

## **Control of voice fundamental frequency in speaking versus singing**

**Ulrich Natke, Thomas M. Donath, and Karl Th. Kalveram**

Institute of Experimental Psychology, Heinrich-Heine-University Düsseldorf, Germany

**Abstract.** In order to investigate control of voice fundamental frequency ( $F_0$ ) in speaking and singing, twenty-four adults had to utter the nonsense word [ta:tatas] repeatedly, while in selected trials their auditory feedback was frequency-shifted by 100 cents downwards. In the speaking condition the target speech rate and prosodic pattern were indicated by a rhythmic sequence made of white noise. In the singing condition the sequence consisted of piano notes, and subjects were instructed to match the pitch of the notes. In both conditions a response in voice  $F_0$  begins with a latency of about 150 ms. As predicted, response magnitude is greater in the singing condition (66 cents) than in the speaking condition (47 cents). Furthermore the singing condition seems to prolong the after-effect which is a continuation of the response in trials after the frequency shift. In the singing condition, response magnitude and the ability to match the target  $F_0$  correlate significantly. Results support the view that in speaking voice  $F_0$  is monitored mainly supra-segmentally and controlled less tightly than in singing.

## I. INTRODUCTION

Auditory control of voice fundamental frequency ( $F_0$ ) has been investigated with the frequency shift paradigm in many studies before (an overview can be found in Donath, Natke, and Kalveram, 2002). From the very beginning, continuous vocalization has been studied (e.g. Elman, 1981; Burnett, Senner and Larson, 1997; Larson, 1998). In this approach, subjects usually produced a vowel for a duration of about 5 seconds. After a randomly determined period, auditory feedback provided via headphones was shifted in frequency. Typically subjects reacted with a change in voice  $F_0$  in the opposite direction of the frequency shift (“opposing response”), which indicates a closed-loop negative feedback system compensating deviations between intended and perceived pitch. However, inter-individual differences in responses were large and responses in the direction of the frequency shift (“following responses”) were also observed. There is some evidence that opposing responses occur with a short latency of 100-150 ms, whereas following responses have a longer latency of 250-600 ms (Larson, 1998). The first response therefore indicates a negative feedback system stabilizing voice  $F_0$  automatically. The second response may reflect a voluntary mechanism, which adjusts voice  $F_0$  to match an (supposed) external reference.

In order to investigate voice  $F_0$  control in speaking, which is characterized by rapidly repeated onsets and offsets of phonation, Natke and Kalveram (2001) and Donath *et al.* (2002) utilized a paradigm which was used previously in investigating the control of vowel duration with delayed auditory feedback (e.g. Kalveram, 1989; Jäncke, 1991; Natke, 1999). In this paradigm, subjects utter a nonsense word repeatedly and auditory feedback is modified in randomly selected trials. These trials are compared to preceding trials in which auditory feedback is not altered. The prosodic pattern of the nonsense word is varied, so that effects of feedback manipulation on short unstressed and long stressed syllables in different positions within the word can be investigated.

The two studies cited above show that in speaking only opposing responses occur with a latency of about 160 ms due to frequency-shifted auditory feedback. Because of this latency, voice  $F_0$  cannot be controlled auditorily within short syllables and responses in long stressed syllables come into effect late. The response takes place over successive syllables, even when phonation stops between them. Furthermore an after-effect in trials after termination of frequency shift is found indicating that the response persists for several seconds. Therefore the system is able to adjust voice  $F_0$  in later syllables by monitoring current syllables. The auditory-vocal system controlling voice  $F_0$  therefore seems to operate mainly at the supra-segmental level, not at the syllabic level.

In speaking, response magnitudes of about 50 cents were found, although a complete compensation would have required a change of 100 cents in voice  $F_0$ . These magnitudes correspond to magnitudes found in continuous vocalization (mostly 30-60 cents using frequency shifts of 100-600 cents). Thus there seems to be a limiting property of the audio-vocal system, which prevents responses from exceeding a certain magnitude. Natke and Kalveram (2001) and Donath *et al.* (2001) explain these findings by pointing out that for the comprehension of speech sound duration, formants and formant transitions seem to be more important than voice  $F_0$ , at least in languages like English or German. Control of voice  $F_0$  at a supra-segmental level but not of absolute pitch within syllables is important to encode non-verbal information. Therefore, a tight control of voice  $F_0$  is unnecessary in speaking.

