6 Relationships between auditorily and visually evaluated speech characteristics

Melanie Willing

6.1 Introduction

Talking to an unknown person without seeing him or her, we often try to imagine the outer appearance of that person from the voice as the only indication. But we find from telephone conversations, or radio reporters, when we finally meet these persons, that sometimes the look is totally different from our expectations. On the other hand we are quite often right in our imagination and the voice of a person *fits* to the appearance; that is the voice perfectly describes the look of the speaker.

The question is whether there are certain speech characteristics that are related to the speakers' outer appearance. Can parameters like formant frequencies or fundamental frequency predict the physical appearance of a person? Can body structures like height and weight explain speech characteristics like pitch?

There are some studies which examined whether the height of pitch or formants correlate with height and weight of a speaker. There are evolutionary based theories like John Ohala's frequency code (Ohala 1983 and 1994) which establish a connection between animal sounds and human language. The theory of the frequency code says that animals and humans use low frequencies for signaling anger and superiority while high pitched voices signal submissiveness and amiability. Animals try to signal their readiness to combat beside the height of pitch via visual means. For example, lions erect their manes, cats arch their backs and peacocks erect their feathers in a way that it looks like a big wheel. These means advise a potential aggressor against the strength of the signaler. Animal sounds also signal the degree of aggressiveness. Low pitched roaring lions show a bigger potential of aggressiveness than a puppy whimpering with a high pitch. Therefore, the frequency code theory concludes that low frequencies signal a tall speaker while high pitched tones signal a small speaker.

Fundamental frequency as well as the formants might therefore give inference to the height and weight of a speaker. The fundamental frequency is a product of the vibrating vocal folds. The bigger and heavier the vocal folds are, the slower the frequency of vibration is. The second indicator for height and weight of a speaker could be the formants, which depend on the length of the vocal tract. Taller people have longer vocal tracts and thus formants might give a direct conclusion to the outer appearance of a speaker.

A number of phoneticians have surveyed whether the formant frequencies or the height of the fundamental frequency can give information about the outer appearance of a speaker. Künzel (1989) made acoustical examinations of fundamental frequency and found no significant correlations between a speaker's height, weight and his pitch. González (2004) only found a weak relationship between a speaker's physics and the formants. Fitch (1997) found that the formant dispersion significantly correlates with the height and weight of rhesus macaques. Acoustical analyses made by van Dommelen & Moxness (1995) showed significant correlations between speech rate and the weight of a speaker. This effect was only significant for the male group of speakers, with heavier speakers having a lower speech rate.

In their study van Dommelen & Moxness (1995) also made a perceptual experiment where they found again an effect of sex. The height and weight of male speakers was recognized by male listeners better than by female listeners.

In the following, an investigation is described which had the aim to discover whether people can classify other people's voices regarding their sex, height, weight and age. Two perceptual experiments were made in which participants should attach voice stimuli to pictures of possible speakers, having different physical appearances. Audiovisual tests like these also have been done by Bonaventura (1935) and Lass & Harvey (1976) but no other studies could be found having audiovisual components.

6.2 Method

6.2.1 Voice stimuli

In the run-up to the experiments a speech corpus was generated which contains 30 sentences read by 63 male and 63 female speakers. The height and weight of these people were intentionally deviant from the standard gathered by the German government. The speakers were at least 5% over or under the average height for the certain age classes. Regarding the weight, a speaker was classified as heavy, when he or she was 20% over the ideal weight (here the body mass index was taken as a reference). The thin speakers had to be 5% under their ideal weights. The threshold for thin speakers had to be raised because it was very difficult to find people being 20% under their ideal weights.

Height and weight of all speakers were measured, and additionally a photograph was taken from everybody showing a full-length frontal view.

A detailed description of the speech corpus can be found in Baumeister & Willing (2010).

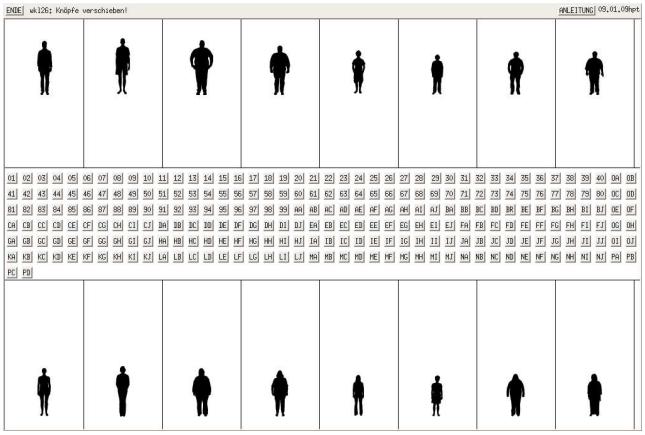

Fig. 6.1: The first experiment before it has been run. In the upper part all male speakers are listed, in the lower part all female speakers. The left half shows the tall speakers, on the right the small speakers are shown. In each size group thin speakers were placed on the left while heavy speakers were placed right. This resulted in pairs of speakers being classified into sex, height and weight. The left member of each pair represented the younger one.

