, p. 135). Vocal involution from male climacterium occurs at a later age. The elasticity in the tissues decreases and the mucosal epithelium tends to atrophy.
Elderly people show atrophy of the internal vocalis muscle by which their glottis 'often has a bowed appearance' (Luchsinger and Arnold, 1965, p. 136), 'this means that, to achieve phonation, greater effort has to be exerted to bring the vocal cords together, and a rather harsh voice is often the result.' (Laver and Trudgill, 1979, p. 10)
In old age fatty tissue - so-called lipoid substances - arise in the upper larynx and around the muscular fibrils of the vocal cords (Inhofer, 1913; Ferreri, 1959).
In extreme old age, changes in voice derive from a complex of endocrinal processes. 'The mucal fluid supply often becomes disturbed (...), tissues become less elastic and cartilages become calcified and ossified' (Laver and Trudgill, 1979, p. 10).

All these factors and more can contribute to the quality of phonation and to changes in quality of the aging voice. (Other explanations, e.g. those considering the characteristic audition changes or socio-emotional factors, as

proposed by Henton (1982) are not taken into account in this report.) But these phenomena did not lead to a great amount of perception studies. On the contrary, almost all research on the relation between age and speech has been concerned with the acoustically measurable aspects of phonation. In particular fundamental frequency (Fo) got much attention.
With increasing age both males and females show - at least until adulthood - a trend of decrease in Fo. From about the age of sixty the average Fo of men tends to shift upwards, possibly due to feminization. Virilization in women after 50, due to an increase in the testosterone-estrogen ratio after the menopause, would led us to expect a downward shift in Fo, but here contradictory results are obtained. Charlip (1968) does not find systematic alterations associated with age, whereas McGlone and Hollien (1963) find a small increase in Fo, as is also reported by Stoicheff (1981) for non-smoking women of old age.
For more conclusive statements more comparative data must be gathered with control of variables such as subjects in pre- or postmenopause and of smokers and non-smokers.
It is not a simple thing to ascertain without longitudinal studies how the female voice develops naturally during the menopause. Damstee (1973, p. 93) states: there are so many grandmothers with low and creaky voices that one may well wonder if there are still mature women left who have never been treated with androgenic preparations. (The influence of testosteron with which a great many women are treated is very individual.)

Other supposed clues of the aging voice which have been analysed acoustically are the stability (or variability) of pitch and intensity, and the rate of speaking and phonation. For an extended review see Helfrich (1979).

Notwithstanding the overemphasis that is placed in research on a few objectively measurable acoustical aspects of the voice, undefined perceptual features of voice quality and voice dynamics are supposed to be the most important clues of age. Nevertheless very little research is done about these perceptual features. Except descriptions of the 'breaking voice' of puberty, work on this subject is virtually limited to the description of older people's voices.

According to Laver and Trudgill (1979, p. 10) 'vocal indications of puberty often include whispery voice' and Luchsinger and Arnold (1965, p. 132) add that 'In addition to the lowering of the average speaking pitch, the voice is frequently husky during mutation, or it may sound weak'.
The voice of old age is generally described in negative terms: monotonous, shrill, thin, sharp, hollow, weak, hoarse, rough, breathy, tremulous trembling. It is not clear to which extent these characteristic attributes are based on stereotypes. Stereotypes, defined as a collection of beliefs about a group, may or may not be a valid description of reality, but the labels are hardly ever acoustically defined. And if efforts have been made to define them the results are not unambiguous. According to McGlone and Hollien (1963, p. 164) Bach, Lederer and Dinolt should have described the 'senile' voice as 'monotonous, flat, and occasionally shrill'. (This citation could

not be found in the original publication.) McGlone and Hollien (1963) tried to verify this description but come to the conclusion that 'pitch variability of speech apparently changes little with advancing age' (p. 169).
In an experiment of Ptacek and Sander (1966) listeners indicated 'that the older speakers tended to use lower pitches and less pitch variability' (p. 276).
Mysak (1959) on the contrary finds that the measure of pitch variability 'reflected a general trend toward greater variability as a function of age' (p. 52).
Most authors agree that the voice of old age has a smaller vocal range. But also there is a shift of the vocal range toward the lower or higher frequencies. Lowering of pitch range can happen through vocal mutation after the menopause and/or because the vocal cords become thicker, e.g. through the occurrence of lipoid substances. A shift to higher frequencies, on the other hand, can be caused by atrophic changes within the vocal cords. But in the male also feminization after the climacterium can result in utilization of falsetto-voice.
Other voice clues of old age are vocal loudness and speaking rate.
Diminishing vitality is demonstrated in the reduction of intensity which is perceived as a reduction in vocal loudness.
And in an investigation of Ptacek and Sander (1966) listener-judges report rate of reading as an obvious clue for differentiating aged from younger persons. Ptacek and Sander measured in the same study words per minute and found significant rate differences for both impromptu speaking and oral reading between aged and younger subjects. This in contrast to Mysak (1959, p. 52) who did not find this difference for impromptu speaking, although he did measure a reduction of time measures (words per minute; phonation/time ratio) as a function of age for reading.

Concluding we can say that on the whole listeners are able to estimate with fair success the age of a speaker on the basis of vocal cues alone. In the impression of the listener this is done directly from the voice in all its complexity, because to the listener voice is an inextricable pattern of pitch, loudness, timbre and rhythm.
About what happens in the aging larynx quite a lot is known physiologically. From the acoustic effects resulting from the anatomic alterations, in particular pitch attributes have been analysed and, the importance accounted to fundamental frequency is indicated in quite a lot of acoustic studies (see Appendix A).
But changes in voice quality and voice dynamic features are likely to carry much more information about, among other things, age. Research in this area however, has been neglected; empirical data on the perceptual aspects of the relation between age and speech are very scarce. (See Hartman and Danhauer, 1976; Ptacek and Sander, 1966; Ryan and Burk, 1974; Shipp and Hollien, 1969).
Even more lacking, virtually non-existent, are studies on the relationship between objectively measured acoustic stimuli and the perceptually communicated percepts of the particular individual characteristics i.c. age. From the foregoing it is clear that much more research on the identification of relevant parameters on both levels must be done before their relationship can be studied systematically.

3. EXPERIMENT

The experiment consists of an evaluative rating by 29 listeners based on 2 minutes oral reading text by 50 Dutch speakers from 18 - 87 years-of-age. On basis of voice alone speakers' ages are judged. The speakers are divided into subgroups according to their ratings on 12 speech therapeutic terms and furthermore they are rated on 10 bipolar semantic scales.

