

Effects of frequency-shifted auditory Feedback on Voice F_0 Contours in Syllables

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

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

Abstract. Previous studies have shown that, during continuous vocalization, voice fundamental frequency (voice F_0) is modified by frequency shifted auditory feedback. In this study, the effects of frequency-shifted auditory feedback on voice F_0 contours were determined for the first two syllables of the non-sense word [tatatas]. Results show that voice F_0 is auditorily controlled with a long latency and responses are not interrupted by onset and offset of phonation itself. Furthermore, after-effects were found in voice F_0 in trials after termination of frequency shift, which indicates that the response persists for several seconds. It is argued that the purpose of the auditory-vocal system is not to control voice F_0 precisely within single syllables, but rather on supra-segmental level in the context of prosody.

I. Introduction

Previous studies investigating the auditory-vocal system with the technique of frequency-shifted auditory feedback demonstrated that voice fundamental frequency (voice F_0) is monitored and controlled through a closed-loop negative feedback system with a relatively long latency of >100 ms and a non-linear limiting property, which prevents corrective responses from exceeding a certain magnitude. When artificially increasing or decreasing the pitch of auditory feedback during vocalization, subjects change their voice F_0 in the opposite direction to compensate for the difference between perceived and intended pitch. However, no study has shown voice F_0 changes exceeding 60 cents for feedback frequency shifts between 100 and 600 cents (Elman, 1981; Burnett, Senner & Larson, 1997; Larson, 1998; Natke & Kalveram, 2001). Also, no significant correlation between frequency shift and response magnitude has been found (Larson, 1998), which has led to the argument for a non-linear limiting property (Larson, Burnett & Kiran, 2000).

This current understanding of the nature of the auditory-vocal system is based almost exclusively on a paradigm in which the pitch of auditory feedback is modified unexpectedly while subjects vocalize continuously for several seconds (Elman, 1981; Burnett et al., 1997; Larson, 1998; Jones & Munhall, 2000). Although this paradigm has provided much insight into the auditory-vocal system, it is not directly comparable to vocalization during speech. Speech is characterized by rapidly repeated onsets and offsets of phonation during the combination of vowels, voiced consonants and voiceless consonants to form words and sentences. Therefore the question arises whether the characteristics of voice F_0 control established so far also apply to speech, and which new aspects of F_0 control can be found by investigating phonation during speech. Utilizing a paradigm introduced by Natke and Kalveram (2001), in which subjects utter a non-sense word with frequency shifted feedback, this study determined voice F_0 contours in syllables of the non-sense word [‘ta:tatas]. In this way segmental and supra-segmental characteristics of voice F_0 control including latency, magnitude, speed, and duration of responses were measured.

The behavioral importance of precise voice F_0 production can be shown by examining the different informational aspects of prosody, of which F_0 is one dimension besides duration and amplitude. For a review of different aspects of prosody, see Cutler, Dahan and van Donselaar (1997). First, there is evidence that prosodic information can have an effect on syntactic analysis. There is also evidence that listeners search actively for accented words to comprehend discourse structure because sentence accent varies depending on semantic context. For example, based on sentence context and the importance of words, a different noun would be stressed in the sentence "Joey gave a book to Mike.". Tone languages such as Thai or Cantonese illustrate the potential of voice F_0 for lexical processing. In these languages, variations of F_0 in single syllables can distinguish between different meaning.

Prosody also conveys information about the emotional state of the speaker. Thus it becomes biologically important to realize the intended prosody. Yet another aspect of prosody is that individuals increase or decrease the similarity between their own and others' prosody depending on the situation. A socio-linguistic explanation for this phenomenon is that by matching prosody one can demonstrate social identification (Giles & Powesland, 1975). Since prosody plays a role in so many aspects of communication, it appears to be most important to control voice F_0 precisely.

A negative feedback system in auditory-vocal control would serve to stabilize voice F_0 by comparing the auditorily perceived pitch with an internal reference and adjusting voice F_0 in the opposite direction of deviations. Accordingly, the previously mentioned studies consistently found so called *opposing responses* as a result of changes of the pitch of auditory feedback. Although the existence of a negative feedback system has been proven many times, the question remains why mean group responses fell within a range of approximately 30 to 60 cents while frequency shifts in most studies were 100 cents (one semitone) or even larger. Burnett, Freedland, Larson and Hain (1998) suggested that the control system might only be used as a fine-tuning mechanism for small magnitude F_0 fluctuations.

In a study by Larson (1998), subjects were instructed to change their voice F_0 either in the same or the opposite direction of the frequency shift or ignore the frequency shift. He reported that the first response was usually opposing with a latency of 100-150 ms, not strongly influenced by instructions and therefore automatic, while the second response appeared to be a reflection of voluntary mechanisms with latencies between 250 and 600 ms. Opposing responses with similar latencies have been found in other studies as well: 104-200 ms (Burnett et al., 1997), 228 ms (Burnett et al., 1998), and 243 ms (Larson et al., 2000).

Besides opposing responses, also so called *following responses* have been found (e.g., Burnett, Senner & Larson, 1997; Larson, 1998). In the context of a negative feedback system, such responses would not be very useful because as voice F_0 is being correct in the same direction like the perceived deviation, the feedback F_0 would move away even further from the internal reference. However, following responses would be useful when adjusting voice F_0 to match an external reference. For example, when the F_0 of a piano note is higher than one's own voice F_0 , voice F_0 has to be increased until it matches the external frequency. It is still unclear when or why following responses occur. Larson (1998) reports that following responses seem to occur more often when the frequency shift is larger.