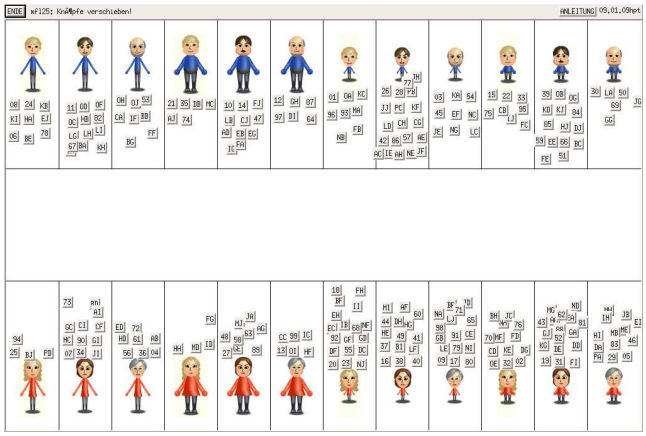

Fig. 6.2: The 2nd experiment with a possible classification. The adjustment of the pictures is the same as in experiment 1, except the addition of a middle age group, positioned between the young and old Mii figures.

6.2.2 Experiment 1

From the database two syntactically different sentences were chosen for the perceptual experiments:

1. Können wir nicht Tante Erna besuchen? (engl. *Can't we visit aunt Erna?*)
2. Achte auf die Autos! (engl. *Mind the cars!*)

The graphical material for the first experiment consisted of the pictures which were taken of the speakers. All speakers were classified into groups of sex, age, weight and height. This resulted in 16 groups: eight male and eight female groups, each having four groups of tall speakers and four groups of small ones. Moreover, each of these groups was divided into groups of heavy and thin speakers. Finally, all speakers were classified by age: the group of young speakers ranged from 17 to 44 years while the group of old speakers contained ages from 45 to 88 years.

A photograph from one speaker of each group was taken that represented the characteristics of that group as well as possible. Then a silhouette of that picture was made.

The perception test was based on a graphical user interface (Pfitzinger 2010) with 16 pictures and 252 little buttons carrying the voice stimuli (Fig. 6.1). By clicking on a button a test sentence was played. Clicking and holding allowed a button to be moved to one of the pictures.

5 male and 7 female participants between 21 and 27 years were asked to attach the sentences to possible speakers. They did not know which physical characteristics were shown by the certain pictures.

During the course of the experiment it was observed that many participants had difficulties in recognizing the properties of the pictures. The sex of the speaker could hardly be recognized via the silhouettes and therefore female voices were assigned to pictures of male speakers and the other way round. Furthermore, the age of the speakers could not be recognized correctly on simple shadow images. The participants had to orient themselves on the clothes of the speakers, which brought up some new confusion. Some older speakers wore clothes which look more juvenile than the clothes of the young ones, see for example the first two pictures in the left upper corner of the first experiment. Due to his short trousers, the right speaker looks more youthful than his younger counterpart on his left side.

In order to avoid that possible wrong classifications are caused by bad visual material other possibilities were sought to display the physical characteristics height and weight as well as sex and age.

6.2.3 Experiment 2

One alternative was found in avatars of the *Nintendo Wii* paddles.[1] In these games it is possible to create characters and to adjust height, weight, sex, age, as well as the color of eyes and hair. These so called *Mii* characters were used in the 2nd experiment to display the certain groups of sex, height, weight and age. A middle age group was created because it was assumed that participants would find it hard to classify voices of 40-year-old speakers as young, as they had to do in the first experiment. The middle age group comprised ages from 35 to 55 years. Due to the third age group the number of pictures raised to 24 (Fig. 6.2).

31 volunteers (15 male and 16 female) between 22 and 37 years participated in the 2nd perception test.

6.3 Hypotheses

Some assumptions were made about the results. It was hypothesized that the factors sex and age would be reflected better in the voice than body characteristics such as height and weight. There are known differences between male and

1 My Avatar Editor, http://myavatareditor.blogspot.com/, [Online, accessed 8.8.2009]

Classification by listeners \ Actual groups	mtly	mtlo	mthy	mtho	msly	mslo	mshy	msho	ftly	ftlo	fthy	ftho	fsly	fslo	fshy	fsho
mtly	**29**	13	15	7	12	6	9	8	·	·	·	·	·	·	·	·
mtlo	22	**9**	14	9	8	11	13	4	·	·	·	·	·	·	·	·
mthy	6	18	**23**	24	4	3	6	17	·	·	·	·	·	·	·	·
mtho	8	24	16	**27**	2	13	13	42	·	·	·	·	·	·	·	·
msly	5	1	3	·	**35**	5	15	·	·	1	·	·	1	3	1	3
mslo	6	4	7	3	19	**18**	10	4	2	1	2	·	1	·	2	·
mshy	13	13	8	9	15	17	**13**	4	·	·	·	·	·	·	·	·
msho	11	18	15	21	3	25	8	**21**	·	1	·	·	·	·	·	·
ftly	·	·	·	·	·	·	·	·	**23**	4	16	6	13	5	16	1
ftlo	·	·	·	·	·	·	1	·	20	**14**	13	11	14	8	15	8
fthy	·	·	·	·	·	1	2	·	6	13	**9**	21	7	8	5	4
ftho	·	·	·	·	·	·	5	·	4	24	10	**22**	6	18	8	21
fsly	·	·	·	·	·	·	·	·	28	4	23	·	**21**	5	23	1
fslo	·	·	·	·	·	1	1	·	7	9	10	4	22	**19**	13	17
fshy	·	·	·	·	·	·	·	·	8	8	12	6	11	9	**13**	8
fsho	·	·	·	·	1	1	4	·	3	22	5	31	4	24	6	**36**

Table 6.1: Actual speaker properties compared to the classification by listeners in experiment 1 (in %). The following abbreviations are used: m = male, f = female, t = tall, s = small, h = heavy, l = light, y = young, o = old.

female voices and very often it is possible to differentiate between a young and an old voice. Therefore, for the identification of speaker properties the following was assumed:

1. The factor *sex* will be best identified.
2. Identification of *age* will follow that of *sex*.

There are two experiments with different audiovisual designs presented in this paper. Since in experiment 1 some problems were observed the listeners had when recognizing sex and age from the silhouettes, the following was assumed:

3. The listeners will be able to identify *sex* and *age* better on the stylised Mii figures from experiment 2 than from the silhouettes from experiment 1.