3.1 Rating form

The rating form exists of two parts:
1) a set of semantic differential scales and
2) twelve terms from the speech therapist's terminology
The age of the speakers is assessed twice; before (AR1) and after (AR2) the ratings on the scales and terms are given.

1) The scales
Twelve bipolar seven-point scales were selected from a battery of 35 used by Fagel, van Herpt and Boves (1983) for global perceptual ratings of speakers. This subset was chosen because literature suggested that the pole-terms might pertain to age. In table 1 the English equivalents of the Dutch scale terms and their factorloadings as resulting in the factor-solution in the study of Fagel et al. is given (If less than 10% of the variance of the scale concerned is explained by a factor, the value of that factorloading is omitted.) The last two columns show the communalities (h^2) i.e. the percentage of the total variance of a variable explained by the 5 factors and, as a measure of reliability the mean correlations between eight rater groups for each scale ($\bar{r}_8$).

Table 1. Selected bipolar scales
(For an explanation see text).

English equivalents of Dutch scale terms	F-1	F-2	F-3	F-4	F-5	h^2	$\bar{r}_8$
S01 spiritless-vivacious	0.80					0.84	0.91
S02 expressive-expressionless	-0.80	-0.32				0.82	0.93
S03 colourless-sonorous	0.70		0.45			0.78	0.90
S04 tense-relaxed	0.49	0.33		-0.36		0.59	0.91
S05 polished-slovenly		-0.75				0.74	0.93
S06 distinct-indistinct	-0.35	-0.59	-0.39			0.63	0.91
S11 husky-not husky			0.66			0.57	0.94
S07 dull-clear	0.44		0.63			0.67	0.91
S08 powerful-weak	-0.37	-0.33	-0.63			0.73	0.86
S09 steady-unsteady	-0.36	-0.45	-0.49			0.66	0.93
S12 high- low for a (wo)man				-0.67		0.46	0.92
S10 quick-slow					0.69	0.49	0.88

With this chosen set of rating scales we hope to be able to describe 'age characteristics' on the same dimensions as found by Fagel et al. The scales S01 up to S04 are clear representatives of the dimension 'Voice Dynamics' (F-1); S05, S06 and S09 of 'Articulation Quality' (F-2) which has to do with intelligibility and preciseness of articulation. S03, S07, S08, S09 and S11 of 'Voice Quality' (F-3) which is strongly associated with perceived clarity as well as with subjective strength.
These three factors explain in the study of Fagel et al. 56% of the total variance. The authors also identified a Pitch-factor (F-4) and a Tempo-factor (F-5) which together explain only 6% of the total variance. Nevertheless we incorporated S12 and S10 in our battery because both pitch and tempo seem to pertain to the perception of age.

In a pilot investigation S11 'husky - not husky' and S12 'high - low for a (wo)man' did not differentiate and were eliminated from the battery. 'Husky - not husky' did not function, probably because the speakers did not differ enough from each other with respect to this parameter. 'High - low' is a notoriously difficult scale, maybe because of the variability of the fundamental frequency or because it is difficult to assess pitch independently from loudness or because the judges have to use different reference points for men and women.

2) Speech therapeutic terms
The second part of the rating-form gathers information in the categories coded in table 2.

Table 2. Judgment-categories

Category	Code	Qualifications
Resonance	T01	Good resonance
	T02	Bad resonance
	T03	Predominantly middle register
	T04	Predominantly chestregister
Phonation	T05	Good phonation
	T06	Bad phonation
	T07	Predominantly hypokinetic
	T08	Predominantly hyperkinetic
	T09	Dyskinetic
Respiration	T10	Good breath support
	T11	Bad breath support
	T12	Crams too many words in one exhalation
	T13	Superficial breathing
	T14	Both T12 and T13

Judgments are given by the raters in the form of a true/false qualification. All raters, students in speech therapy in their third year, were familiar with these terms, nevertheless an instruction of the following type was given:
Concerning respiration. Someone continues speaking on one exhalation when the last word (or words) he or she utters sound strained. We speak of shallow or superficial breathing when a speaker inhales too often and at 'wrong' moments in a sentence.

3.2 Speakers

Fifty Dutch male and female speakers from 18 - 87 years-of-age were distributed over 3 age categories as follows:

18 - 39 years	7 female	9 male
40 - 59 "	9 "	8 "
60 - 87 "	11 "	6 "

A lower boundary of 17 years was chosen because at that age every speaker has gone through vocal mutation.

3.3 Speech material

The speech material consisted of an identical text of about 2 minutes duration read aloud by the speakers.
It would have been preferable to use fragments from free conversation because, as mentioned in 2.0, contradictory results are obtained concerning the relation between oral reading rate and speaker's age. But in this case an oral reading text was chosen because it was considered more important to control for between-speaker differences in lexicon and syntax, as this would give information about the age of the speakers.
To enable the listeners to judge breath-control a long compound sentence was introduced into the text.
Before reading the required text the readers had ample opportunity to familiarize themselves with this text.

3.4 Raters

The rating experiment was carried out with a group of 29 female students (aged 19 - 31 years) from the Training Course of Speech Therapists in Amsterdam.
A sample of 30 raters was planned since that would lead to an effective reliability of 0.90 or higher for all bipolar scales according to a procedure described in Fagel et al. (1983). All raters were female because it is known that men and women tend to use somewhat different yardsticks; women seem to use the extremes of the scales in a semantic differential more readily than men. This is especially important for scales like 'spiritless - vivacious' and 'dull - clear' on which female speakers are generally rated closer to the positive extreme, and 'powerful - weak' and 'steady - unsteady' on which

male speakers are mostly rated closer to the positive extreme of the scale (Boves, Fagel and van Herpt, 1982).
Students in their third year of training had been selected because knowledge of terminology of speech science was required for the second half of the evaluation. From an unpublished analysis of the data of the experiment by Fagel et al. (1983) is known that the ratings on the 10 bipolar seven-point scales are not significantly influenced by this sort of training of the judges. (The present experiment has been replicated with Speech Therapy students in their first year of training, to check this finding. Data still have to be processed.)

3.5 The rating procedure

The fifty fragments are presented in one session with a tea-break between the voices of men and those of women. The voices in both parts are presented in random order with respect to age and preceded by three trial-samples to familiarize the listeners with the text, allotted time, scales and scoring categories. The listeners were given two minutes to judge each speech fragment. The complete session required almost 2.5 hours.