Natke and Kalveram (2001) demonstrated that the closed-loop negative feedback mechanism found in experiments using continuous vocalization also applies to speech. In their study, frequency shift was randomly activated before subjects uttered a three-syllable non-sense word. Thus the complete word was produced with frequency-shifted auditory feedback. Opposing responses were found in long-stressed, but not unstressed (short) first syllables. In the second syllables, opposing responses were found in both unstressed and long-stressed syllables. The authors concluded that long-stressed syllables are long enough for voice F_0 to be affected within them, whereas latencies of responses are too long for voice F_0 to be affected within unstressed and therefore short syllables. Furthermore, they suggested that in second short syllables, re-

sponses can be observed because latencies are equal to or less than the duration of preceding syllables.

In the study of Natke and Kalveram, average voice F_0 was calculated for entire syllables and latencies of response were not determined. However, the authors inspected F_0 contours of a limited number of trials and found onsets of responses approximately in the middle of long-stressed first syllables. Since the average response calculated this way reflects also part of the syllable in which there is no change in F_0 , the peak response in long-stressed first syllables was likely greater than the reported 25 cents. Overall the study revealed that the mechanism controlling voice F_0 works at the syllabic level in speech production and most likely applies to natural speech as well. There is also some evidence that the mechanism of voice F_0 control is not limited to relatively artificial settings, but is active during spontaneous speech as well. Natke, Grosser and Kalveram (2001) utilized frequency shifts of half an octave downwards and upwards during spontaneous speech in stuttering and non-stuttering persons. Non-stuttering persons responded with a decrease of global voice F_0 of 36 cents under the upward shift condition, but with no change under the downward shift condition.

Continuing the previous line of research, in this study the effects of frequency-shifted auditory feedback on voice F_0 contours in successive syllables were determined. In this way, temporal aspects of auditory-vocal F_0 control on segmental and supra-segmental level in syllable production were investigated.

II. Method

A. Subjects

Twenty-two adults (11 women, 11 men) between 19 and 29 years of age participated ($M = 23$ years, $S.D. = 3.4$ years). Exclusionary criteria were a self-reported current speech disorder and a mother tongue other than German. A hearing screening assured that subjects had 20 dB HL or

better pure-tone thresholds for the frequencies 0.25, 0.5, 1, 2, 4 kHz bilaterally (audiometric test: Hortmann DA 323, Neckartenzlingen, Germany).

B. Apparatus

The subject's voice was frequency-shifted with a commercial device (DFS 404, Casa Futura Technologies, Boulder, Colorado, USA), which works digitally (32 kHz, 14 bits) and had been modified with a relay to enable remote-switching between non-altered auditory feedback and frequency-shifted auditory feedback. Auditory feedback was provided through sealed headphones (Blackhawk, DSP 5DX, Flightcom, Portland, Oregon, USA), which have a built-in microphone to pick up subjects' voices and, according to the manual, attenuate air-conducted sound by approximately 24 dB SPL. Feedback volume was adjusted so that a sinusoidal tone of 440 Hz with 75 dB (A) at the microphone led to a headphones feedback volume of 70 dB (A). Subjects reported that this volume was comparable to auditory feedback during a normal conversation.

In order to mask bone conducted sound, low-pass-filtered white noise ($f_c = 900$ Hz) was presented at an intensity of 65 dB (A) during the experiment. Since there was no simple approach to physical measurement of masking efficiency, our own and the subjects' observations served as an indicator. When the artificial feedback was turned off suddenly while the masking noise continued, subjects reported that to them it seemed as if they had stopped speaking because they could no longer hear their voice. This indicates that the masking level was sufficient to mask any speech sounds not presented through headphones. However, the masking level was still low enough for subjects to hear their voice clearly and distinctly over the noise.

The use of masking noise is debatable because it also masks the auditory feedback itself and thus the independent experimental variable. However, with masking subjects hear only the electronically processed feedback, as opposed to a mix of processed and bone-conducted sound. The latter had previously been described by subjects as a doubling of their voices, which presents a more serious interference with the independent experimental variable. Although in this study

masking noise was used, Burnett et al. (1998) did not find that the presence or absence of masking noise had an effect on responses to frequency shift.

Vibrations of vocal folds were recorded with an electroglottograph (EGG; Laryngograph, Kay Elemetrics, Pine Brook, New Jersey, USA). A commercial personal computer with stereo sound-card digitized the analogue EGG signal and stored it with a sampling rate of 11,025 Hz and sampling resolution of 16 bits. In addition, the computer switched the frequency shift device between non-altered and frequency-shifted auditory feedback. In addition to the microphone used for feedback, a second small microphone attached to subjects' clothes picked up the acoustical speech signal. The acoustical signal was recorded along with the EGG signal. In this way the source of potential artifacts in the EGG signal could later be established more easily by inspecting the acoustical signal.

C. Procedure

Subjects had to speak the non-sense word [ˈta:tatas] at their habitual voice F_0 level. Furthermore, they had to speak at a speed and with a stress pattern indicated by a tone sequence presented through headphones. The tone sequence consisted of three 440 Hz sinusoidal tones. The first tone for the long-stressed syllable had a duration of 400 ms, the following two tones were each 200 ms long. Tones were separated by 40 ms pauses. The tone sequence was played twice, after which the subjects had 3.5 s to utter the word. Subjects were instructed to speak clearly, with normal volume, and with monotone voice (i.e., no voice F_0 variation). However, subjects were not informed that the feedback would be modified from time to time. Also they were not specifically instructed to maintain their voice F_0 as a consequence to modified feedback. Playback of tones and recording of data was automated, so that subjects sat alone in a sound-insulated, but not reflection-free room.