Ohala's frequency code (Ohala 1983, 1994) predicts that the height of pitch has an influence on the perception of the speaker's outer appearance.

4. Speakers with a low pitch will be perceived as tall and heavy.
5. Speakers with a high pitch will be perceived as small and thin.

6.4 Results

In the following the results of the two analyses are presented. Due to the large number of different results, these are grouped into thematic sections.

6.4.1 Identification of sex, height, weight and age

For the analysis of the identification of the different properties each classification was compared to the actual properties of each speaker. It is remarkable that in an overview the results of both experiments are extremely similar. The property of sex was recognized the best under both experimental conditions (97.7% in experiment 1 and 98.5% in experiment 2). The properties of height and age were correctly identified second best. From the silhouettes from experiment 1, height was recognized correctly by 60.4% and age by 60.0% of the listeners. In the second experiment, the factor age was identified somewhat better (60.9%) than height (58.5%). Surprisingly, the property of weight has been identified correctly by 52.2% in both experiments.

Furthermore, it was analysed whether the listeners could identify some speakers in all the categories together. Tables 6.1 and 6.2 show matrices of confusion in which the classification of the listeners is compared to the actual values of the speakers.

Chi-square tests showed significant values for the distinguishability of the factor levels of *sex* for both experiments ($\hat{\chi}^2 = 2777.310 > 10.828 = \chi^2_{(0.001;1;two\text{-}tailed)}$ *** in the first experiment and $\hat{\chi}^2 = 7344.407 > 10.828 = \chi^2_{(0.001;1;two\text{-}tailed)}$ *** in the second). That is, listeners could mostly identify the correct gender of the speakers. Tables 6.1 and 6.2 show, that the judgements of gender of some speakers were continuably confused. Small men were sometimes classified as female speakers. In cases were women were identified wrongly the factor size had no influence since tall as well as small women were classified as men.

In the classification of size, the listeners have been consistent. There were significant values for experiment 1 ($\hat{\chi}^2 = 129.959 > 10.828 = \chi^2_{(0.001;1;two\text{-}tailed)}$ ***) and experiment 2 ($\hat{\chi}^2 = 243.907 > 10.828 = \chi^2_{(0.001;1;two\text{-}tailed)}$ ***). Men as well as women were classified as tall or small irrespective of the actual values of the speakers.

The factor of weight showed similar significance ($\hat{\chi}^2 = 33.698 > 10.828 = \chi^2_{(0.001;1;two\text{-}tailed)}$ *** in experiment 1 and $\hat{\chi}^2 = 78.139 > 10.828 = \chi^2_{(0.001;1;two\text{-}tailed)}$ *** in experiment 2). The assignment of the voice stimuli to heavy and thin speakers was again very consistent.

The factor of age showed significance in both experiments ($\hat{\chi}^2 = 139.097 > 10.828 = \chi^2_{(0.001;1;two\text{-}tailed)}$ ***

Classification by listeners \ Actual groups	mtly	mtlm	mtlo	mthy	mthm	mtho	msly	mslm	mslo	mshy	mskm	msho	ftly	ftlm	ftlo	fthy	fthm	ftho	fsly	fslm	fslo	fshy	fshm	fsho
mtly	**25**	3	·	5	7	1	19	4	·	24	2	·	·	·	·	·	·	·	·	·	·	·	·	·
mtlm	21	**15**	8	14	11	10	9	8	1	7	5	·	·	·	·	·	·	·	·	·	·	·	·	·
mtlo	·	7	**10**	3	8	10	1	7	21	2	4	11	·	·	·	·	·	·	·	·	·	·	·	·
mthy	13	3	1	**11**	8	·	8	6	·	8	8	·	·	·	·	·	1	·	·	·	·	·	·	
mthm	6	16	20	22	**19**	18	4	3	1	1	6	29	·	·	·	·	·	·	·	·	·	·	·	·
mtho	·	8	18	5	13	**24**	·	1	5	·	1	18	·	·	·	·	·	·	·	·	·	·	·	·
msly	7	2	·	5	1	·	**32**	10	1	27	5	·	1	·	·	1	1	·	·	·	·	1	1	·
mslm	11	13	7	10	4	3	9	**10**	2	10	11	3	·	·	1	·	·	·	·	·	·	·	·	·
mslo	1	5	9	3	5	9	1	8	**33**	1	9	6	·	·	1	·	·	·	·	·	·	·	1	·
mshy	10	5	1	7	2	·	12	16	1	**15**	6	·	·	·	1	·	1	·	·	·	1	·	·	·
mshm	6	18	15	13	14	6	5	10	5	4	**11**	11	·	·	·	·	·	·	·	·	·	·	·	·
msho	·	6	12	3	9	19	·	7	21	·	11	**21**	·	·	1	·	·	·	·	·	1	·	·	·
ftly	·	·	·	·	·	·	·	·	·	·	·	·	**29**	2	·	17	·	·	17	3	·	12	·	·
ftlm	·	·	·	·	·	·	·	·	·	·	2	·	6	**22**	4	9	23	3	5	15	2	8	6	2
ftlo	·	·	·	·	·	·	·	·	1	1	2	·	·	6	**12**	2	4	23	·	10	10	·	19	27
fthy	·	·	·	·	·	·	·	·	·	1	·	·	12	3	·	**12**	2	·	5	5	·	6	1	·
fthm	·	·	·	·	·	·	·	2	·	·	3	·	3	18	11	5	**19**	5	2	10	4	7	10	4
ftho	·	·	·	·	·	·	·	3	2	·	4	·	·	3	23	1	13	**16**	·	3	18	·	5	6
fsly	·	·	·	·	·	·	·	·	·	·	·	·	19	1	·	19	·	·	**40**	5	1	25	1	·
fslm	·	·	·	·	·	·	·	·	·	1	·	·	7	18	2	7	6	·	12	**14**	3	11	8	2
fslo	·	·	·	·	·	·	·	·	1	·	1	·	·	6	13	1	4	24	1	12	**28**	2	18	31
fshy	·	·	·	·	·	·	·	·	·	·	·	·	18	3	·	17	2	·	14	5	·	**23**	1	1
fshm	·	·	·	·	·	·	·	1	1	·	4	·	4	13	15	9	18	·	4	13	9	6	**15**	4
fsho	·	·	·	·	·	·	·	2	5	·	5	·	·	3	17	·	8	29	·	5	23	·	15	**23**