3.6 Reliability

3.6.1 The scales

The reliability of a scale can be expressed in a number of ways. For reasons of comparison we calculated the same measure as Fagel et al. (1983), the so-called effective reliability α (Rosenthal, 1973) of each scale.

$$\alpha = n\bar{r}/(1+(n-1)\bar{r}),$$

in which $\bar{r}$ is the mean correlation between raters and n is the number of raters.

Table 3. Reliability of rating scales

Nr	Scales	α
S01	spiritless-vivacious	0.964
S02	expressive-expressionless	0.969
S03	colourless-sonorous	0.952
S04	tense-relaxed	0.924
S05	polished-slovenly	0.957
S06	distinct-indistinct	0.959
S07	dull-clear	0.965
S08	powerful-weak	0.975
S09	steady-unsteady	0.965
S10	quick-slow	0.972

As can be seen, α is dependent on the number of raters involved. To achieve a high reliability we used 29 raters which resulted in an $\alpha > 0.92$ for all scales (see table 3). This fits in with Fagel et al. who report that 'for all scales an effective reliability of 0.90 or higher resulted when at least 25 raters were involved, except for the scales 'loud-soft' and 'quick-slow' that require about 30 raters to reach an effective reliability of 0.90.' (p.322.)

3.6.2 The speech therapeutic terms

The reliability formula used for the scales is not suitable for answers of the false/true type. In order to get some insight into the measure of agreement between the listeners in their judgments on the therapeutic terms chi-square tests were used to ascertain if the distributions within the categories are based on chance.
The distribution of the observations in the categories is shown in table 4 in percentages. The probability of the distributions being attributable to chance is < 0.001 in every case, so that the conclusion may be drawn that the ascription of the qualifications is not arbitrary.

Table 4. Distribution of observations in percentages over categories. (Meaning of codes in table 2)

Category-codes	N observations		Percentages	
Resonance	1223			
T01 (Good)	546		44.6	
T02 (Bad)	677		55.4	
T03 (Middle)		357		56.6
T04 (Chest)		274		43.4
Phonation	1375			
T05 (Good)	564		41.0	
T06 (Bad)	811		59.0	
T07 (Hypo)		167		20.6
T08 (Hyper)		471		57.9
T09 (Dis)		175		21.5
Respiration	1379			
T10 (Good)	525		38.1	
T11 (Bad)	854		61.9	
T12 (Shallow)		258		30.2
T13 (Superf.)		394		46.1
T14 (12+13)		203		23.7

4. RESULTS

4.1 Rating 1 versus rating 2

Age ratings are given twice. The first age rating (AR1) is an assessment made by the listeners after hearing the speakers during a few seconds. The second assessment is made after finishing the scoring on scales and therapeutic terms of each speaker, i.e. after hearing the speakers during two minutes.
The two sets of ratings are highly correlated ($r = 0.98$) but although the difference between the means of AR1 and AR2 is less than one year (see table 5), comparison of AR1 and AR2 of the same individuals by means of a dependent t-test showed the differences to be systematic. In almost all cases the second rating is higher than the first one. This holds for all ages, for female as well as for male speakers, and just as well for overrating the age as for underrating.
This result suggests that the parameters relevant for age assessment can be 'extracted' from speech samples of only a few seconds.

In the following paragraphs calculations concerning AR1 are left out because the differences between the first rating and the second are small and systematic.

Table 5. Mean age ratings (AR1 and AR2) and mean true age (AGE) of male and female speakers of each speaker-age group.

Age-group	Female				Male				Total			
	N	AR1	AR2	AGE	N	AR1	AR2	AGE	N	AR1	AR2	AGE
18-39	7	26.1	26.3	25.1	9	34.0	34.1	26.2	16	30.6	30.7	25.8
40-59	9	43.6	44.2	50.1	8	44.1	44.7	48.8	17	43.9	44.4	49.5
60-87	11	63.4	64.4	77.6	6	64.8	65.8	74.3	17	63.9	64.9	76.4
Total	27	47.2	47.8	54.8	23	45.6	46.1	46.6	50	46.4	47.0	51.1

4.2 Age ratings

First, the results show that there is a close relation between rated and real ages. The correlation between AR2 and AGE is 0.82.
Nevertheless in a t-test with paired observations of all speakers the difference between AGE and AR2 is significant ($p < 0.001$).
Closer analysis revealed that the rating errors made are a function of age of the speakers. The rating errors expressed as the mean of the absolute differences between AR2 and AGE for the subsequent age-groups are 6.3, 10.0 and 12.7 years respectively. ($p < 0.001$).
Also, as can be seen from table 5, there is a central tendency in the

ratings: older people are rated too young, younger people too old. However this error is not significant for female speakers in the youngest age group (18 - 31 years) which is exactly the same age group as the one we recruited the female listeners from. The correlation AR2 - AGE for this group is 0.97. This result suggests an interaction of the variables age and sex between listeners and speakers.
The picture is further complicated by a speaker-sex-effect. In all age groups the mean rated age of men is higher than that of women.

4.3 Age and therapeutic terms

With the aid of t-tests in the case of two judgment-categories and with an one-way analysis of variance followed by Tukey's Honestly Significant Difference Test in the case of three therapeutic subcategories we analyzed the different speech therapeutic qualifications on differences between rated and true ages. (See table 6)

Table 6. Mean age of subjects categorized according to speech therapeutic terms. (Meaning of the codes in table 2.)
All differences are significant ($p < 0.01$) except between subgroups indicated with °.

Categories		main groups		Categories	Subgroups	
		AR2	AGE		AR2	AGE
RES.	T01 (Good)	42.86	45.98	T03 (Middle)	45.46	50.12°
	T02 (Bad)	49.25	53.63	T04 (Chest)	49.48	50.14°
	Diff.	-6.39	-7.65	Diff.	-4.02	-0.02
PHON.	T05 (Good)	40.41	42.66	T07 (Hypo)	44.57	46.82
	T06 (Bad)	51.09	55.60	T08 (Hyper)	52.90°	57.86°
	Diff.	-10.68	-12.94	T09 (Dys)	52.40°	57.73°
RESP.	T10 (Good)	42.44	45.99	T12 (Shallow)	50.22°	55.42°
	T11 (Bad)	49.42	53.03	T13 (Superf.)	48.41°	50.24
	Diff.	-6.98	-7.04	T14 (12+13)	50.48°	55.54°