To ensure that subjects uttered the non-sense word correctly, a training phase preceded the actual experiment. During the training phase, the tone sequence was played repeatedly in a loop

while subjects simultaneously spoke along until they produced the word correctly at least five times in succession, as judged by the experimenter. Generally, subjects reached the criterion within ten trials.

The experimental phase consisted of 30 trials, each approximately 7 s long. The frequency of auditory feedback was shifted downward by 100 cents in 20% of all trials, resulting in 6 frequency-shifted trials. In these trials, the frequency shift was activated at the beginning of the tone sequence, so that auditory feedback, but not the tone itself was frequency-shifted from the very beginning of subjects' utterances. The frequency shift was deactivated without subjects' notice after they had uttered the non-sense word. Trials with frequency shift were arranged randomly with the limitation that at least two trials without frequency shift had to precede a frequency-shifted trial and the last trial could not be frequency-shifted (see below for explanations).

Additionally, a public speaking condition was introduced in a randomized fashion to investigate potential effects of a stressor (the presence of two listeners) on the auditory-vocal system. These results will be reported elsewhere (Donath, Natke & Kalveram, 2001).

D. Data Analysis

See figure 1 for a schematic illustration of data analysis.

(Figure 1)

Data were analyzed using MATLAB (The MathWorks Inc., Natick, Massachusetts, USA). The raw EGG signal was first highpass-filtered (Butterworth-type, $f_c = 2$ kHz). Since the glottal closing is characterized by a steep increase of the EGG signal (Childers & Krishnamurthy, 1984), these increases were assessed by determining the maximums of the EGG signal's first derivative. The period between two closing instants equals the duration of one phonation cycle and its recip-

rocal equals the momentary F_0 . The first detected closing instant defined the onset of phonation. F_0 contours were obtained for the vowels of the first two syllables.

Since the glottal closing instants and the F_0 values associated with them were spaced irregularly in time, F_0 contours with a fixed resolution of 0.1 ms steps were interpolated, so that averaging across trials and subjects became possible. In this way, for each vocalization a F_0 contour was obtained. The F_0 contours of single subjects were then truncated at the end, so that equal lengths were obtained. The shortest vowel duration determined the length of all F_0 contours for one subject. This approach was chosen to ensure that all portions of F_0 contours for one subject 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 21 trials (7 PRE-trials, 6 FAF-trials, and 8 POST-trials) were discarded because of insufficient length or artifacts, resulting in a total of 383 trials for the first syllable and 388 trials for the second.

F_0 contours of vowels in words before (PRE), during (FAF), and after (POST) trials with frequency-shifted auditory feedback were analyzed. Accordingly, the truncated F_0 contours for each subject were averaged across the six trials preceding, the six trials during, and the six trials after frequency-shifted auditory feedback. The remaining 12 trials were not analyzed. Trials preceding frequency-shifted auditory feedback were used as reference trials, to which FAF- and POST-trials were compared for each subject. This approach was based on the assumption that subjects' voice F_0 should have returned to normal levels in PRE-trials because these were separated from FAF-trials by at least one trial without any frequency shift.

When examining subjects' F_0 contours, it became apparent that voice F_0 is not constant within single vowels. After averaging, voice F_0 still shows fluctuations, which might indicate that subjects have an individual voice F_0 contour during production of vowels. After a fast initial change, some subjects show a slow increase or decrease of voice F_0 over the course of vowel production. In case subjects have individual, non-random voice F_0 contours in vowels, the F_0 variation solely

due to frequency-shifted auditory feedback can be isolated by calculating the voice F_0 contour differences between FAF- and PRE-trials as well as POST- and PRE-trials. In other words, the voice F_0 variation shared by PRE-, FAF-, and POST-trials is removed, leaving only random variation and variation introduced by frequency-shifted feedback. Another advantage of using trials during the experiment as a reference is based on the large intra-individual variation of voice F_0 . For example, Coleman and Markham (1991) found that habitual voice F_0 varies as much as plus/minus three semitones, and Jones and Munhall (2000) observed a significant steady increase in voice F_0 along 140 trials even without alteration of auditory feedback. Therefore this approach should have minimized the error introduced by slow fluctuations of voice F_0 over the course of several trials.

In order to eliminate the inter-individual variance due to the different pitches of the male and female subjects, whose habitual voice F_0 ranged from 99 Hz to 251 Hz, and based on the reasoning described above, the differences in cents between F_0 contours in PRE-trials and FAF-trials as well as PRE-trials and POST-trials were calculated for each data point. The resulting contours reflect the deviation of momentary F_0 during frequency shift and after its termination from normal productions under non-altered auditory feedback, independently of the subjects' habitual voice F_0 .