Table 6.2: Actual speaker properties compared to the classification by listeners in experiment 2 (in %). The following abbreviations are used: m = male, f = female, t = tall, s = small, h = heavy, l = light, y = young, m = middle-aged, o = old.

in the first experiment and $\hat{\chi}^2 = 3575.913 > 18.467 = \chi^2_{(0.001;4;\textit{two-tailed})}$ *** in the second experiment).

A detailed analysis of the factor *age* in the second experiment shows that age was identified correctly relatively often (Tab. 6.3). Young speakers have been classified as old in only 0.3% of all cases and old speakers have been identified as young in only 1.7% of all cases.

6.4.2 Comparison of the sentences

Two syntactically and prosodically different sentences were used in the experiments. The interrogative sentence has a low f0-contour in the middle of the phrase and a raising pitch at the end (L*+H H%), while a late peak and a falling f0 can be expected in the declarative sentence (H*+L L%). In order to see whether this fact had an influence on the evaluation of the voice stimuli it was important to analyse the classifications of the sentences separately.

Figures 6.3 and 6.4 show the classifications of each participant in experiment 1 and 2. Listed here is how consistent the listeners were for both sentence types. That is, did they classify both sentences as e.g. a small, thin and young man or did they perceive the same voice in the second sentence as tall, heavy and old?

		actual age groups		
		young	middle-aged	old
evaluated age groups	young	**27.1**	5.7	0.3
	middle-aged	14.8	**20**	5.8
	old	1.7	10.9	**13.7**

Table 6.3: Accumulated classifications of the factor age in experiment 2. Readings in %.

It is obvious that the factor *sex* of both sentence types was classified most consistently by all subjects. In only very few cases, a voice stimulus was identified as male for one sentence and as female for the other. But looking at the results of experiment 1 (which used the silhouettes as the visual component) more closely, it turns out that participants *mck23* and *was21* were only consistent to 88% and 89%, respectively, in their classifications of sex. I.e. they classified 11–12% of the stimuli as opposed sexes for the different sentence types. These two listeners are seriously below the overall average of 97.2%.

In the second experiment, in which the voice stimuli were to be classified into Mii-figures, there were no participants who deviated so strongly from the general average. The fac-

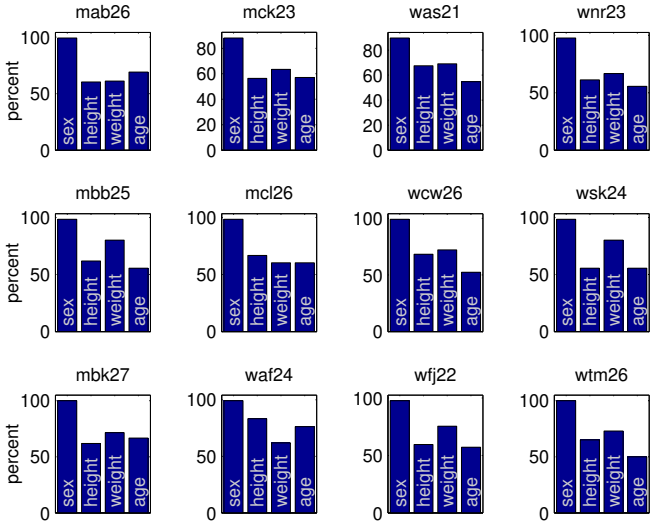

Fig. 6.3: Consistent classifications in experiment 1, readings in %.