For all main categories the 'good' speakers are significantly younger than the 'bad' speakers both on rated and on true age. In other words, listeners consider speakers to be younger when their resonance (T01/02), respiration (T10/11) and/or especially their phonation (T05/06) is good. Their assumptions seem to be justified since the real AGE of the 'good' speakers is in reality also significantly lower than that of the bad speakers (Table 6). Speakers who use the chestregister (T04) are rated four years older than those who speak in the middle register (T03). In this case, however, there

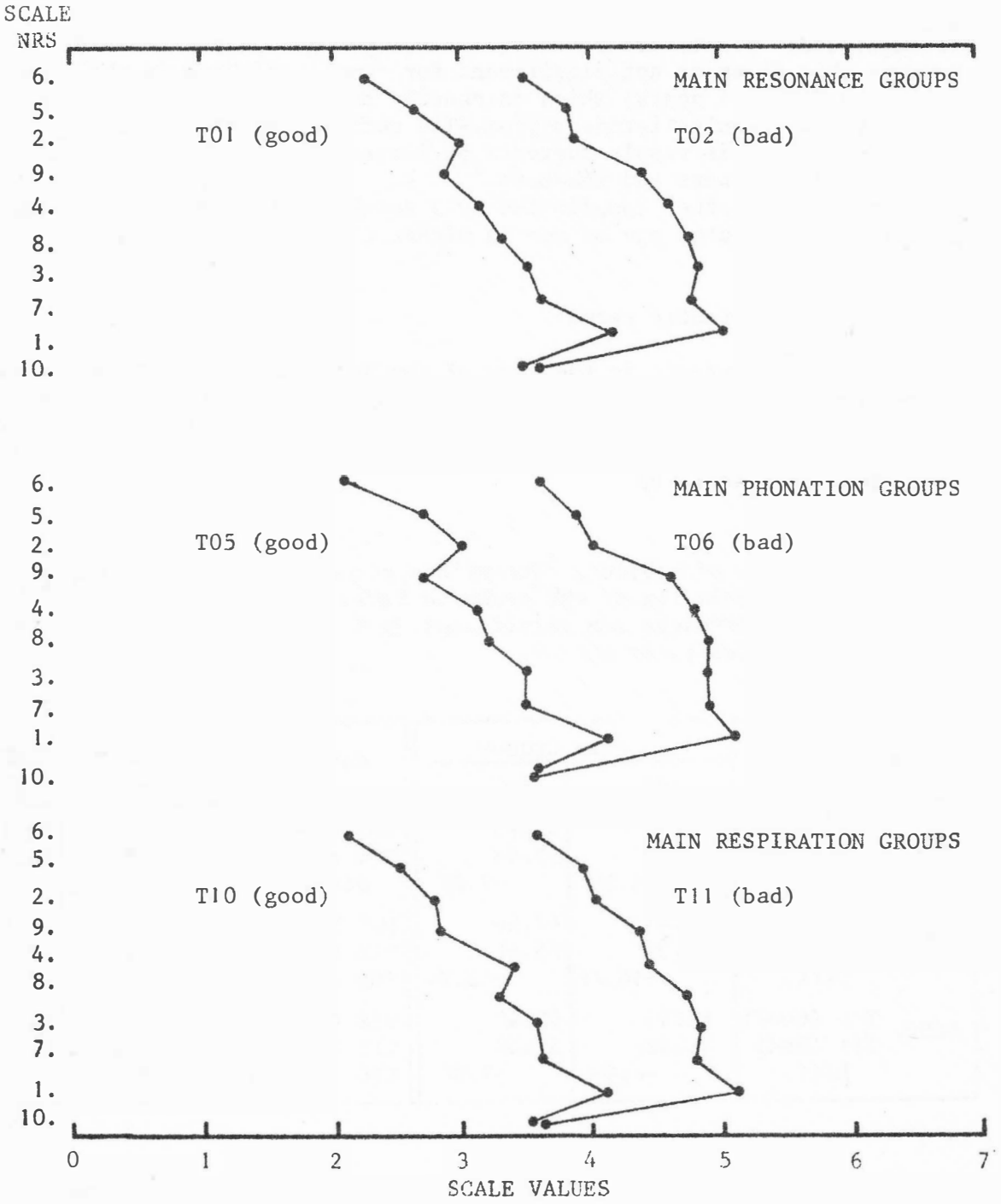

Fig. 1. Mean scores on ten bipolar scales of 'good' and 'bad' speakers on three therapeutic concepts.
All scales are oriented with the positive poles to the left.

is no difference in real age between both groups. This attribution error suggests that rating a higher pitched voice younger than a low-pitched voice is a stereotype.
The group with bad phonation is subdivided in three groups (T07, T08 and T09). There is no difference in real or rated age between the hyper- and dyskinetic groups, but both are estimated about eight years older than the hypokinetic group. The difference is in reality even more, namely 11 years, so these concepts are apparently used in the attribution of age.
Between the three subgroups of speakers with bad breath-support (T12, T13 and T14) there is no significant difference in mean age ratings although the mean real age of the 'superficial' breath group is five years younger. In this aspects vocal age seems not to coincide with chronological age

4.4 Scales and therapeutic terms

Relating the scales to the therapeutic terms reveals a high correlation between both ratings. For all scales, except 'quick-slow' (S10), the mean scores are significantly more on the positive pole of the scale for speakers rated as 'good' in Resonance or Phonation or Breathing. The profiles (see fig. 1) differentiate clearly between 'good' and 'bad' groups of speakers. The correspondence of the three profiles is caused by the interrelation of the categories Resonance, Phonation and Respiration: 12 from 50 speakers were scored as 'good', 24 speakers as 'bad' on all three categories. (For further analyses the true/false qualifications will be scaled.) The 'good' and 'bad' groups, on their turn differentiate significantly both on rated and true age, as shown in table 6.
Nevertheless the correlations between the ten bipolar scales and AR2 are, although all significant at less than 1%, never higher than 0.44. Those between AGE and the scales are all in the same direction but on the average 0.05 lower (Table 7)

Scales	AR2	AGE
S01	0.204	0.120
S02	0.193	0.112
S03	0.267	0.184
S04	0.296	0.295
S05	0.210	0.153
S06	0.271	0.252
S07	0.371	0.300
S08	0.248	0.231
S09	0.439	0.417
S10	0.170	0.121

Table 7. Correlations between age and scales. All scales are oriented with the positive pole to the left. $p < 0.01$ for all correlations.