Based on previous research summarized in the introduction and the closed-loop negative feedback system, one-tailed hypotheses about the differences between PRE-trials and FAF-trials and between PRE-trials and POST-trials were formulated. In order to test for differences in the initial and final portions of voice F_0 contours, they were arbitrarily divided into 25 ms intervals. Means for these intervals were calculated individually for subjects. However, since subjects' vocalizations were of unequal lengths, it had to be established, which of the last 25 ms intervals was based on a reasonably high number of subjects. In our graphs, 0 ms correspond to onset of phonation, which was defined by the first detected glottal closing. The voice F_0 contours begin at the point for which momentary voice F_0 values became available for all subjects, and therefore the

graph reflects the entire group. The length of a glottal cycle is almost 10 ms in men and momentary voice F_0 values were “assigned” to the end of each cycle before interpolation and subsequent plotting. Therefore momentary voice F_0 is not defined for the first few ms after onset of phonation. Consequently, the interval 0-25 ms after onset of phonation was not analyzed. By investigating the second 25 ms interval, it was ensured that the average voice F_0 used for testing is based on all subjects.

Wilcoxon rank sign tests were calculated for the intervals 25-50 ms ($N \geq 21$), 225-250 ms (long-stressed syllables, $N \geq 20$), and 75-100 ms (short unstressed syllables, $N \geq 17$) after onset of phonation.

It was hypothesized that several intervals would have a higher voice F_0 compared to PRE-trials. A compensatory response due to the downward frequency shift should occur once auditory feedback is processed and muscular changes take effect. Therefore the last interval of the first long-stressed syllable in FAF-trials as well as the first and last interval of the second unstressed syllable in FAF-trials should have a higher voice F_0 . The compensatory response should last until corrected by the feedback not shifted in frequency. Therefore the first interval of the first long-stressed syllable in POST-trials should also have a higher voice F_0 . In this case the latency of the response will cause the first portion of the vowel to have a higher voice F_0 before returning to normal. Additional p -values were calculated as effect measures for intervals of syllables for which no changes in voice F_0 were assumed. (When interpreting p -values as effect measures, a high p -value indicates a lower probability for a difference and vice versa.) Since a normal distribution could not be assumed for responses, non-parametric Wilcoxon rank sign tests were calculated; one-tailed to test our hypotheses and two-tailed for other intervals.

The (average) latencies of responses in first syllables in FAF-trials and first syllables in POST-trials were determined based on a non-parametric approach to the change-point problem. The Castellán change-point test for continuous variables is related to the Mann-Whitney-Wilcoxon test and is described concisely in Siegel and Castellán (1988). This test estimates the (single) point

of change in the distribution of a sequence and is more precise compared to approaches which estimate latency by determining the point at which a curve exceeds an arbitrarily chosen threshold (e.g. 2 S.D.). The latter approaches always overestimate latencies. The calculation of the test statistic posed a special problem: The time series of momentary fundamental frequencies to which the test was applied is the result of averaging across subjects and trials. Therefore the actual number of data points can only be estimated, and data points are spaced irregularly in time. A very conservative approach was chosen. The minimum number of data points, based on which the time series could be created, is the average number of phonation cycles occurring over its length. This N is only a fraction of the actual number of data points, on which the contours are based, but yields a more realistic Z -score.

Since six tests were calculated, the significance level $\alpha = 5\%$ was corrected according to Bonferroni to $\alpha' = \alpha/6 \approx 0.008$. Additionally, the distribution of individual responses in second syllables during frequency shift and first syllables after termination of frequency shift were explored. Therefore potentially existing groups of individuals exhibiting different response magnitudes because of low or high degree of audio-vocal control might be found. The correlation (Spearman-rho) between the two was calculated as well. In this way, a possible relationship between the degree to which vocal F_0 is auditorily controlled and the magnitude of adaptation, as indicated by effects in utterances following frequency-shifted trials, might be established.

III. RESULTS

A. First syllable during frequency shift

Figure 2 shows the voice F_0 deviation of FAF-trials from PRE-trials for first syllables. There is no significant difference between PRE-trials and FAF-trials in the interval 25-50 ms after vowel onset ($Z = -1.494$, $n = 21$, $p_{2-tailed} = 0.714$). Voice F_0 begins to increase during frequency-shifted auditory feedback after 157 ms, as determined with the Castellan change-point test ($Z = -5.541$, n

= 42, $p_{1\text{-tailed}} < 0.001^*$). The response velocity is 339 cents/s in the interval 150-250 ms after vowel onset, as measured by calculating the slope with a linear regression. In the interval 225-250 ms after vowel onset, mean voice F_0 is 28.3 cents higher than in PRE-trials ($Z = -3.380$, $n = 20$, $p_{1\text{-tailed}} < 0.001^*$).

(Figure 2)

B. Second syllable during frequency shift

Figure 3 shows the voice F_0 deviation of FAF-trials from PRE-trials for second syllables. In the interval 25-50 ms after vowel onset, mean voice F_0 is 51.5 cents higher than in PRE-trials ($Z = -3.924$, $n = 21$, $p_{1\text{-tailed}} < 0.001^*$). In the interval 75-100 ms after vowel onset, mean voice F_0 is 51.1 cents higher than in PRE-trials ($Z = -3.517$, $n = 17$, $p_{1\text{-tailed}} < 0.001^*$).

(Figure 3)

C. First syllable after termination of frequency shift

Figure 4 shows the voice F_0 deviation of POST-trials from PRE-trials for first syllables. In the interval 25-50 ms after vowel onset, mean voice F_0 is 20.6 cents higher than in PRE-trials ($Z = -3.099$, $n = 22$, $p_{1\text{-tailed}} < 0.001^*$). The Castellan change-point test indicates that voice F_0 begins to decrease after 171 ms ($Z = 4.854$, $n = 42$, $p_{1\text{-tailed}} < 0.001^*$). There is no significant difference between PRE-trials and POST-trials in the interval 225-250 ms after vowel onset ($Z = -1.791$, $n = 20$, $p_{2\text{-tailed}} = 0.074$).