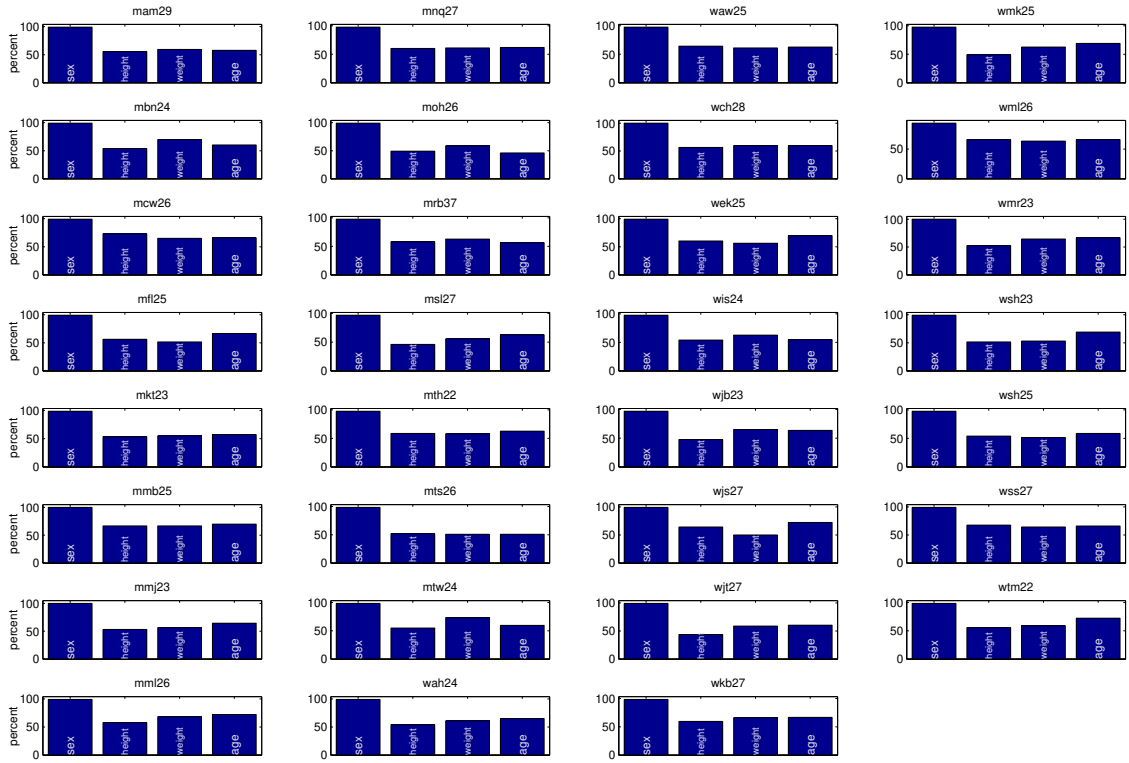

Fig. 6.4: Consistent classifications in experiment 2, readings in %.

tor of sex was classified consistently by an average of 98.0%.

The categories of height and weight were judged more consistently in experiment 1. At an average of 63.9% both sentences were classified to the same height and at an average of 69.6% to the same weight. In the second experiment the body size was judged the same by an average of 56.4% and weight by 60.4%.

Regarding the category of age, it was found that the sentences were classified more frequently in the wrong age groups in the first experiment than in the second. However, with an average of 59.2% in experiment 1 and 63.1% in experiment 2, this difference is only very small.

6.4.3 Acoustical analysis of f0 and physical appearance

The fundamental frequency was measured and compared with the four factors of sex, height, weight and age.

First, f0 was extracted from the speech signals with *get_f0* (Talkin 1995), the resulting f0 values were ordered and 0-values removed. In order to compare male with female voices it was decided to describe the fundamental frequency in semi tones (st). Based on a reference frequency of 50 Hz the values for f0 were logarithmised. Then, the highest 25% and lowest 25% of the values were used to estimate two mean values as a robust estimation of f0 range.

It isn't surprising that the factor of sex was highly significant ($p < 0.001 ***$) for all mean values of f0 since there are well-known differences in fundamental frequency for men and women. The data also showed a change of fundamental frequency for aging speakers. The factor age was highly

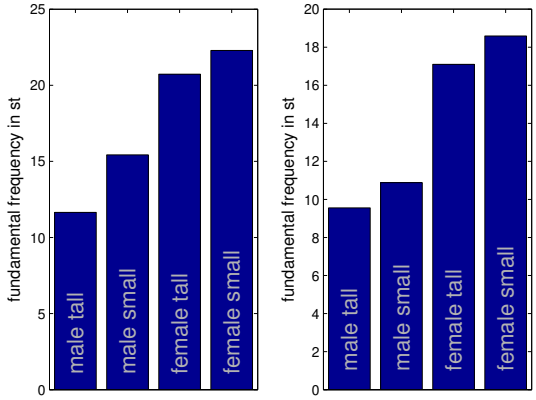

Fig. 6.5: Lower mean values for f0 in comparison to body size. Left: sentence 1 (interrogative), right: sentence 2 (declarative).

significant ($p = 0.001 ***$) for the lower f0 value and significant for the upper f0 value ($p = 0.008 **$). Furthermore, the factor of size had a highly significant influence on the fundamental frequency in the lower f0 value ($p < 0.001 ***$) and a significant influence in the upper one ($p = 0.009 **$). The taller a speaker is, the lower the pitch is. Fig. 6.5 shows the relationship between size and f0 in detail.

Tall men had a mean f0 of 11 st in the interrogative sentence; small men spoke considerably higher with a mean f0 of 15 st. In the second, declarative sentence, small men spoke only 1.4 st higher than tall men.

A similar tendency could be observed for women also, but here the difference was not as big as that for the men. In the interrogative the f0 for small women was 1.6 st higher and in the declarative the difference was 1.5 st.

The factor of weight had no significant influence on the fundamental frequency values.

6.4.4 Perception of f0 and its classification

Figures 6.6 and 6.7 show the influence of the lower mean f0 on the classification of size and weight in both experiments. The analysis was performed with the lower mean f0 value because it is not influenced as strongly by surrounding values as the upper mean f0 value, and it is closer to the value of indifference of a speaker.