The significance of the scales suggests that acceptable scales were used. The values being low means that the scale values cannot be used for the prediction of age without some form of manipulation. Further analysis, e.g. stepwise multiple regression analysis and the delimiting of weighting factors may be necessary. However, the scales can give an insight into the factors that place the speaker groups in perceptual space. E.g. the differences between two subgroups of bad phonation, hypo- (T07) and hyperkinetic (T08), are significant on scales S01 up to S04, S07 and S08. Except on S04 the hyperkinetic group is assessed on all scales more positively (i.e. in the direction of the scores for good phonation). In other words a hyperkinetic speaker is perceived as more vivacious, more expressive and more sonorous, as clearer and more powerful but as less relaxed than a hyperkinetic speaker - but both are in all these respects less so than the speaker with good phonation (T05). In figure 2 the relative differences of these six differentiating bipolar scales are represented in scale units.
And again, differences in ages between the three groups are in the same direction: AR2 - T05: 40.41 T08: 44.57 T07: 52.90 years
AGE - T05: 42.66 T08: 46.82 T07: 57.73 years

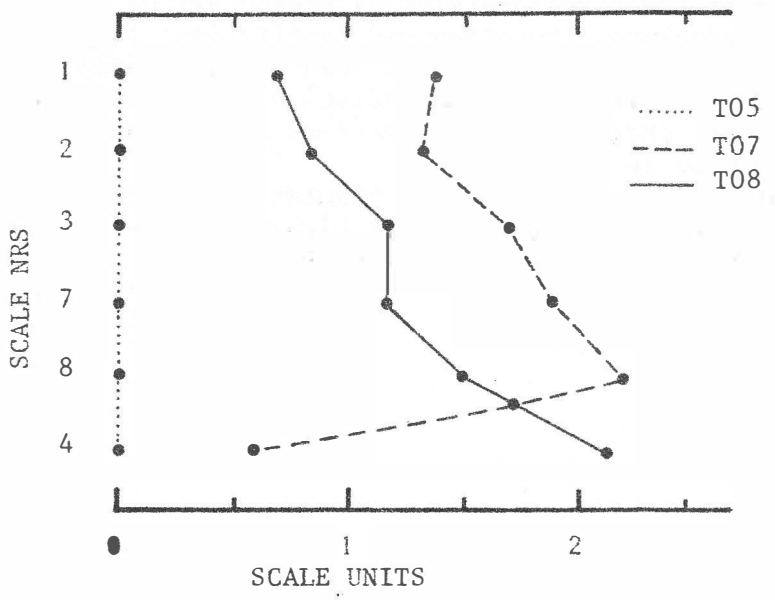

Fig. 2. Relative differences - in scale units - between hypo- (T07) and hyperkinetic (T08) speakers compared on six bipolar scales with 'good' phonating speakers (T05). All scales are oriented with the positive poles to the left.
All differences are significant ($p < 0.001$)

5.0 CONCLUDING REMARKS

First of all, it must be emphasized again that neglect of the perceptual attributes of - among other things - age judgment from voice stimuli, is unwarranted.
Listeners appear to be able to estimate speakers' ages fairly well from vocal parameters. In general, speakers who score high on voice quality features and on voice dynamic features are supposed to be younger than speakers who do worse. A supposition which appears justified since there is a high correlation between chronological age and its estimation in this study. None the less, the prediction of the true age of one individual remains precarious, because there are several interfering and interacting variables. For instance, there is a central tendency i.e. older people are rated too young, younger people too old. But at the same time there seems to be an interaction between speakers' age and listener's age, i.e. somebody's age is best estimated by someone of the same age. Since the listeners in the present study were all of the youngest age group, this age-interaction phenomenon might also account for our finding that the size of the rating errors are a function of age of the speaker. Moreover there can be a listener-speaker sex interaction, although it is also possible that - for sociological causes - only women make this rating-error (estimating men's age too high). Differences in ratings may also occur as a result of the material used (oral reading versus impromptu speaking), of the time allotted to the listeners to give their ratings (seconds versus minutes) or of the type of listener taking part in the experiment (trained versus untrained).
The differentiation of the perceptual ratings in the present study give confidence that this type of descriptive tools are useful in establishing criteria for defining speaker characteristics.
The bipolar scales did function well but the relation to the therapeutic terms is not quite clear. This is caused by the correlation between the categories Resonance, Phonation and Respiration. Further analyses on the relation of scales and terms will be performed but are hampered by the fact that the therapeutic terms are merely scored as True or False. Other rating procedures for this type of terminology are preferable e.g. having the terms rated in small clusters of between 2 and 4 on 7- and 4-point scales, as proposed by van Bezooijen (1984), or simply by using as a score the percentage of judges that give a positive rating.
From the battery of bipolar scales the scale 'high-low' was eliminated because it did not function properly (see 3.1). In further research a solution for this notoriously difficult scale must be found to enable the description of speakers on the pitch dimension of perceptual space. Given the amount of available physical data concerning fundamental frequency it is desirable that subsequent studies are conducted relating acoustic measures connected with Fo both to age and sex and to the perception of age and sex.
As indicated already in the last sentence, in subsequent studies age-indices will be investigated in combination with other biologically determined markers in speech. Much is still unknown about the exact nature of the cues which are used in the attribution process, but listeners assess biological characteristics such as sex, age and physique with reasonable validity, reliability and accuracy from vocal cues as has been shown, among others by

Lass et al. (1976, 1978, 1980). Since these cues have perceptual importance they should be useful in establishing criteria for the evaluation of coded speech as well as for defining general speaker-characteristics which can be used in a stepwise process of automatic recognition.

ACKNOWLEDGEMENT

We would like to thank Heleen Deighton-van Witsen, Wil Fagel and Louis Pols for reading an earlier version of this article and for their comments on it.

REFERENCES

Allport, G.W. (1934). Judging personality from voice. J. of Soc. Psych., 5, pp. 37-55.

Bach, A.C., Lederer, F.L. & Dinolt, R. (1941). Senile changes in the laryngeal musculature. Arch. Otolaryng., 34, 47-56.

Bezooijen, R. van (1984). The auditory description of long-term speech characteristics: Some practical remarks. Proceedings 8, p. 1-12. Catholic University, Nijmegen.

Boves, L., Fagel, W.P.F. & Herpt, L.W.A. van (1982). Opvattingen van vrouwen en mannen over de spraak van mannen en vrouwen. De Nieuwe Taalgids 75(1), 1-23.

Charlip, W.S. (1968). The aging female voice: selected fundamental frequency characteristics and listeners judgements. Dissertation, Purdue University. (Dissertation Abstracts, 1968, 29 B, 1198-B.)

