

# Average Speaking Pitch vs. Average Speaker Fundamental Frequency – Reliability, Homogeneity, And Self Report Of Listener Groups

*Sven Grawunder and Ines Bose\**

Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

\*Martin Luther University, Halle, Germany

grawunder@eva.mpg.de, bose@sprechwiss.uni-halle.de

## Abstract

Speaker fundamental frequency often stands as equivalent to the auditory measurement of average speaking (vocal) pitch. Previously found effects regarding higher estimations of average speaking pitch vs. average F0 for female voices have been fully replicated in an identification experiment involving 13 subjects. For female voices the average value of a given group of experts measuring average speaking pitch would be 2-3 semitones higher than the acoustically measured speaker fundamental frequency. Self reports of listeners' certainty of their judgements seem not to correspond at all with any given variance in the estimations and ratings. In a complementary discrimination experiment (16 subjects) this interval seems to indicate a threshold of the same size of 2 -3 semitones for pitch discrimination in speech.

## 1. Introduction

Speaker fundamental frequency (SFF) varies widely depending on individual physiological dispositions, actual emotional state but also with different speaking styles and text genres. In previous phonetic work [3, 15, 13] SFF and its auditory pendant speaking pitch have often been assumed to be identical or at least isomorphic. Auditory estimation of average speaking pitch (ASP) is best practice in clinical diagnostics of voice but also in voice and speech training, speech education, and forensic speaker evaluation [1, 3]. Thus ASP is considered as an indicator of subjective (tonal) perception, as well as an overall impression of a speaking voice. In this way we consider it as a relevant and weighty component in the perception of prosodic features. To be sure, the average vocal F0 values (as an acoustic measure) [1, 9] is also a fairly widely used parameter.

The perceptual (auditory) method of average speaking pitch measurement consists of a subjective evaluation by a trained and experienced listener. The specific process of perception, i.e. how the listener decides what to focus on over time in the voice of a speaker, is still opaque [2]. From reports by such evaluators, we know that the average speaking pitch is a more 'virtual' parameter. Listeners actively follow the pitch movement during the utterance (silently or half-loud); only then do they decide on the virtual center of the perceived tones [7]. The concept of such evaluation also includes the imagination of a modal value, i.e. the focussed tone would represent the pitch which occurs most frequently [1]. Experts are able to ignore loudness and textual features as well the nontypical sequences (e.g. creaky voice) often found at the beginning and end of an utterance (cf. [10]). The musical scale serves as referential base so that the listeners would focus on a musical note.

In a previous pilot study involving 6 expert listeners we had

observed an effect of average pitch identifications being approx. 2.5st higher than the average SFF [5]. The current study aims to replicate this effect with a larger group of listeners and with a more homogeneous background (training phase). Additionally we intend to exclude other factors, such as in-group variance or the arduousness of the perception task.

## 2. Identification Experiment: Auditory Measurement Of Average Speaking Pitch

### 2.1. Material

30 audio recordings of news broadcast from different German radio stations were investigated altogether. The samples have an average duration of about 80 sec. The recordings were made by means of a minidisc recorder (Sony MR7). The areal selection of (Federal German) radio stations was random, but with the requirement of equal proportions for gender and station type (public vs. private) [5].

### 2.2. Method

#### 2.2.1. Auditory measurement

For this study 13 expert listeners (10 female / 3 male, between 19 and 27 years of age, and of normal hearing) were asked individually to judge the samples in two separate runs (female/male speakers). Each of the experts (students of speech education) is musically trained and has experience in obtaining the average speaking pitch from a subject's running speech due to a training phase of 2 months. The experts were allowed to judge after repeated listening and with the aid of a piano or a tuning fork. The average speaking pitch estimations were provided as musical notes which have been afterwards transformed into numerical values (semitones, hertz). In addition to the perceived musical note the experts were also asked for a self report about the difficulty or ease of measurement per speaker (5 degree scale).

#### 2.2.2. Acoustic measurement

For detection of F0 we choose the PRAAT [6] standard setting of an autocorrelation algorithm with a time step window of 0.01sec (100 pitch values per second) and a voicing threshold of 0.7. In order to enable a comparison between the two measurements (Hz for acoustic, musical note for auditory) we have chosen to use the logarithmic semitone-scale as a common scale of reference, which has been shown to be appropriate for these tasks of equality ratings (cf. eg. [8]).