However in singing, tight control of voice  $F_0$  seems preferable. Whereas in speaking continuous changes of voice  $F_0$  occur, singing is usually characterized by matching absolute values of pitch mainly stepwise, corresponding to musical notes (apart from exceptions such as performing a glissando, singing with vibrato or singing blue notes making up the blues

scale, where pitch is not fixed precisely but varies). In speaking an external reference for voice  $F_0$  does not exist, but in singing for instance in a choir or accompanied by musical instruments, an external reference is provided. A deviation between the  $F_0$  of the own voice and the external reference must be compensated. This is easily achieved by trained singers; for example trained singers can match a reference tone of 440 Hz with an accuracy of less than 1 Hz (Sundberg, 1987). Also solo singing requires some form of a reference, so that individual notes of scales can be sung with the correct pitch. Even untrained singers are usually aware of when they do not produce the right pitch, resulting in the characteristic “bad” glissando while trying to match their voice  $F_0$  with the target  $F_0$ . This reference must be represented internally. Consequently, because a reference for voice  $F_0$  exists in singing, a frequency shift should be compensated almost completely. Evidence for that is reported by Burnett, Senner and Larson (1997) who had trained singers sing musical scales. In some subjects a complete compensation of unanticipated frequency shifts of 100 cents was observed. However, no group data is reported. In another study (Parlitz and Bangert, 1999) subjects sang straight notes and complete compensations of frequency shifts were observed, too.

In summary, it is suggested that in speaking an internal reference for pitch plays a subordinate role for controlling voice  $F_0$  (leading to the limiting property), whereas the audio-vocal system is able to operate very effectively in controlling voice  $F_0$ , if a precise reference is provided as in singing. Therefore the response to frequency-shifted auditory feedback should be greater in singing than in speaking. The present study was designed to test this hypothesis.

## II. METHOD

### A. Subjects

Twenty-four adult German native speakers (9 women, 15 men) between 21 and 33 years of age participated in this study ( $M=25.9$  years,  $S.D.=3.92$  years). None of the participating subjects showed a hearing deficit of more than 20 dB HL (audiometric test: Hortmann DA 323, Neckartenzlingen, Germany). No subject reported a current speech or language disorder.

### B. Apparatus

Voice frequency was shifted using a commercial device (DFS 404, Casa Futura Technologies, Boulder, Colorado, USA), which works on a digital basis (sampling frequency: 32 kHz, sampling resolution: 14 bits). The device had been modified by installing a relay to enable remote-switching between non-altered auditory feedback and frequency-shifted auditory feedback. Auditory feedback was provided through sealed headphones of a headset (Blackhawk, DSP 5DX, Flightcom, Portland, Oregon, USA), which attenuate air-conducted sound by approximately 24 dB SPL. Feedback volume was calibrated in a way that a sinusoidal tone of 440 Hz with 75 dB (A) at the microphone led to a feedback volume of 70 dB (A) in the headphones. Subjects perceived this volume as a normal feedback volume. In order to mask bone conduction, during the whole experiment low-pass-filtered white noise ( $f_c=900$  Hz) was binaurally added at an intensity of 65 dB (A). This level was low enough to hear the own voice clearly and high enough to mask bone conduction effectively.

In order to determine voice  $F_0$ , the vibrations of the vocal folds were recorded directly with an electroglottograph (EGG; Laryngograph, Kay Elemetrics, Pine Brook, New Jersey, USA). Control of the experiment and data acquisition was automatized by a commercial

personal computer with a soundcard. The computer switched the frequency shift device between non-altered and frequency-shifted auditory feedback and digitized the EGG signal with a sampling rate of 11,025 Hz and a sampling resolution of 16 bits.

### **C. Procedure**

Subjects had to utter the nonsense word ['ta:tatas] with a speech rate indicated by a rhythmic sequence presented via headphones before speaking. The first long-stressed syllable was represented by a sound with a duration of 400 ms. Each of the following two sounds represented the unstressed syllables and were 200 ms long. Sounds were separated by pauses of 40 ms. The sequence was presented twice, afterwards 3.5 s were left blank to speak the word. The sequence was presented automatically while the subject sat alone in a sound isolated chamber.

Two conditions were realized which differed regarding the sound sequence and the instruction. In the “speaking condition”, the sequence consisted of white noise. Therefore, no kind of reference pitch for voice  $F_0$  was provided. Subjects were asked to speak clearly and with normal volume, but no specific instruction regarding the fundamental frequency was given. In the “singing condition”, the sequence consisted of piano notes. Subjects were instructed to match the pitch of the piano notes while singing the nonsense word. However, the piano notes were presented before subjects’ utterance, not during it. The fundamental frequency of the piano notes was varied based on the sex of the subjects: females: 233 Hz, males: 123 Hz. These frequencies were chosen because they are very close to the average voice  $F_0$  of women and men (reference) and resemble the musical notes a#3 and b2. In neither condition subjects were informed that the feedback would be modified from time to time.