Damstee, P.H. (1973). De pathologische stembandfunktie. Stafleu's Wetenschappelijke Uitg. mij b.v., Leiden.

Endres, W., Bambach, W. & Flösser, G. (1971). Voice spectrograms as a function of age, voice disguise and voice imitation. J. Acoust. Soc. Am., 49, 1842-1848.

Fagel, W.P.F., Herpt, L.W.A. van & Boves, L. (1983). Analysis of the perceptual qualities of Dutch speaker's voice and pronunciation. Speech Communication 2, 315-326.

Ferreri, G. (1959). Senescence of the larynx. It. Gen. Rev. Oto-rhinolaryng., 1, 640-709.

Hartman, D.E. & Dannauer, J.L. (1976). Perceptual features of speech for males in four perceived age decades. J. Acoust. Soc. Am., 59, p. 713-715.

Helfrich, H. (1979). Age markers in speech. p. 63-107. In: K.R. Scherer & H. Giles, Social markers in Speech. Cambridge University Press.

Henton, C.G. (1982). Aspects of sex-specific language difference at the phonatic and phonological levels. Thesis. Phonetics Laboratory, Oxford.

Herzog, H. (1933). Stimme und Persönlichkeit. Zschr. Psychologie, 130, 300-379.

Inhofer, R. (1913). Die Ermüdung der Stimme (Phonastenia). Verlag Kurt Kabitzsch, Würzburg.

Lass, N.J., Hughes, K.R., Bowyer, M.D., Waters, L.T. & Bourne, V.T. (1976). Speaker sex identification from voiced and filtered isolated vowels. J. Acoust. Soc. Am., 59, 675-678.

Lass, N.J. & Davis, M. (1976). An investigation of speaker height and weight identification. J. Acoust. Soc. Am., 60, 700-703.

Lass, N.J. Beverly, A.S. Nicosia, D.K. & Simpson, L.A. (1978). An investigation of speaker height and weight identification by means of direct estimation. J. of Phonetics, 6, 69-76.

Lass, N.J., Almerino, C.A., Jourdan, L.F. & Walsh, J.M. (1980). The effect of filtered speech on speaker race and sex identification. J. of Phonetics, 8, 101-112.

Lass, N.J., Brong, G.W., Ciccolella, S.A., Walters, S.C. & Maxwell, E.L. (1980). An investigation of speaker height and weight discriminations by means of paired comparison judgments. J. of Phonetics, 8, 205-212.

Lass, N.J., Hendricks, C.A. & Iturriaga, M.A. (1980). The consistency of listener judgments in speaker height and weight identification. J. of Phonetics, 8, 439-448.

Laver, J. & Trudgill, P. (1979). Phonetic and linguistic markers in speech. 1-32. In: K.R. Scherer & H. Giles, Social markers in Speech. Cambridge University Press.

Luchsinger, R. & Arnold, G.E. (1959). Lehrbuch der Stimm- und Sprachheilkunde. Springer, Berlin.

Luchsinger, R. & Arnold, G.E. (1965). Voice-Speech-Language. Clinical Communicology: Its Physiology and Pathology. Wadsworth Publishing Co., Inc. Belmont, California.

McGlone, R.E. & Hollien, H. (1963). Vocal pitch characteristics of aged women. J. Speech Hearing Research, 6, 164-70.

Mysak, E.D. (1959). Pitch and duration characteristics of older males. J. Speech Hearing Research, 2, 46-54.

Pear, T.H. (1931). Voice and Personality. London: Chapman & Hall.

Ptacek, P.H. & Sander, E.K. (1966). Age recognition from voice. J. Speech Hearing Research 9: 273-277.

Rosenthal, R. (1973). Estimating effective reliabilities in studies that employ judges' ratings. J. Clinical Psychol., Vol. 29, 342-345.

Ryan, W.J. & Burk, K.W. (1974). Perceptual and acoustic correlates of aging in the speech of males. J. Comm., 7, 181-192.

Sanford, F.H. (1942). Speech and personality. Psychological bulletin, 39, 811-845.

Shipp, F.T. & Hollien, H. (1969). Perception of the aging male voice. J. Speech Hearing Research, 12, 703-10.

Stern, H. (1932). Ueber einige Ursachen and Folgen der Alterserscheinungen in der Stimmgebung. Mschr. Ohrenheilkunde, 66, p. 665.

Stoicheff, M.L. (1981). Speaking fundamental frequency characteristics of nonsmoking female adults. J. Speech Hearing Research, 24, 437-441.