One semitone, the twelfth part of an octave corresponds to a frequency ratio of  $1 : \sqrt[12]{2}$ . The margin of 100Hz (G2↑) was taken as point of reference. In this way we can also use the

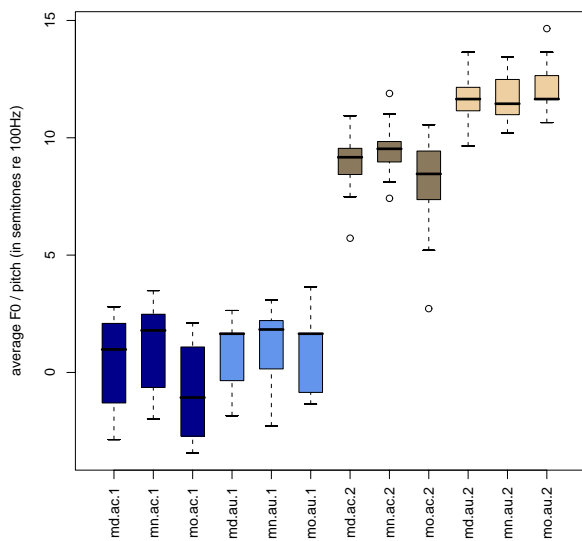


Figure 1: *Acoustic values of SFF vs. auditory values of ASP values by 19 subjects on 15 female and 15 male voices; 1=male, 2=female, ac=acoustic, au=auditory, mn=mean, md=median, mo=mode.*

semitone-scale for nominal representation of pitch tones.

### 2.3. Results (Identification experiment)

#### 2.3.1. Comparison of auditory and acoustic estimations

A characterization regarding the relation between acoustically and auditorily measured values was achieved by comparing (simple and averaged) mean and median values of SFF with means of mean and median ASP values. For male voices total mean of SFF correlates strongly with the mean value of pitch evaluations (cf. Figure 1). If we consider differences of means, we find the best convergence of median of SFF together with mean and median of pitch in male voices. However, for female voices the experts tended to measure average speaking pitch 2-3st higher than the acoustically measured SFF. Following Baken & Orlikoff [1] we also considered the mode value of SFF as a possible candidate of best correspondence. As for ASP, the mode value refers to the most frequent tone value within the group of listeners given for one sample. Surprisingly, the acoustic mode value shows greater distance from auditory measurements than mean and median even for male voice.

#### 2.3.2. Inter-listener homogeneity

In order to assess the homogeneity of pitch evaluations per gender group of speakers, a non-parametric variance analysis was adopted. After determining ranks for the individual values separately for experts and stimulus (speaker gender), the middle ranks per objects (values per speaker) were determined. Then the mean absolute difference between individual ranks and middle rank per object were compiled. In this way the null hypothesis was formulated that if the ratings of experts is homogeneous, the mean absolute differences in the two (gender) groups should also be homogenous. This was then tested by means of

the Mann-Whitney U-Test. The results ( $U = 89.0, N_{male} = N_{female} = 15, exactP = 0.345$ ) show homogeneous expert ratings for both genders of speakers, in so far as they show the same in-group variability.

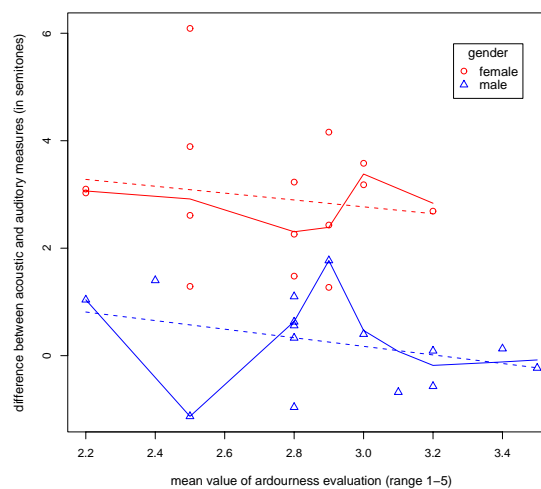


Figure 2: *scatterplot of self report ratings about certainty vs. difference between median values of ASP and SFF for 13 raters / 30 samples. Rating 1(= "very easy") to 5(= "very difficult").*

#### 2.3.3. Self report regarding arduousness

The ratings for self reports about judgement certainty range only between 1 and 4, i.e. excluding 5 as the most 'uncertain' ("difficult") degree. Figure 2 shows a xy-plot of acoustic-to-auditory-measurement deviation. Although the distribution seems to suggest a higher variance in ratings of male voices there is no significant difference between these and the female voices. This indicates that the deviation with female voices is not associated with a greater uncertainty in listeners' judgments (Figure 2).