In a training phase preceding each of the two phases of the experiment, subjects had to utter or sing the nonsense word simultaneously, while the target white noise or target piano notes, respectively, were presented in a loop. This was done until they produced the word correctly at least five times in succession, as judged by the experimenter. Generally, subjects reached this criterion within ten trials.

The experimental procedure consisted of 30 trials for the speaking condition and 30 trials for the singing condition. In 20% of the trials frequency was shifted downwards by 100 cents (one semitone), resulting in 6 frequency-shifted trials per condition. The frequency shift was activated at the beginning of the rhythmic sequence, in a way that auditory feedback, but not the sequence sounds themselves were frequency-shifted from the very beginning of subjects’ utterances. Subjects then produced the entire word with frequency-shifted auditory feedback. Frequency shift was deactivated before the next trial began. Trials with frequency shift were randomly selected with the limitations that at least two trials with non-altered auditory feedback had to precede a trial with frequency shift and that the last trial of an experimental phase was always conducted with non-altered auditory feedback. Therefore it was guaranteed that before each trial with frequency shift there was always a trial with non-altered feedback, which did not immediately follow frequency shift. This way trials immediately before frequency shift could be used as the reference to calculate deviations in voice  $F_0$ , while trials immediately after frequency shift could be used to investigate the after effect. The sequence of the two conditions was randomized as well.

### **D. Data Analysis**

Data analysis was almost identical to Donath, Natke, and Kalveram (2002), in which the method is described in more detail. Based on the EGG-signal, momentary voice  $F_0$  and vowel duration were calculated. Contours of  $F_0$  were obtained for vowels of the first two

syllables of the nonsense word, the first being long stressed and the second being unstressed. First  $F_0$  contours in a fixed resolution of 0.1 ms were determined. This was achieved by linear interpolation of the momentary frequency of each vocal fold period. In some cases there were irregular high-frequency fluctuations in voice  $F_0$  contours due to problems with analyzing EGG-signals with a poor signal-to-noise ratio. Therefore voice  $F_0$  contours were smoothed with a moving average over 10 ms. Phonation onset was defined as the first closing instant detected in the EGG-signal. Mean voice  $F_0$  contours for individual subjects were truncated at the end according to the shortest vowel duration produced by this subject. Thus, all portions of the contour were based on the total number of trials. To avoid  $F_0$  contours being truncated excessively based on unusually short vowels, trials in which vowel duration was 25% shorter compared to all other trials were discarded. A total of 30 trials (speaking condition: 18, singing condition: 12) were discarded because of insufficient length or artifacts, resulting in a total of 834 valid trials. Despite the training, one subject produced the second syllable with an unusual vowel duration shorter than 20 ms during the actual experiment, so the data of this subject were excluded from analysis of the second syllable.

Mean voice  $F_0$  contours of vowels in words before (PRE), during (FAF), and after (POST) trials with frequency-shifted auditory feedback were calculated. Remaining trials were discarded. PRE-trials were used as reference trials, to which FAF- and POST-trials were compared for each subject. In order to eliminate inter-individual variance due to the subjects' habitual voice  $F_0$  (especially due to the sex of the subjects), for each data point the difference in cents was calculated between  $F_0$  contours in PRE- and FAF-trials as well as in PRE- and POST-trials. Resulting contours therefore reflect the deviation from normal production of voice  $F_0$  during frequency shift and after its termination. Finally, individual subjects' contours were averaged to obtain the group data.

Based on previous findings (see Donath *et al.*, 2002) it is hypothesized that voice  $F_0$  is higher in the final portion of the first long stressed syllable and the initial and final portion of the second unstressed syllable in FAF-trials. Due to the after-effect voice  $F_0$  should also be higher in the initial portion of the first syllable in POST-trials. The initial portion of syllables was defined as the interval 25-50 ms after vowel onset and the final portion was defined as the interval 200-225 ms in the first long stressed syllable and 75-100 ms in the second short unstressed syllable. These intervals were chosen to be as late as possible and yet featuring voice  $F_0$  data for most subjects. Means for these intervals were calculated for each subject and one-tailed Wilcoxon-signed-rank-tests were calculated to test for responses. A significance level of  $\alpha = 5\%$  was chosen and corrected according to Bonferroni to  $\alpha' = \alpha/8 = .00625$ . For intervals for which no differences were assumed, additional  $p$ -values were calculated two-tailed and interpreted as measures of effect. Response latencies for syllables, in which changes in voice  $F_0$  were indicated by differences between the initial and final portion, were determined using the Castellan change-point test (Siegel and Castellan, 1988).

Furthermore it is hypothesized that the response in the singing condition is greater than in the speaking condition. The difference between the speaking and singing condition was tested by comparing the response magnitudes in the interval 25-75 ms of the second syllable because in this interval the response is at its maximum and relatively stable as found by Donath *et al.* (2002).