(Figure 4)

D. Second syllable after termination of frequency shift

Figure 5 shows the voice F_0 deviation of POST-trials from PRE-trials for second syllables. There are no significant difference between PRE-trials and POST-trials in the interval 25-50 ms ($Z = -1.791$, $n = 20$, $p_{2-tailed} = 0.140$) and 75-100 ms after vowel onset ($Z = -2.123$, $n = 17$, $p_{2-tailed} = 0.033$).

(Figure 5)

E. Relationship between response magnitude during frequency shift and magnitude of effects after termination of frequency shift

To investigate the distribution of individual response magnitudes, the average response magnitude in the second short syllable during frequency-shifted auditory feedback was plotted for each subject (Fig. 6). The second syllable was chosen because the response is at its maximum and relatively stable, presumably reflecting the maximum of exercised control over phonation based on auditory feedback.

Exact Kolmogorov-Smirnov goodness of fit tests were calculated to test for a violation of normal-distribution. For response magnitudes in second syllables under frequency shift, p is 0.912 ($Z = 0.529$, $n = 21$). For magnitudes of after-effects in first syllables after termination of frequency shift, p is 0.523 ($Z = 0.578$, $n = 21$).

To investigate a potential relationship between maximum response magnitude, which might reflect the amount of individual auditory-vocal control, and after-effects in the first syllable after termination of frequency shift, F_0 deviations of the beginning (25-100 ms) of the first syllable after termination of frequency shift were compared to the response magnitudes in second syllables during frequency shift. As seen in figure 6, there appears to be no obvious relationship between the two. Spearman-rho is 0.173 ($p_{2-tailed} = 0.454$, $n = 21$).

(Figure 6)

IV. DISCUSSION

The data in this study show several features of the auditory-vocal system in respect to F_0 control in speech. The data support the view of a negative-feedback system with a limiting property during production of vowels. Latency and duration of responses indicate that the auditory-vocal system is used to control voice F_0 on supra-segmental level, rather than within single syllables. Furthermore, voice F_0 adjustments based on auditory feedback remain for several seconds during pauses of speech.

A. Response Latency

Adjustment of voice F_0 occurs after an interval, within which auditory feedback is processed, the efferents to the larynx are modified, and physical changes of vocal folds take place. The long response latency of 157 ms prevents auditorily controlled realization of intended F_0 in short syllables and during the initial portion of long syllables. Therefore auditory feedback would not be very useful for adjusting voice F_0 being realized during speech, but more useful in the regulation of voice F_0 in later segments or the learning of transformations between larynx configurations and subglottal air pressure and resulting voice F_0 . The latter function of auditory feedback would be the basis for producing the correct voice F_0 at the onset of phonation during speech or singing. The first function could serve to adjust for temporarily changed conditions of the larynx and to control supra-segmental prosody.

It can be assumed that prosodic information is encoded in relative changes of voice F_0 rather than in absolute pitch; otherwise inter- and intra-individual variations in baseline voice F_0 would make prosodic decoding impossible. By monitoring actually realized voice F_0 of syllables, adjust-

ments can be made for later syllables, i.e., on a supra-segmental level. In this way, the relative differences in voice F_0 that encode information are produced more reliably.

B. Response magnitude

The maximum voice F_0 response in the opposing direction is approximately 50 cents in this study, even though a complete compensation would require a change of 100 cents. This maximum magnitude of response agrees well with earlier findings (Elman, 1981: approximately 40-60 cents; Burnett et al., 1997: approximately 30 cents; Larson, 1998: approximately 30 cents; Natke & Kalveram, 2001: 15-65 cents), even though different paradigms such as continuous vocalization and speaking of non-sense words and frequency shifts varying from 100-600 cents were employed. These magnitudes of less than a semitone may seem small, but the compensation was easily audible in the audio records and falls well within natural prosodic variations. For example, at the end of questions voice F_0 increases around 100 cents (Bosshardt, Sappok, Knipschild & Hölscher, 1997).

Although previous findings and these results support the closed-loop model, no study has yet demonstrated a correlation between frequency shift magnitude and resulting response magnitude. This is not surprising because most studies have used frequency shifts of 100 cents and more, and only one used frequency shifts as low as 25 cents (Burnett et al., 1998), while responses seem to be limited to 30-60 cents. Therefore the range in which a correlation between frequency shift magnitude and resulting response magnitude might be observed was not investigated systematically.

Several hypotheses, which do not necessarily exclude each other, are offered here to address the question why the response does not exceed approximately 30-60 cents regardless of frequency shift magnitude. First, besides auditory feedback, mechano-receptors in the vocal fold mucosa or proprio-receptors in the vocal fold muscles might play a role in voice F_0 regulation. Research on this topic is limited because of the invasive techniques necessary for experimental

investigation. Reduced voice F_0 control was demonstrated after anesthetization of the laryngeal nerve sup. (Tanabe, Kitajima & Gould, 1975) and after anesthetization of the undersides of the mucosa of the vocal folds (Sundberg, Iwarsson & Billström, 1993). The different feedback channels would carry different information during frequency shift: proprioceptive and mechanoreceptive feedback would report actual voice F_0 , whereas auditory feedback reports a deviation. The integration of feedback channels might limit the maximum response in a non-linear fashion.