First, the data of experiment 1 is analysed. For men a high f0 seems to be associated with small speakers. A mean f0 of 14.0 st was assigned to small and thin men. T-tests show significance for all other classifications. Voices perceived as belonging to tall and heavy men showed a mean f0 of 10.5 st. The f0 values of tall and heavy men highly significantly differ from the f0 values for small and thin men ($\hat{t} = 3.496 >= t(57;0.001) = 3.470***$). F-tests showed homogeneous variances. F0 values of men classified as tall and thin (lower mean f0 of 11.0 st) were significantly different from the values for small and thin men ($\hat{t} = 2.999 >= t(56;0.005) = 2.923**$). F0 values of men classified as small and heavy (lower mean f0 of 11.5 st) were significantly different from the classifications as small and thin men ($\hat{t} = 2.393 >= t(55;0.050) = 2.004*$).

Regarding the classification of the female speakers it turned out that high f0 values were associated rather with low weight than with body size. A mean lower f0 of 19.9 st was found with women classified as being tall and thin. A mean lower f0 of 21.6 st was observed with women classified as small and thin. The difference between these two classifications is only marginally significant ($\hat{t} = 1.707 >= t(67;0.100) = 1.668(*)$. The classifications as small and thin women differ significantly from the classifications as tall and heavy women who had mean values of 17.9 st ($\hat{t} = 3.604 >= t(62;0.001) = 3.454***$). The classifications as small and thin vs. small and heavy women (lower mean f0 of 18.2 st) are also significantly different ($\hat{t} = 3.234 >= t(64;0.002) = 3.223**$).

Although Figures 6.6 and 6.7 are rather similar, the data of experiment 2 is shown. For the male speakers there is only one significant difference ($\hat{t} = 2.654 >= t(60;0.020) = 2.390*$) in the classification as tall and heavy men (lower mean f0 of 10.3 st) and the small and thin men (lower mean f0 of 12.7 st). Men classified as tall and thin had a lower mean f0 of 11.3 st. Men classified as being small and heavy had 11.5 st.

For the female speakers there was a significant difference ($\hat{t} = 2.108 >= t(60;0.050) = 2.000*$) in the classifications as small and thin women (lower mean f0 of 20.7 st) vs. tall and heavy women (lower mean f0 of 18.3 st). Again, no other differences were statistically significant. A lower mean f0 of 19.2 st occurred for tall and thin women and 19.6 st for small and heavy women.

6.5 Discussion

This work aimed at examining whether listeners can perceive physical characteristics such as height and weight from simply hearing a person's voice. Furthermore, it should be tested whether perception tests with audiovisual components are beneficial for this kind of examination.

During the performance of the experiment with the shadow images, some problems came up which led to the decision to conduct a second, different experiment. However, the general trend of the results of both experiments turned out to be rather similar. This is astonishing because it was hypothesized that at least sex and age would be identified less correctly in the first experiment, since the silhouettes didn't give much information about sex and age. Therefore, hypothesis 3 (p. 37) cannot be verified since the factors of sex and age have not been identified better on the Mii avatars of experiment 2 than on the shadow images of experiment 1. Unfortunately, there were only 12 participants in experiment 1, a number which does not actually allow statements of significance. Maybe the difference in the general view would look differently, if more listeners had participated in experiment 1.

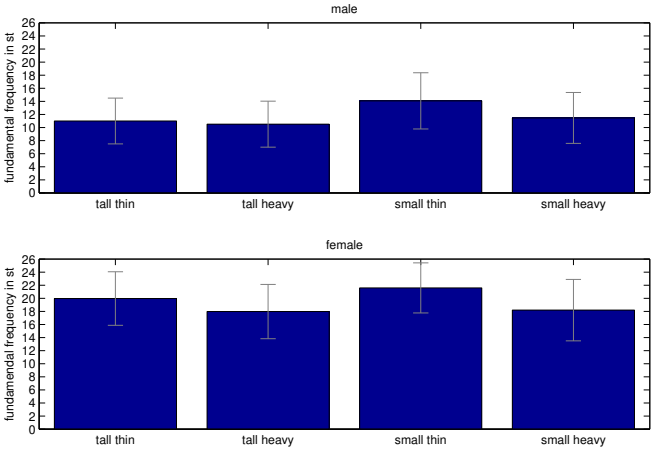

Fig. 6.6: Lower mean f0 and standard deviation as well as its classification of size and weight in experiment 1.

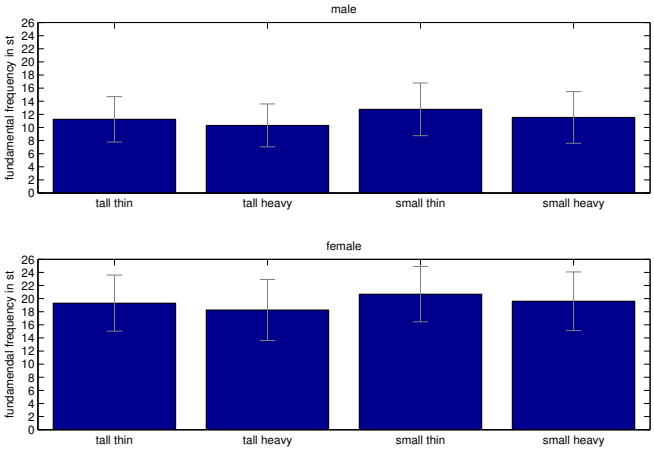

Fig. 6.7: Lower mean f0 and standard deviation as well as its classification of size and weight in experiment 2.