The same interval was chosen in order to investigate whether a relationship exists between the ability to accurately match a note which is provided externally and the magnitude of compensation under frequency shift. For each subject the response magnitude was calculated as the deviation of voice  $F_0$  contours of FAF-trials from PRE-trials in the interval 25-75 ms of the second syllable in the speaking as well as in the singing condition. The same interval in PRE-trials in the singing condition was chosen to calculate the

deviation of voice  $F_0$  from the fundamental frequency of the piano notes (target  $F_0$ ), which were presented before the utterance to provide the target  $F_0$ . The absolute value of this deviation indicates the accuracy of note matching. Spearman’s rho correlation was calculated for the response magnitude and the absolute deviation from the target  $F_0$ .

### III. RESULTS

#### A. Response latency and response magnitude during frequency shift

Table 1 reports the vowel duration of the first (=long stressed) syllable and the second (=unstressed) syllable of the nonsense word [‘ta:tatas] averaged over PRE-, FAF- and POST-trials. The vowel duration of the stressed syllable was nearly three times longer than the vowel duration of the unstressed syllable. Whereas in the unstressed syllable the vowel duration in the speaking and in the singing condition do not differ ( $p_{2\text{-tailed}} = .280$ ,  $Z = -1.080$ ,  $n = 24$ ), the vowel duration of the long stressed syllable is approximately 29 ms longer in the singing than in the speaking condition ( $p_{2\text{-tailed}} = .002$ ,  $Z = -3.148$ ,  $n = 24$ ). Therefore the prolongation of vowels in singing, which sometimes is obvious, was found in the present experiment, however restricted to long stressed syllables.

**Tab. 1:** Mean vowel duration in speaking vs. singing of the first (=long stressed) syllable and the second (=unstressed) syllable of the nonsense word [‘ta:tatas] (standard deviation in brackets). Vowel duration was averaged over all PRE-, FAF- and POST-trials.

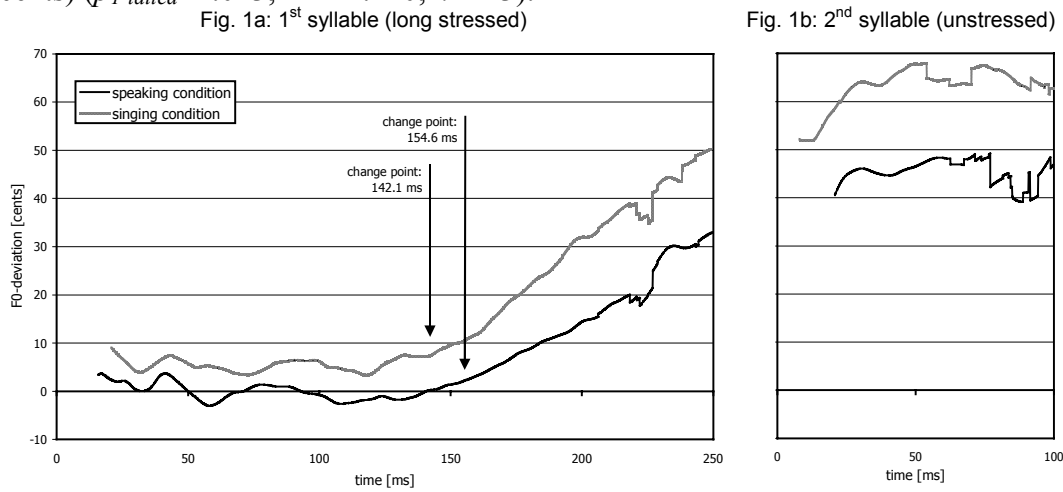
condition	vowel duration (S.D.) [ms]	
	1 <sup>st</sup> syllable (long stressed)	2 <sup>nd</sup> syllable (unstressed)
speaking	249.5 (33.5)	89.2 (22.9)
singing	278.4 (36.5)	93.4 (27.5)

Figure 1 shows the deviation of voice  $F_0$  contours of FAF-trials from PRE-trials. In all graphs, 0 ms corresponds to the first detected glottal closing, which defines onset of phonation. The solid line refers to the speaking condition, the dotted line refers to the singing condition.