Another possibility might be that during speech (not singing), efferents controlling the larynx and modifying voice F_0 are only modified within close boundaries which reflect the small variations in F_0 occurring naturally in prosody. However, mean variations of 30-60 cents seem low compared to a S.D. of 15.87 Hz (corresponding to approximately 250 cents) of voice F_0 during reading of an unemotional, narrative text (Eady, 1982). Only the elimination of proprioceptive feedback combined with frequency shift of auditory feedback could isolate the role of auditory feedback and indicate its limitations for regulation. For example, Sundberg et al. (1993) demonstrated that anesthetization of the vocal folds reduces voice F_0 control. Although this approach decreases external validity considerably because of a complete loss of rather than a variation of proprioceptive feedback, it might be established to which degree (overlearned) motor programs limit voice F_0 variation based solely on auditory feedback.

When exploring individual subject's (average) responses in the second short syllable, it appears that response magnitudes are normally distributed around a mean magnitude of approximately 50 cents, ranging from 13 to 92 cents (Fig. 3). Therefore it is not reasonable to assume a simple distinction between individuals who rely heavily or little on auditory feedback for voice F_0 control. Rather there seems to be a continuum for the degree of auditory-vocal control. It could be that subjects whose voice F_0 is affected more strongly by auditory feedback also exhibit a stronger correlation between response magnitude and frequency shift magnitude. A future study might address this.

It might also be that voice F_0 in speech simply is not controlled very tightly. First, an external reference frequency for regulating voice F_0 is missing while speaking. When a reference tone of 440 Hz is provided, trained singers can match it with an accuracy of more than 1 Hz, which is a difference of approximately 4 cents (Sundberg, 1987). Furthermore, it has been shown that an unpredictable frequency shift can be completely compensated for in singing (Burnett et al., 1997; Parlitz & Bangert, 1999). This indicates that the compensatory mechanism for voice F_0 can be very effective when a reference tone is represented internally, as in singing.

When speaking syllables, an intended voice F_0 may also be given for single syllables. However in speaking, relative changes in voice F_0 would seem to be more important than realization of absolute voice F_0 . Voice F_0 varies much more across gender and age than within one individual during normal speaking. Nevertheless non-verbal information can be extracted from speech of children, women, and men with different voice registers. Therefore relative changes must encode non-verbal information and not absolute voice F_0 . Second, for precise comprehension of syllables and individual words, one would assume that voice F_0 is less important compared to formants and formant transitions, at least in languages such as English or German. Thus, it seems less important for the speaker to monitor and regulate fundamental frequency within syllables. This of course does not apply to tone languages, in which syllable voice F_0 contours have lexical function. Consequently, although a compensatory mechanism for voice F_0 at the syllabic level may exist, it seems that for languages such as English and German control of voice F_0 on supra-segmental level is more important (e.g., for word focus or conveyance of emotional states).

C. Response Duration

Since in this study subjects had to utter a word, features of the supra-segmental nature of the auditory-vocal system could be examined. While a partial correction of a mismatch between intended and auditorily perceived voice F_0 occurs in first syllables with a latency, second syllables show a higher F_0 from the very beginning. Since syllables were separated by a voiceless conso-

nant, phonation stopped between them. Therefore control of voice F_0 is continuous within single words and not interrupted by the onset and offset of phonation itself. It would seem plausible if such a continuous control scheme also applied to natural running speech.

Also in trials after termination of frequency shift, the beginning of first syllables still has a significantly higher F_0 , even though auditory feedback is normal at that point and there had been a pause in speech for approximately 6 seconds. This indicates that the auditory-vocal system regulates voice F_0 of syllables in words produced several seconds later, based on information gathered in single syllables. Therefore there is evidence for a relatively fast, i.e. within a single word, and persisting adjustment of voice F_0 production based on auditory feedback. A detected deviation between intended and perceived voice F_0 is not only partially corrected, but also leads to adjustments of later vowels, which are separated by a pause. In trials after termination of frequency shift, the intended voice F_0 is actually exceeded as a result of this adjustment. Thus there is clear evidence for the supra-segmental nature of auditory-vocal control. With normal auditory feedback in such trials, the mismatch is detected and voice F_0 decreases after a latency of approximately 170 ms, which is comparable to the 157 ms latency in frequency-shifted trials.

Voice F_0 is increased only by approximately 20 cents after termination of frequency shift, which is less than the approximately 50 cents during frequency shift. This may be the result of a ‘decay’. Over time, the magnitude of voice F_0 correction to be integrated with intended F_0 may decrease. This might be due to a memory effect; the voice F_0 correction ‘fades’ as time passes. Future studies might investigate the characteristics of the temporal persistence of modifications based on auditory feedback.

It should be noted that these after-effects are different from the sensorimotor adaptation to auditory frequency shift demonstrated by Jones and Munhall (2000). They changed the pitch of auditory feedback in one cent steps from trial to trial until they reached a shift of 100 cents, maintained this shift for 20 trials, and finally returned feedback to normal pitch for 10 trials. Subjects gradually increased their voice F_0 , although not to the extent necessary for a complete

compensation. The last trials provided evidence for a sensorimotor adaptation. When subjects had heard their voice's pitch higher than their true voice F_0 and feedback became unexpectedly normal (i.e., seemingly lower than before), their voice F_0 increased, and vice versa. The authors state that "subjects acted as if a remapping between perceived and produced pitch had taken place" and compare their adaptation data to classic prism experiments (e.g., Held, 1965), in which subjects make errors in the opposite direction of visual prism displacement after a training period.