#### 2.3.4. Listener consistency & inter-listener reliability

The applied musical scale implies the adoption of reliability procedures regarding (adjusted) interval based rating scales. According to Wirtz & Caspar [14] such scales can be correlated by means of the product-moment correlation and represented as a matrix in a correlogram [4](Figure 3). Reliability needs to be tested with regard to group mean (auditory mean/median), with regard to the acoustic measure (acoustic mean/median) as external point of reference, as well as the relation of reliability between individual listeners (inter-listener). Figure 3 depicts the correlations for all pitch estimations of female voices. The correlation between two individual listeners is very high in some cases (R & J), in other cases (G & Q) there is no correlation, and in some cases (G & M) listeners were estimating contrarily so that we see a negative correlation. But the majority of pitch estimations clusters positively either around the acoustic median or the (auditory) median.

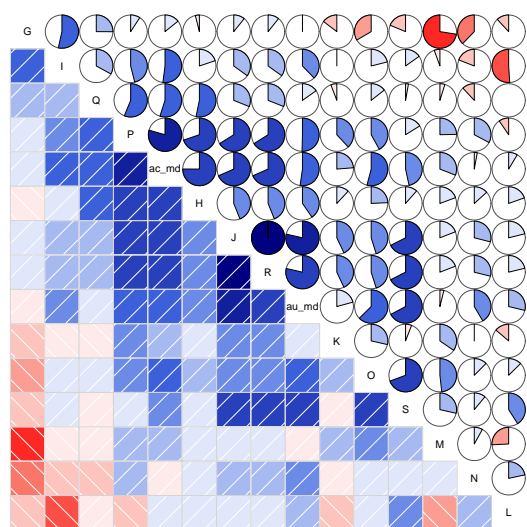


Figure 3: *Correlogram of acoustic measures and auditory average measures (median values) as well as individual auditory measures of 13 listeners for 15 female voices. The degrees of correlation (negative in red, positive in blue) are represented either as degrees of shading (lower panel) or completeness of the pie (upper panel). The letters (G to S) represent the estimations per listener (ac\_md=acoustic median; au\_md=auditory median).*

## 2.4. Discussion

The results regarding the group of 13 subjects fully replicate the previous results involving only 6 experts. However, the current larger corpus including the ASP estimations of all 19 subjects shows more homogenous measures than the previous one [5]. The distribution of certainty ratings shows no clear relation to the distance of ASP estimation and SFF. We would therefore exclude a lack of certainty ('guessing') as an explanation for the found mismatch. Hence, we have to assume a universal effect in the perception of speaking pitch of female speakers. This effect would lead to a deviant estimation of average speaking pitch versus speaking fundamental frequency in the size of 2-3 semitones.

## 3. Discrimination Experiment: Average Speaking Pitch vs. Reference Tone

### 3.1. Hypotheses

In addition to the identification experiment described above a discrimination experiment was designed. Here we proceeded under the following assumptions: (1) listeners should have more difficulty evaluating female voices than male voices, (2) the accuracy of discrimination of female voices should not correlate with the subjective confidence (certainty) of judgements, (3) an upward shifted reference tone for female voices should be rated as equal to ASP, and to the contrary (4) an upward shifted reference tone for male voices would be perceived as too high. The identification experiment also determined the interval of 2.5 semitones as a probable threshold. Therefore a null hypothesis

would deny any effect within the given range of  $\pm 2.5$  semitones. Otherwise one would need to assume (as an alternative null hypothesis) a match of shifted reference tone (comparison tone) and equality discrimination ratings.

### 3.2. Material & Method

The same data corpus of radio news presentations as in the previous experiment served as stimuli. One male and one female sample were excluded and the maximum time per sample was reduced to 25sec. Thus, 28 samples (14 male / 14 female speakers) were presented as stimuli. The samples were opposed to reference tones which had been partially shifted ( $-2.5st$ ;  $\pm 0st$ ;  $+2.5st$ ).