In the interval 25-50 ms after vowel onset (Fig. 1a) mean voice  $F_0$  of the first syllable does not differ between PRE- and FAF-trials, neither in the speaking ( $p_{2\text{-tailed}} = .797$ ,  $Z = -0.257$ ,  $n = 23$ ) nor in the singing condition ( $p_{2\text{-tailed}} = .189$ ,  $Z = -1.314$ ,  $n = 23$ ). In the speaking condition the response begins after 154.6 ms, in the singing condition after 142.1 ms, as calculated with the Castellan change-point test. In the interval 200-225 ms after vowel onset, mean voice  $F_0$  of the first syllable is 17.4 cents higher in FAF-trials in the speaking condition ( $S.D. = 34.0$  cents;  $p_{1\text{-tailed}} = .005$ ,  $Z = -2.571$ ,  $n = 24$ ), while in the singing condition voice  $F_0$  is 35.5 cents higher than in PRE-trials ( $S.D. = 28.8$  cents;  $p_{1\text{-tailed}} < .001$ ,  $Z = -3.954$ ,  $n = 23$ ).

In both conditions, mean voice  $F_0$  of the second syllable is also higher in FAF-trials compared to PRE-trials (Fig. 1b). In the speaking condition the difference in the interval 25-50 ms after vowel onset is 45.4 cents ( $S.D. = 37.9$  cents;  $p_{1\text{-tailed}} < .001$ ,  $Z = -3.832$ ,  $n = 23$ ), in the singing condition 64.6 cents ( $S.D. = 31.1$  cents;  $p_{1\text{-tailed}} < .001$ ,  $Z = -4.106$ ,  $n = 23$ ). In the interval 75-100 ms after vowel onset, mean voice  $F_0$  of the second syllable is 50.2 cents higher in FAF-trials in the speaking condition ( $S.D. = 38.8$  cents;  $p_{1\text{-tailed}} < .001$ ,  $Z = -3.582$ ,  $n = 19$ ), while in the singing condition voice  $F_0$  is 64.2 cents higher than in PRE-trials ( $S.D. = 27.0$  cents;  $p_{1\text{-tailed}} < .001$ ,  $Z = -3.883$ ,  $n = 20$ ).

In the interval 25-75 ms of the second syllable, which was chosen for comparison of the speaking and singing condition, response magnitude in the singing condition ( $M = 66.1$  cents,  $S.D. = 30.4$  cents) is higher than in the speaking condition ( $M = 46.7$  cents,  $S.D. = 37.3$  cents) ( $p_{1-tailed} = .013$ ,  $Z = -2.220$ ,  $n = 23$ ).

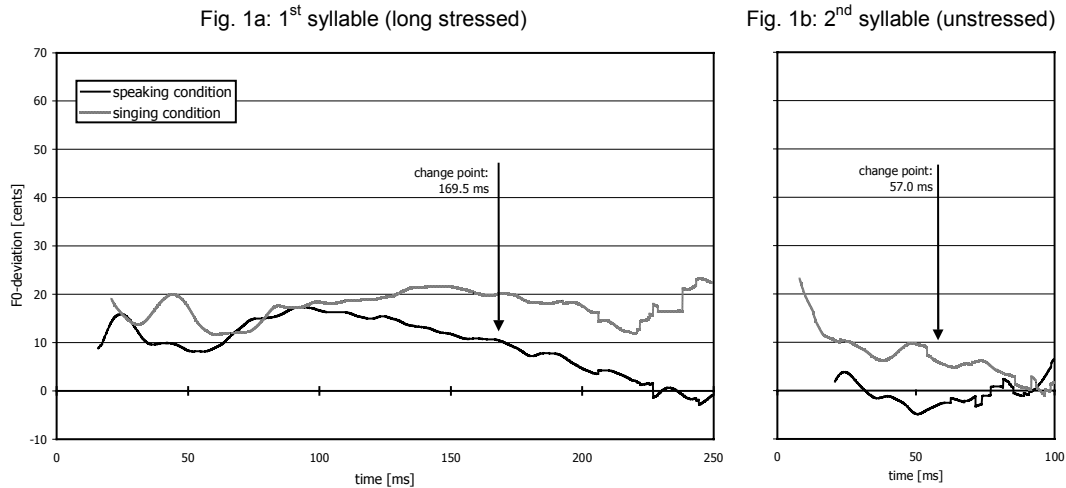


**Fig. 1:** Average deviation of voice  $F_0$  contours of the first syllable (=long stressed, Fig. a) and the second syllable (=unstressed, Fig. b) of the nonsense word [‘ta:tatas] during frequency shift from trials before frequency shift. **—** speaking condition, **—** singing condition. 0 ms is the onset of phonation, as defined by the first glottal closing. The contour begins at the point for which voice  $F_0$  data were available for all subjects and ends at 250 ms resp. 100 ms, when  $n$  was still greater than 19. Steps at the end of these averaged graphs are the result of single subjects stopping their vocalization. Change points were determined using the Castellan change-point test.