Interestingly, there is no relationship between the magnitude of responses in frequency-shifted trials, which might indicate the individual level of auditory-vocal control, and the magnitude of after-effects at the beginning of the following word. This indicates that individuals can have different levels of auditory-vocal control and 'adjustment decay rates', and that therefore these two aspects of the auditory-vocal system are mediated through independent mechanisms.

Although no subjects exhibited so called following responses during frequency shift, two subjects showed after-effects in the opposite direction. Their voice F_0 is approximately 60 and 20 cents lower in trials after termination of frequency shift compared to trials before frequency shift. At this point it is not clear whether this is a random effect or might reflect that some individuals exhibit a deviant voice F_0 control. More data would be needed to investigate this effect systematically and test it statistically. A future study might address this.

The very fast adjustment found in this study might reflect the need for quick and also persisting changes in control over laryngeal muscles, in order to consistently achieve intended voice F_0 because of the complex neuro-physical processes involved in phonation. As has been previously suggested by Natke and Kalveram (2001), the relatively long latency prevents auditory feedback from efficient F_0 control within syllables. Voice F_0 in short syllables can not be controlled auditorily at all, and voice F_0 in long syllables can only be adjusted after a rather long latency. Therefore the picture that emerges is that auditory feedback is primarily used to adjust voice F_0 of following syllables, that is on a supra-segmental level.

V. CONCLUSION

This study demonstrates that the paradigm of frequency shifted auditory feedback during speaking of non-sense words can address several new aspects of auditory-vocal control during speaking. Rather than using auditory feedback for adjustment of voice F_0 in real-time, the results point towards a role of the auditory-vocal system in supra-segmental monitoring and control of voice F_0 . Based on auditory feedback, quick adaptations can be made to ensure that intended voice F_0 is produced and information is encoded properly in prosody.

ACKNOWLEDGMENTS

This research was supported by the Deutsche Forschungsgemeinschaft (DFG), grant no. Ka 417/13-3.

References

- Bosshardt, H. G., Sappok, C., Knipschild, M., and Hölscher, C. (1997), "Spontaneous imitation of fundamental frequency and speech rate by nonstutterers and stutterers," *Journ. Psych. Ling. Res.* **26**, 425-448.
- Burnett, T. A., Freedland, M. B., Larson, C. R. and Hain, T. C. (1998), "Voice F0 responses to manipulations in pitch feedback," *J. Acoust. Soc. Am.* **103**, 3153-3161.
- 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.
- Childers, D. G. and Krishnamurthy, A. K. (1984), "A critical review of electroglottography," *CRC Critical Reviews in Biomedical Engineering* **12**, 131-61.
- Coleman, R. F. and Markham, I. W. (1991), "Normal variations in habitual pitch," *J. Voice* **5**, 173-177.
- Cutler, A., Dahan, D. and van Donselaar, W. (1997), "Prosody in the comprehension of spoken language: A literature review," *Lang. Speech* **40**, 141-201.
- Donath, T., Natke, U. and Kalveram, K. T. (2001), "Magnitude and latency of fundamental frequency response within syllables under frequency shifted auditory feedback and public speaking," In B. Maassen, W. Hulstijn, R. D. Kent, and P. H. H. M. Van Lieshout (Eds.) *Speech Motor Control in Normal and Disordered Speech. Proceedings 4th International Speech Motor Conference (61-64)*. Nijmegen, The Netherlands:Uitgeverij Vantilt, 2001.
- Eady, S. J. (1982), "Differences in the F0 patterns of speech: Tone language versus stress language," *Lang. Speech* **25**, 29-41.
- Elman J. L. (1981), "Effects of frequency-shifted feedback on the pitch of vocal production," *J. Acoust. Soc. Am.* **73**, 45-50.
- Giles, H. and Powesland, P. E. (1975), "Speech style and social evaluation," London, England: Academic Press.
- Held, R. (1965), "Plasticity in sensory-motor systems," *Sci. Am.* **213**, 84-94.

- Jones, J. A. and Munhall, K. G. (2000), "Perceptual calibration of F0 production: Evidence from feedback perturbation," *J. Acoust. Soc. Am* **108**, 1246-1251.
- 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.
- Larson, C. R., Burnett, T. A., and Kiran, S. (2000), "Effects of pitch-shift velocity on voice F0 responses," *J. Acoust. Soc. Am.* **107**, 559-564.
- Natke, U., Grosser, J. and Kalveram, K. T. (2001), Fluency, fundamental frequency, and speech rate under frequency-shifted auditory feedback in stuttering and nonstuttering persons. *J. Fluency Disord.* **26**, 227-241.
- Natke, U. and Kalveram, K. T. (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 Pangert, 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., Iwarsson J., and Billström A. H. (1993), "Significance of mechanoreceptors in the subglottal mucosa for subglottal pressure control in singers," Presented at the 22nd Annual Symposium Care of the Professional Voice, Philadelphia.
- Sundberg, J. (1987), "The science of the singing voice," Dekalb, IL: Northern Illinois University Press.
- Tanabe M., Kitajima K., and Gould W. (1975), "Laryngeal phonatory reflex: The effect of anesthetization of the internal branch of the superior laryngeal nerve – Acoustic aspects," *Ann. Otol. Rhinol. Laryngol.* **84**, 206-212.