Here the acoustic median F0 (see 2.3.1) was taken as a basis for each sample. All stimuli were presented in random order using the multiple forced choice environment of PRAAT. A speaker sample was interrupted by a 4sec break after about two thirds of the sample length; at the end of the sample the shifted ping tone was played after another pause of 4sec. Play back repetitions were allowed. The subjects were of the same type of subjects (with even partial overlaps) as for the identification experiment. In the course of the experiment the participants were asked to state whether the perceived average speaking pitch of the presented stimulus vis-à-vis the following reference tone was equal, lower, higher or "dissimilar" (unspecific unequal). Subsequently the responses were processed as equal, unequal-higher, unequal-lower, unequal-dissimilar (see Figure 4). Additionally we asked for a "confidence" rating (5 degree scale) as a self report about the raters' certainty of their judgements.

### 3.3. Results (Discrimination experiment)

#### 3.3.1. Equality ratings

We report the responses only in bipolar opposition (equal vs. unequal) since the individual ratings show too ambiguous results (cf. Figure 3). In cases of equal reference tone and median SFF, half of the listeners rated the ASP as equal (male speakers 47.5%; female speakers 52.5%) the other half as unequal to the tone. Similarly, in cases of downward shifted reference tones, listeners tended to perceive tone and pitch as equal (male 48.4%; female 46.8%). In contrast, in cases of upward shifted reference tones a clear majority of listeners rated the stimuli as unequal (male 68.8%; female 72.5%).

#### 3.3.2. Gender effects

Apparently no clear-cut gender asymmetry in pitch estimations of male and female voices could be observed. Even if we would exclude "dissimilar" ratings, this would not change the tendencies in the individual categories of the experiment.

### 3.4. Discussion

The null hypothesis can be rejected, but only on the basis of paradox results. We would have expected such ambiguous results specifically for a constellation where the effect of perceptual shift of female voices would collide with a reference tone shift of the same size.

The expected effects regarding a "neutralization" of perceptual shifted tones (see hypotheses in 3.1) could not be observed. Only under circumstances of high shifted reference tones did a majority of listeners rate the average pitch as unequal (male 38.5%; female 57.5% as lower). In general, the raters seemed to be excessively demanded, given also the high

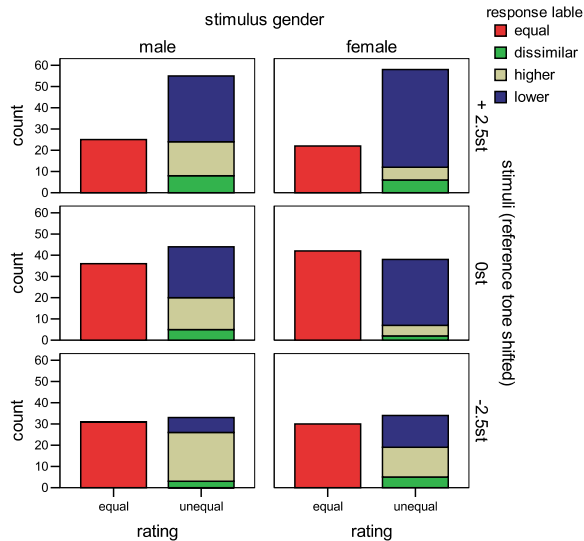


Figure 4: Ratings of 16 subjects' discrimination regarding perceived the ASP and a given reference tone (RT); 80 cases of 2.5st upward shifted RT (top), 80 cases of 0st shifted RT (mid), and 64 cases of 2.5st downward shifted RT (bottom) with median F0 as a basis

rates of “equal”-judgements. Both would indicate that the chosen interval of 2.5st makes a threshold for discrimination of such complex harmonic sounds as the human voice. This of course needs to be tested and further investigated since the just notable difference (JND) of average speaking pitch seems to be situated outside of the chosen shifting interval for down-shifted reference tones but within for upward-shifted tones.

#### 4. General Discussion & Conclusion

In summary, we argue for a distinction between auditory and acoustically measured average speaking pitch (F0). According to our data even the acoustic mode value does not correspond with its auditory pendants, which has been assumed before [3]. For a non-trained listener we assume here an overall impression of speaking pitch. In this way we consider the further investigation also relevant for exploration of perceptual biases towards pitch judgements of human speech, including those for intonation. Hence, further investigations need to check systematically for such influence factors as pitch variance (range, micro-prosodic variation; cf. [11, 12]) within the given sample and as spectral voice components [10]. As a practical issue and result we suggest a development of ASP-SFF distance interval thresholds in order to access in-group consistency and inter-listener reliability in test situations.