## B. Response latency and response magnitude after termination of frequency shift

In Figure 2 the voice  $F_0$  deviations of POST-trials from PRE-trials are shown. In the interval 25-50 ms after vowel onset (Fig. 2a),  $p$ -values do not reach the Bonferroni corrected significance level, but indicate the tendency that mean voice  $F_0$  of the first syllable is 10.9 cents higher in POST-trials than in PRE-trials in the speaking condition ( $S.D. = 21.2$  cents;  $p_{1-tailed} = .014$ ,  $Z = -2.200$ ,  $n = 24$ ) and 16.8 cents higher in the singing condition ( $S.D. = 33.1$  cents;  $p_{1-tailed} = .016$ ,  $Z = -2.143$ ,  $n = 24$ ). In the interval 200-225 ms after vowel onset mean voice  $F_0$  does not differ between POST- and PRE-trials in the speaking condition ( $p_{2-tailed} = .145$ ,  $Z = -1.457$ ,  $n = 24$ ). The response ends after 169.5 ms. However in the singing condition, voice  $F_0$  is 15.5 cents higher ( $S.D. = 18.5$  cents;  $p_{2-tailed} < .001$ ,  $Z = -3.254$ ,  $n = 23$ ).

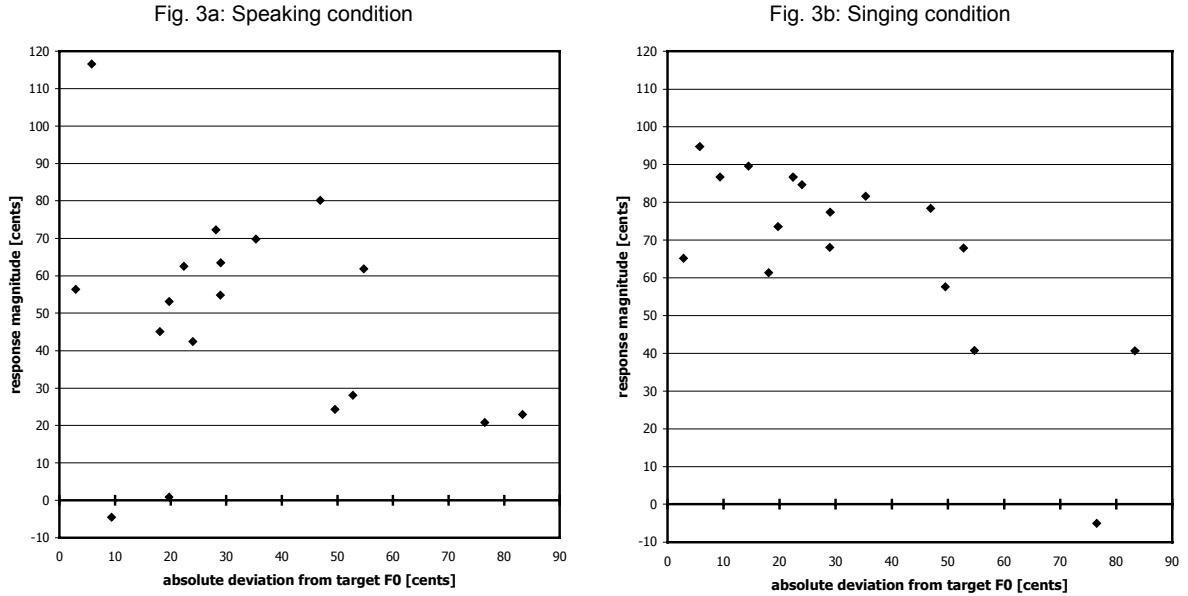
In the speaking condition, mean voice  $F_0$  of the second syllable (Fig. 2b) does not differ between POST- and PRE-trials in the interval 25-50 ms after vowel onset ( $p_{2-tailed} = .465$ ,  $Z = -0.730$ ,  $n = 23$ ). In the singing condition voice  $F_0$  is 8.2 cents higher ( $S.D. = 16.3$ ;  $p_{2-tailed} = .007$ ,  $Z = -2.677$ ,  $n = 23$ ). In the interval 75-100 ms after vowel onset, mean voice  $F_0$  does not differ between POST- and PRE-trials, neither in the speaking condition ( $p_{2-tailed} = .629$ ,  $Z = -0.483$ ,  $n = 19$ ) nor in the singing condition ( $p_{2-tailed} = .279$ ,  $Z = -1.083$ ,  $n = 20$ ). The response ends after 57.0 ms in the singing condition.



**Fig. 2:** Average deviation of voice  $F_0$  contours of the first syllable (=long stressed, Fig. a) and the second syllable (=unstressed, Fig. b) of the nonsense word ['ta:tatas] after termination of frequency shift from trials before frequency shift. **—** speaking condition, **—** singing condition. See caption for figure 1 for additional explanation.