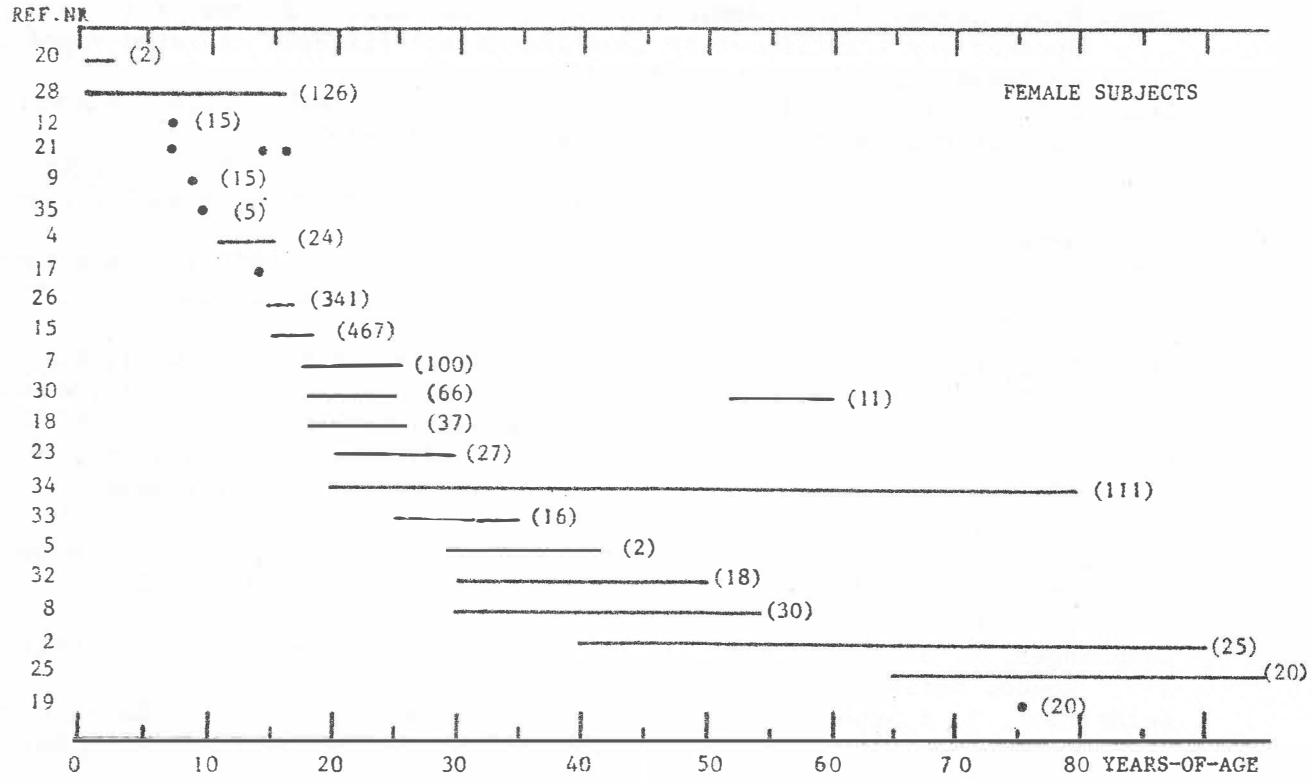

Fig. I. Age groups of female subjects in studies on fundamental frequency. Age in years is indicated on the horizontal axis. Between brackets the number of subjects included in each study. The numbers along the vertical axis refer to the references given in the appendix.

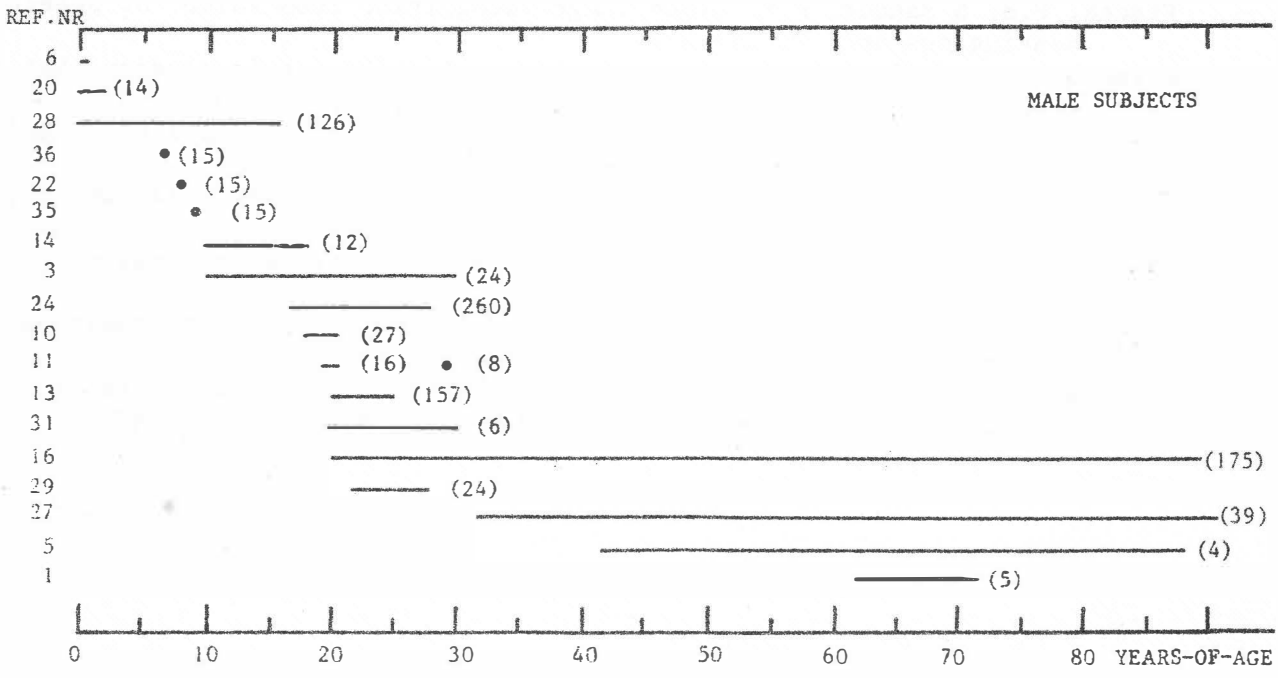

Fig. II. Similar to fig. I - Male subjects.

APPENDIX A

In figure I and II a summary is given of acoustic studies of vocal pitch characteristics of male and female subjects of different ages. The horizontal lines in both figures represent the age ranges of the subjects used in the studies which are indicated by the reference numbers along the vertical axis. We do not report actual data of Fo measurements because the studies differ in various relevant variables. E.g. in several cases it is not clear what kind of speech material is used (oral reading or impromptu speaking). Variables that are controlled only in a few investigations are (amount of) smoking of cigarettes, or (amount of) stress. Concerning the subjects it is not always clear if the measures are determined pre- or post-menopause/climacterium, or if the women did or did not use the pill, and with young subjects if they are male or female.

Notwithstanding the fact that this is but a limited summing up of variables involved and even though there are many studies on the subject, it will be clear that there is a great need of longitudinal studies, preferably executed cross-culturally

REFERENCES OF APPENDIX A

1. Bach, A.C., Lederer, F.L. & Dinolt, R. (1941). Senile changes in the laryngeal musculature. Arch. Otolaryngology, 34, 47-56.
2. Charlip, W.S. (1968). The aging female voice: selected fundamental frequency characteristics and listener judgements. Ph.D.Diss., Purdue University.
3. Curry, E.T. (1939). An objective study of the pitch characteristics of the adolescent male voice. Ph.D.Diss., State University of Iowa. (Shortened version in: Speech Monographs, 1940, 7, 48-62: The pitch characteristics of the adolescent male voice.)
4. Duffy, R.J. (1958). The vocal pitch characteristics of eleven-, thirteen-, and fifteen-year-old female speakers. Ph.D.Diss., Univ. of Iowa.
   (Shortened version in: Language and Speech, 1970, 13, 14-24: Fundamental frequency characteristics of adolescent females.)
   Note: The unpublished data for the mean fundamental frequencies of the 11, 13 and 15 year-olds of Duffy's study of 1958 contain a slight calculation error. Table 1, p. 16 of the article of Duffy based on this dissertation contains the correct figures.
5. Endres, W., Bambach, W. & Flösser, G. (1971). Voice spectrograms as a function of age, voice disguise and voice imitation. J. Acoust. Soc. Am., 49, 1842-1848.
6. Fairbanks, G. (1942). An acoustical study of the pitch of infant hunger wails. Child Development, 13, 227-232.
7. Fitch, J.L. & Holbrook, A. (1970). A modal vocal fundamental frequency of young adults. Arch. Otolaryngology, 92, 379-382.