Then, hypothesis 3 could perhaps be verified, since already two of the 12 participants (presented in section 6.4.2 on p. 38) classified the two sentences rather inconsistently. They identified only 88% and 89% of the voice stimuli as equal sex for both sentences. That is, in 11–12% of all cases they decided that one sentence was spoken by a male speaker while they identified the other sentence (by the same speaker) as a female voice. Inconstancies like these did not occur in the 2nd experiment and we can hypothesize that these two listeners were so bad in their judgments because they could not recognize the sex of the shadow images. However, with only 12 participants in experiment 1 no statistical analysis is able to reveal significant differences between the two experiments to finally affirm hypothesis 3.

Hypothesis 1 is verified, since the factor of sex was recognized under both experimental conditions the best. The easiest task when imagining a speaker, it would seem, is deciding if he or she is male or female. However, there were also some speakers whose sex was identified wrongly by some listeners but the general view clearly shows that sex could be identified the most correctly.

Hypothesis 2, stating that identification of speaker age is more consistent than of speaker height, cannot be verified. The numbers of correct identifications for age and height were very similar. In experiment 1 height was correctly classified in 60.4% of all cases and age in 60.0%. In the second experiment the factor of age was identified somewhat better (60.9%) than the factor of size (58.5%). But these values are too similar to achieve statistical significance.

Here, an explanation is needed as to why age was identified so inaccurately. It was expected that effects of age would be perceived better than body characteristics like height and weight. The question arises, whether this lack of accuracy could result from wrong experimental conditions. Has the inclusion of a third age group in experiment 2 been a useful support for the participants? Or could it be that, because of the raised number of pictures, the confusion to select one picture grew? But regarding the detailed analysis of the factor of age for the second experiment in Tab. 6.3, we learned that there were very few confusions between young and old speakers. Most of the young speakers have been classified as young and most of the old speakers have been perceived as old. Taking a look at the middle-aged group, there was more confusion, though the majority was classified as middle-aged. However, there were also 10.9% who were classified as old. We can infer that listeners had no problems in identifying young and old speakers. The classification of middle-aged speakers brought up some confusion which happened in both experiments. In the first experiment the middle-aged speakers weren't identified as well, but here the listeners had to choose the categories young or old. It seems that the problem lies in the correct identification of middle-aged voices but not in its classification to pictures of middle-aged speakers.

Naturally, F0 was influenced by sex. But age and size also had an effect. Taller speakers spoke with lower pitch — a result that confirms the frequency code hypothesis of Ohala (1983; 1994). The height of pitch was not influenced by the property of weight, but with respect to size Ohala's hypothesis can be affirmed.

The fact that smaller speakers have higher pitches can also explain why the small men were partially identified as female speakers. Because of their higher f0, some listeners got the impression female speakers.

Remarkable is the fact that the f0 difference between tall and small speakers was bigger for men than for women in sentence 1 (interrogative). The small men spoke with an f0 which was 4 st higher than the f0 of tall men. In sentence 2 (declarative) this difference was only 1.4 st, even though small men spoke with a higher f0 here, too. The question comes up, whether these findings result from a deficient choice of test sentences. The interrogative sentence typically has a final rising intonation contour. But the second sentence also showed significant relations between body size and frequency and therefore it can be excluded that the significant results in sentence 1 are founded rather in the rising intonation than in physical characteristics.

But why is the f0 difference between tall and small speakers higher for men than for women? Ohala's frequency code would explain this circumstance with the evolutionary based necessity that men have to appear powerful and huge in order to be attractive to women, and later to protect the family. For women who had not to fight and dislodge antagonists there is no necessity to appear tall and heavy and therefore it may be, that the difference in f0 between small and tall women is not as big as that for men, even though there is also a significant relationship in size and f0 for female speakers.

Regarding the perception, the results show that listeners behave like the frequency code predicts. The lower the pitch, the taller the speaker is perceived. High-pitched voices were classified as belonging to smaller speakers. For the female speakers it seems, that listeners rely more heavily on the factor of weight. High-pitched voices were assigned to small and thin, as well as to tall and thin women. But as a high f0 was mainly assigned to small and thin men and women, hypotheses 4 and 5 (p. 37) can be affirmed.

It is possible that listeners are liable to prejudices in this case. There is a prototypical image of a speaker in our mind when we hear a low- or high-pitched voice. This fact is also used in comedy broadcasts where a comedian makes jokes by imitating other people. If there is a story about huge and heavy people, the performer will lower his voice, while he will raise the pitch when he imitates small people.

It seems that these prejudices find their explanation in the evolutionary based frequency code. In the stone age it was very important to know whether the counterpart was in a good mood or whether he was aggressive. Since fortitude and aggressiveness were expressed by outer appearance of hugeness it was important to look very tall, in order to protect the family. With the hight of the pitch a huge appearance could be intensified.

The results seem to confirm the frequency code hypothesis. But we found counterexamples to Ohala's theory. There are people with a big body size but also a high fundamen-

tal frequency, there are tall people who are physically as well as psychically not very strong, and there are people with a small outer appearance who have very much physical strength. Exceptions prove the rule?

From this research it can be concluded that there is no perfect relationship between the physical conditions of a person and the acoustic properties of their utterances. 2003 Tillmann & Pfitzinger unexpectedly found a weak relationship between the movements of the articulators when speaking and the perceptual local speech rate, as speakers probably develop very different strategies to realize utterances as desired (Kuehn & Moll 1976). One can probably assume that the influence of physical conditions is also limited and that speaking strategies and other factors are underestimated.