Although we are able to predict tendencies for listener group measurements, a satisfying explanation for these repeatedly attested effects seems to be difficult. Related experiments (cf. [10]) have so far been done by means of simple and complex sinus tones. Although a JND for tone distinction or discrimination of sinus tones would be in the given range (160-200Hz) between 1 and 2 hertz (0.5%) [10], the current results for voices would suggest values between 15 and 20 hertz. A previous study focussing on average speaking pitch had already

determined a tolerance interval of 2 semitones for individual average pitch estimations of both genders [7]. In agreement with this we would prefer to conceptualize average speaking pitch as a range instead of a tone.

See <<http://www.fonetik.de/sff-asp/>> for more details of the analysis.

#### 5. Acknowledgements

We thank all our subjects for their participation in the two experiments. We are grateful to the students of a seminar who engaged in useful discussions and helped to transfer the raw data. We are indebted to Roger Mundry (MPI EvA) for discussing the statistics of homogeneity and interrater reliability, and to Brigitte Pakendorf (MPI EvA) for checking the manuscript. We are also grateful to an anonymous reviewer for various valuable comments. All remaining mistakes and errors are of course to our full responsibility.

#### 6. References

- [1] Baken, R. J., Orlikoff, R. F., 2000. *Clinical measurement of speech and voice*. San Diego: Singular.
- [2] Carlson, R., Elenius, K., & Swerts, M., 2004. Perceptual Judgments of Pitch Range. *Speech Prosody 2004*, International Conference ISCA. Nara, Japan.
- [3] Braun, A., 1994. Sprechstimmlage und Muttersprache. *Göschel (ed.): Z. f. Dial. & Ling., 1, LXI: 170–178*.
- [4] Friendly, M., 2002. Corrgrams: Exploratory Displays for Correlation Matrices. *The Am. Statistic., 56(4): 316–325*.
- [5] Grawunder, S., Bose, I., Hertha, B., Trauselt, F., Anders, L. C., 2006. Perceptive and acoustic measurement of average speaking pitch of female and male speakers in German radio news. *Proc. ICSLP*, Pittsburgh: 885–888.
- [6] Boersma, P., Weenink, D., 2007. *Praat: doing phonetics by computer (Version 4.5.16)* [Computer program].
- [7] Kurka, E., Fredrich, R.-B., 1968. Zur Bestimmung des physiologischen Hauptsprechtons. *Wiss. Z. Univ. Halle, XVII'68 G, Vol. 5, 45–52*.
- [8] Nolan, F., 2003. Intonational equivalence: an experimental evaluation of pitch scales. *Proc. 15th ICPhS, 771–774*.
- [9] Schultz-Coulon, H.J., 1975. Bestimmung u. Beurteilung d. individ. mittl. Sprechstimmlage. *Fol. Phon., 27: 375–386*.
- [10] Terhardt, E., 1998. *Akustische Kommunikation*. Berlin [u.a.]: Springer.
- [11] Traunmüller, H., Eriksson, A., 1995a. The frequency range of the voice fundamental in the speech of male and female adults. 12-Dec-05 retrieved from <<http://www.ling.su.se/staff/hartmut/aktupub.htm>>
- [12] Traunmüller, H., Eriksson, A., 1995b. The perceptual evaluation of F0 excursions in speech as evidenced in liveliness estimations. *J Acoust Soc Am, 97(3): 1905–1915*.
- [13] Vaissière, J., 2005. The Perception of Intonation. In D. Pisoni and R. Remez, editors, *The handbook of speech perception*, 236–263. Oxford: Blackwell Publishing.
- [14] Wirtz, M.A., Caspar, F., 2002. *Beurteilerübereinstimmung und Beurteilerreliabilität*. Göttingen [u.a.]: Hogrefe.
- [15] Wittlinger, I., & Sendlmeier, W.F., 2005. Stimme und Sprechweise erfolgreicher Frauen. In W. Sendlmeier, editor, *Sprechwirkung - Sprechstile in Funk und Fernsehen*, 71–119. Berlin: Logos.