8. Gilbert, H.R. & Weismer, G.G. (1974). The effects of smoking on the speaking fundamental frequency of adult women. J. Psycholinguistic Research, 3, 225-231.
9. Hammond, J.M. (1947). An objective study of the pitch characteristics of eight year-old girls during oral reading. M.A. Thesis, State University of Iowa.
Note: Data from this thesis are used in: Fairbanks, G., Herbert, E.L. & Hammond, J.M. (1949). An acoustic study of vocal pitch in seven- and eight-year-old girls. Child Development, 20, 71-78.
See also: Fairbanks, G. (1950). An acoustic comparison of vocal pitch in seven- and eight-year-old children. Child Development, 21, 121-129.
10. Hanley, T.D. (1949). An analysis of vocal frequency and duration characteristics of selected samples of speech from general American, eastern American and southern American dialect regions. Ph.D. Diss., University of Iowa.
Article based on this dissertation: An analysis of vocal frequency and duration characteristics of selected dialect regions. Speech Monographs, 1951, 18, 78-93.
11. Hanley, T.D., Snidecor, J.C. & Ringel, R.L. (1966). Some acoustic differences among languages. Phonetica, 14, 97-107.
12. Herbert, E.L. (1942). An objective study of the pitch characteristics of seven-year-old girls during oral reading. M.A. Thesis, State University of Iowa.
See note under 9.
13. Hollien, H. & Jackson, B. (1973). Normative data on the speaking fundamental frequency characteristics of young adult males. J. of Phonetics, 1, 177-200.
14. Hollien, H. & Malcik, E. (1967). Evaluation of cross-sectional studies of adolescent voice change in males. Speech Monographs, 34, 80-84.
15. Hollien, H. & Paul, P. (1969). A second evaluation of the speaking fundamental frequency characteristics of post-adolescent girls. Language and Speech, 12, 119-124.
16. Hollien, H. & Shipp, F.T. (1972). Speaking fundamental frequency and chronological age in males. J. Speech Hear. Res., 15, 155-159.
17. Hollien, H., Malcik, E. & Hollien, B. (1965). Adolescent voice changes in southern white males. Speech Monographs, 32, 87-90.
18. Hollien, H., Tolhurst, G.C. & McGlone, R.E. (1982). Speaking fundamental frequency as a function of vocal intensity. Unpublished report to the Dreyfus Foundation.
(Cited by O. de Pinto and H. Hollien, J. Phonetics, 1982, 10, p. 369.)
19. Honjo, I. & Isshiki, N. (1980). Laryngoscopic and voice characteristics of aged persons. Arch. Otolaryngology, 1980, 106, 149-150.
20. Keating, P. & Buhr, R. (1978). Fundamental frequency in the speech of infants and children. J. Acoust. Soc. Am., 63, 567-571.

21. Cited in several articles with different titles:
Krikelas, H.
1. An analysis of the pitch characteristics of seven-year-old girls during oral reading. M.A. Thesis, Stanford University, 1959.
2. An analysis of the fundamental frequency characteristics of 14- and 16-year-old female speakers. M.A. Thesis, Stanford University.

22. Lassman, F.M. (1942). An objective study of the pitch characteristics of eight-year-old boys during oral reading. M.A. Thesis, State University of Iowa.
Note: Data from this thesis are used in: Fairbanks, G., Wiley, J.M. & Lassman, F.M. (1949). An acoustical study of the vocal pitch in seven and eight-year-old boys. Child Development, 20, 63-69.
See also: Fairbanks, G. (1950). An acoustic comparison of vocal pitch in seven- and eight-year-old children. Child Development, 21, 121-129.

23. Linke, E. (1973). A study of pitch characteristics of female voices and their relationship to vocal effectiveness. Ph.D.Diss., University of Iowa.
Article based on this dissertation and under the same title is published in Folia Phoniatrica, 1953, 25, 173-185.

24. Majewski, W., Hollien, H. & Zalewski, J. (1972. Speaking fundamental frequency of Polish adult males. Phonetica, 25, 119-125.

25. McGlone, R.E. & Hollien, H. (1963). Vocal pitch characteristics of aged women. J. Speech Hear. Res., 6, 164-170.

26. Michel, J.F., Hollien, H. & Moore, P. (1966). Speaking fundamental frequency characteristics of 15-, 16- and 17-year-old girls. Language and Speech, 9, 46-51.

27. Mysak, E.D. (1959). Pitch and duration characteristics of older males. J. Speech Hear. Res., 2, 46-54.

28. Oordt, H.W.A. van & Drost, H.A. (1963). Development of the frequency range of the voice in children. Folia Phoniatrica, 15, 289-298.

29. Philhour, C.W.Jr. (1948). An experimental study of the relationships between perception of vocal pitch in connected speech and certain measures of vocal frequency. Ph.D.Diss., University of Iowa. (Cited a.o. in 4, 23, 24 and 25.)

30. Pinto, O. de & Hollien, H. (1982). Speaking fundamental frequency characteristics of Australian women: then and now. J. of Phonetics, 10, 367-375.

31. Pronovost, W. (1942). An experimental study of methods for determining natural and habitual pitch. Speech Monographs, 9, 111-123.

32. Saxman, J.H. & Burk, K.W. (1967). Speaking fundamental frequency characteristics of middle-aged females. Folia Phoniatrica, 19, 167-172.

33. Snidecor, J.C. (1940). Studies in the pitch and duration characteristics of superior speakers. Ph.D.Diss., Iowa City. Article based on this dissertation. J. Speech Hear. Res., 1951, 16, 44-52.

34. Stoicheff, M.L. (1981). Speaking fundamental frequency characteristics of nonsmoking female ad lts. J. Speech Hear. Res., 24, 437-441.

35. Weenink, D.J.M. (1985). Fundamental frequency of nine-year-old speakers. Unpublished data, IFA University of Amsterdam.
36. Wiley, J.E. (1942). An objective study of pitch characteristics of seven-year-old boys during oral reading. M.A. Thesis. State University of Iowa.
See note under 22.