### C. Relationship between accuracy of note matching and response magnitude

In order to investigate whether a relationship exists between the ability to match a note accurately and the magnitude of compensation under frequency shift, absolute deviation from target  $F_0$  and response magnitude as defined in section II.D. was calculated for each subject. Four subjects deviate from the fundamental frequency of the piano notes more than one semi-tone in PRE-trials. With deviations of 190, 287, 440 and 526 cents these subjects show virtually no ability to match an external pitch with their voice. These exceptionally large deviations pose a problem as far as the calculation of correlation is concerned because of the large influence of such outliers in a small sample. Consequently, these four subjects were excluded from the correlation analysis, even though they do show responses to frequency shift. Therefore it should be noted that the results reported here should only be generalized to individuals who show at least some ability to match external notes. Response magnitude and absolute deviation from target  $F_0$  are plotted against each other in Figure 3. Whereas in the speaking condition (Fig. 3a) no significant correlation was found (Spearman's rho = -.004,  $p_{1-tailed} = .494$ ,  $n = 19$ ), in the singing condition (Fig. 3b) response magnitude and absolute deviation from the target  $F_0$  correlate inversely (Spearman's rho = -.469,  $p_{1-tailed} < .022$ ,  $n = 19$ ). Therefore in singing, subjects who show better note matching abilities also show greater compensation during frequency shift.



**Fig. 3:** Individual absolute deviation from the target  $F_0$  (an indication of the accuracy of note matching) and response magnitude in the speaking (Fig. a) and in the singing condition (Fig. b). Response magnitude was calculated as the mean difference between FAF- and PRE-trials over the interval 25-75 ms after onset of phonation in the second short unstressed syllable. Absolute deviation of target  $F_0$  was calculated in the same interval as the absolute deviation of voice  $F_0$  in PRE-trials from the fundamental frequency of the piano notes presented as the target  $F_0$  in the singing condition.

#### IV. DISCUSSION

Regarding voice  $F_0$  control in speaking, the results of the present study confirm those of Donath *et al.* (2002). Auditory feedback is used to control voice  $F_0$  in vowel production by a negative-feedback system. Because of the long response latency of approximately 150 ms, deviations between intended and perceived voice  $F_0$  cannot be compensated during the first half of long stressed syllables. Unstressed syllables are therefore too short for a compensation to occur. In subsequent syllables the response persists and reaches its maximum. Therefore control of voice  $F_0$  is not interrupted by the onset and offset of phonation between syllables of a single word. Adjustments of voice  $F_0$  even persist when speaking is paused for several seconds. Therefore, auditory feedback is not used to regulate voice  $F_0$  within syllables, rather it is used for the adjustment of voice  $F_0$  supra-segmentally in terms of prosody.

Results in the singing condition are comparable to those of the speaking condition, except for response magnitude and latency of after-effect. In the speaking condition, the frequency shift was compensated to an extent of 47% of the frequency shift, demonstrating the limiting property of the auditory-vocal system shown in previous studies. However, in the singing condition the compensation was significantly higher reaching 66% of the frequency shift. The piano notes were presented as the target  $F_0$  just before the utterance, but not during it. Therefore there remained only an internal reference for the  $F_0$ . Results support the hypothesis that an internal reference contributes to the auditory-vocal system being effective in regulating voice  $F_0$ .

However even in the singing condition, no total compensation of the frequency shift was realized. The ability to match the target  $F_0$  provided by the piano notes varied greatly. The ability to match the target  $F_0$  correlates significantly with the response magnitude in the

singing condition. This means that subjects who have better control over their voice and/or better auditory perception thus enabling them to match external target frequencies more accurately, also compensate perceived deviations of their voice  $F_0$  better in singing. Therefore the mean response magnitude of 66% of the frequency shift in the singing condition might reflect the distribution of the ability of note matching in the sample. A future study comes to mind easily: When singers and non-singers are compared regarding their response magnitude during frequency-shifted auditory feedback, singers might show a better compensation in case they match the target notes more accurately than non-singers. Furthermore response latency of singers might be shorter than that of non-singers. This seems possible because musicians have superior pre-attentive auditory processing abilities (Koelsch, Schröger, and Tervaniemi, 1999). Finally it would be interesting to investigate whether an improvement of the ability of note matching leads to greater responses. Such studies would answer the question whether training could influence the system regulating voice  $F_0$ , which seems to work involuntary and reflex-like (Larson, 1998).