6.6 Outlook

Other factors influencing the voice could also be considered. Factors like smoking or high consumption of alcohol can influence the height of the pitch. The speech database contains information whether the speakers smoke, while consumption of alcohol was not asked for. The emotional state a speaker is in is also known to significantly influence f0 values (Amir et al. 2010; Bösel & Pfitzinger 2010). Factors like these could also have been included in the analyses in order to subtract outer influences from the physical properties.

Another influencing factor can be hormones like testosterone. Evans et al. (2006) showed that men with a high level of testosterone have muscular and tall bodies. Dabbs & Mallinger (1999) found that the vocal cords were longer and massier for men with high levels of testosterone. A consequence of this is a slower vibrating rate and thus a lower f0. For further analyses of male voices it would be possible to measure the level of testosterone and interrelate it with the values for height and weight as well as f0.

Further acoustic parameters e.g. formants should be analysed in future work. Fitch (1997) discovered a correlation between formant dispersion and size in primates. He suggests that formants are a more reliable measurement than f0, since f0 is influenced by many other properties.

The speech corpus consists of a large number of speakers, and the number of listeners in the 2nd experiment was also convenient for statistical analyses. But a number of 12 participants in experiment 1 is an absolute minimum, and maybe more participants would have produced other results.

Bibliography

Amir, N., Mixdorff, H., Amir, O., Rochman, D., Diamond, G. M., Pfitzinger, H. R., Levi-Isserlish, T. & Abramson, S. (2010). Unresolved anger: Prosodic analysis and classification of speech from a therapeutic setting. In: *Proc. of the 5th Int. Conf. on Speech Prosody*. Chicago.

Baumeister, B. & Willing, M. (2010). Ein neues Sprachkorpus zur Untersuchung von Beziehungen zwischen Stimme und Alter, Größe sowie Gewicht. Arbeitsberichte (AIPUK) 38, Inst. für Phonetik und digitale Sprachverarbeitung, Univ. Kiel, 21–28.

Bonaventura, M. (1935). Ausdruck der Persönlichkeit in der Sprechstimme und im Photogramm. *Archiv für die gesamte Psychologie 94*, 501–570.

Bösel, J. & Pfitzinger, H. R. (2010). Kulturübergreifende Unterschiede bei Akustik und Wahrnehmung authentischer Emotionen. Arbeitsberichte (AIPUK) 38, Inst. für Phonetik und digitale Sprachverarbeitung, Univ. Kiel, 63–76.

Dabbs, J. M. & Mallinger, A. (1999). High testosterone levels predict low voice pitch among men. *Personality and Individual Differences 27*, 801–804.

Darwin, C. (1972). *The expression of the emotions in man and animals*. Chicago University Press. Originally published 1872.

Evans, S., Neave, N. & Wakelin, D. (2006). Relationships between vocal characteristics and body size and shape in human males: An evolutionary explanation for a deep male voice. *Biological Psychology 72*, 160–163.

Fitch, W. T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *Journal of the Acoustical Society of America 102*, 1213–1222.

González, J. (2004). Formant frequencies and body size of speaker: a weak realationship in adult humans. *J. of Phonetics 32*, 277–287.

Höge, H., Kotnik, B., Kačič, Z. & Pfitzinger, H. R. (2006). Evaluation of pitch marking algorithms. In: *ITG-Fachbericht 192: Sprachkommunikation*. Berlin, Offenbach: VDE Verlag.

Kuehn, D. P. & Moll, K. L. (1976). A cineradiographic study of VC and CV articulatory velocities. *J. of Phonetics 4*, 303–320.

Künzel, H. J. (1989). How well does average fundamental frequency correlate with speaker height and weight? *Phonetica 46*, 117–125.

Lass, N. J. & Harvey, L. A. (1976). An investigation os speaker photograph identification. *Journal of the Acoustical Society of America 59*(5), 1232–1236.

Ohala, J. J. (1983). Cross-language use of pitch: An ethological view. *Phonetica 40*, 1–18.

Ohala, J. J. (1994). The frequency code underlies the sound-symbolic use of voice pitch. In: Hinton, L., Nichols, J. & Ohala, J. (Eds.), *Sound Symbolism*. Cambridge: Cambridge University Press, 325–347.

Pfitzinger, H. R. (2010). CoDIT: A combined discrimination/identification tool for speech perception experiments. Arbeitsberichte (AIPUK) 38, Inst. für Phonetik und digitale Sprachverarbeitung, Univ. Kiel, 29–34.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In: Kleijn, W. B. & Paliwal, K. K. (Eds.), *Speech coding and synthesis*. New York: Elsevier, Ch. 14, 495–518.

Tillmann, H. G. & Pfitzinger, H. R. (2003). Local speech rate: Relationships between articulation and speech acoustics. In: *Proc. of the 15th ICPhS*, vol. 3. Barcelona, 3177–3180.

van Dommelen, W. A. (1993). Speaker height and weight identification: a re-evaluation of some old data. *J. of Phonetics 21*, 337–341.

van Dommelen, W. A. & Moxness, B. H. (1995). Acoustic parameters in speaker height and weight identification: Sex-specific behaviour. *Language and Speech 38*(3), 267–287.

von Kriegstein, K., Warren, J. D., Ives, D. T., Patterson, R. D. & Griffiths, T. D. (2006). Processing the acoustic effect of size in speech sounds. *NeuroImage 32*, 368–375.