There seems to be no relationship between the ability to match the target  $F_0$  and the response magnitude in the speaking condition. Therefore, the ability for precise voice  $F_0$  control principally exists (as shown by the correlation between the ability to match the target  $F_0$  and the response magnitude in the singing condition), but is not used effectively in speaking. As stated in the introduction, it is assumed that voice  $F_0$  is not monitored very tightly in speaking because it is not necessary for comprehension of speech, at least in non-tone languages. Future studies might address differences between tone and non-tone languages in responses to frequency-shifted auditory feedback.

Response latencies in frequency shift trials were comparable in the speaking and the singing condition and are about 150 ms. In trials after termination of frequency shift, voice  $F_0$  seems to be increased from the very beginning of phonation. Voice  $F_0$  therefore is controlled continuously beyond pauses of phonation within words as well as between words which are separated by several seconds. The lasting effect after returning to non-altered auditory feedback indicates that voice  $F_0$  adapts in a feedforward fashion on a supra-segmental level. In the speaking condition, the after-effect comes to an end within the first syllable after about 170 ms as found by Donath *et al.* (2002) previously. Whereas the onset of response is characterized by a relatively fast increase of voice  $F_0$ , the decrease of voice  $F_0$  seems to take place more slowly. In the singing condition the after effect continues to the second syllable. The change point in the second syllable is 57 ms, so the total latency of the after effect in the singing condition adds to 415 ms with the vowel duration of the first syllable of about 278 ms and the estimated duration of the second consonant of 80 ms. In singing the compensation starts as quickly as in speaking, but remains longer, even when there is no deviation between intended and perceived voice  $F_0$  anymore. In singing, of course, it is important to compensate deviations quickly. Therefore it seems somewhat surprising that the after effect persists longer, i.e. the latency of correction is higher. Some singers report that they “glide” into notes right up from below. Maybe the other direction from top to bottom is less natural and therefore causes the higher latency of the after-effect.

## V. CONCLUSION

The study demonstrates how the paradigm of frequency shifted auditory feedback during speaking versus singing of nonsense words can be applied to get insight into segmental and supra-segmental processes controlling voice  $F_0$ . Results show that in speaking voice  $F_0$  is controlled to a lesser extent than in singing and that the degree of note matching ability influences the degree of compensation to frequency shift in singing.

## REFERENCES

- Burnett, T. A., Senner, J. E., and Larson, C. R. (1997), "Voice F0 Responses to pitch-shifted auditory feedback: A preliminary study," *J. Voice* 11, 202-211.
- Donath, Th. M., Natke, U., and Kalveram, K. Th. (2002), "Effects of frequency-shifted auditory feedback on voice  $F_0$  contours in syllables." *J. Acoust. Soc. Am.* 111(1), 357-366.
- Elman J. L. (1981), "Effects of frequency-shifted feedback on the pitch of vocal production," *J. Acoust. Soc. Am.* 73, 45-50.
- Jäncke, L. (1991) "The 'audio-phonatoric coupling' in stuttering and nonstuttering adults: Experimental contributions." In: *Speech Motor Control and Stuttering*, edited by H. F. M. Peters, W. Hulstijn, and C. W. Starkweather, Amsterdam: Elsevier Scientific Publishers, p. 171-180.
- Kalveram, K. Th., and Jäncke, L. (1989) "Vowel duration and voice onset time for stressed and nonstressed syllables in stutterers under delayed auditory feedback condition," *Folia Phoniatr.* 41, 30-42.
- Koelsch, S., Schröger, E., and Tervaniemi, M. (1999) Superior attentive and pre-attentive auditory processing in musicians. *NeuroReport* 10, 1309-1313.
- Larson, C. R. (1998), "Cross-modality influences in speech motor control: The use of pitch-shifting for the study of F0 control," *J. Comm. Dis.* 31, 489-503.
- Natke, U. (1999), "Die Kontrolle der Phonationsdauer bei stotternden und nichtstotternden Personen: Einfluß der Rückmeldelautstärke und Adaptation (Control of phonation in stuttering and nonstuttering persons: Influence of feedback loudness and adaptation)," *Sprache - Stimme - Gehör* 23, 198-205.
- Natke, U., and Kalveram, K. Th. (2001), "Fundamental frequency under frequency shifted auditory feedback of long stressed and unstressed syllables," *J. Speech Lang. Hear. Res.* 44, 577-584.
- Parlitz, D., and Bangert, M. (1998), "Short and medium motor responses to pitch shift: Latency measurements of the professional musician's audio-motor loop for intonation," Paper presented at the 16th international congress on acoustics and the 137th meeting of the acoustical society of America, Seattle, WA.
- Siegel, S., and Castellan, N. J. (1988), "Nonparametric statistics for the behavioral sciences," 2nd ed. Boston: McGraw-Hill.
- Sundberg, J. (1987), "The science of the singing voice," Dekalb, IL: Northern Illinois University Press.