Collected Figure Captions

FIG. 1. Schematic illustration of the different steps during data processing. Step 1: The maximums of the first derivative of the digitized EGG-signal are used to assess the glottal closings. Step 2: The frequency of each glottal period, i.e. the momentary voice F_0 , is determined. The first glottal closing defines the onset of phonation. Step 3: A linear interpolation in a time pattern of 0.1 ms steps enables subsequent averaging of contours across trials and subjects. Step 4: PRE-, FAF-, and POST-trials are averaged for each subject. Step 5: The difference in cents between PRE- and FAF-trials as well as PRE- and POST-trials is determined. In this way, contours are created which reflect the variation based solely on frequency shift. Step 6: The contours are averaged across subjects.

FIG. 2. Deviation of voice F_0 contour of the first (long) syllable of the non-sense word [‘ta:tatas] during frequency shift from trials before frequency shift, averaged across all subjects. S.D. across all subjects were plotted as bars with an arbitrarily chosen interval of 10 ms to visualize the variation of the contour. 0 ms is the onset of phonation, defined by the first glottal closing. The contour begins at the point for which voice F_0 data becomes available for all subjects (see II-D for a more detailed description). The contour ends at the point for which n becomes less than 8. The two p -values are the results of Wilcoxon tests for interval deviations from 0. The latency of response was determined with the Castellan change-point test.

FIG. 3. Deviation of voice F_0 contour of the second (short) syllable of the non-sense word [‘ta:tatas] during frequency shift from trials before frequency shift. See caption for figure 1 for additional explanation.

FIG. 4. Deviation of voice F_0 contour of the first (long) syllable of the non-sense word [‘ta:tatas] after termination of frequency shift from trials before frequency shift. See caption for figure 1 for additional explanation.

FIG. 5. Deviation of voice F_0 contour of the second (short) syllable of the non-sense word [‘ta:tatas] after termination of frequency shift from trials before frequency shift. See caption for figure 1 for additional explanation.

FIG. 6. Data from all subjects, illustrating individual response magnitudes in the second (short) syllable during frequency shift and magnitude of after-effects at the beginning of the first (long) syllable after termination of frequency shift. Magnitudes during and after frequency shift were calculated as mean differences to PRE-trials over the interval 25-100 ms after onset of phonation.

Figure 1

Donath, T. M.

JASA

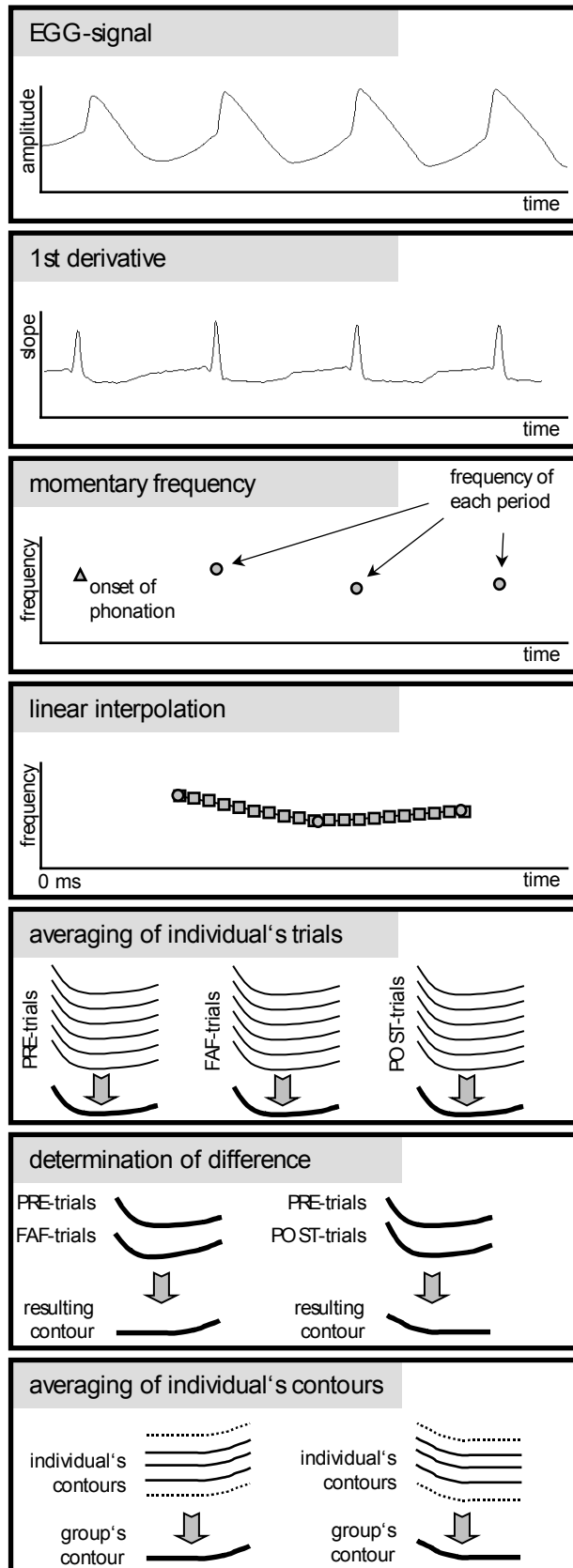


Figure 2

Donath, T. M.

JASA

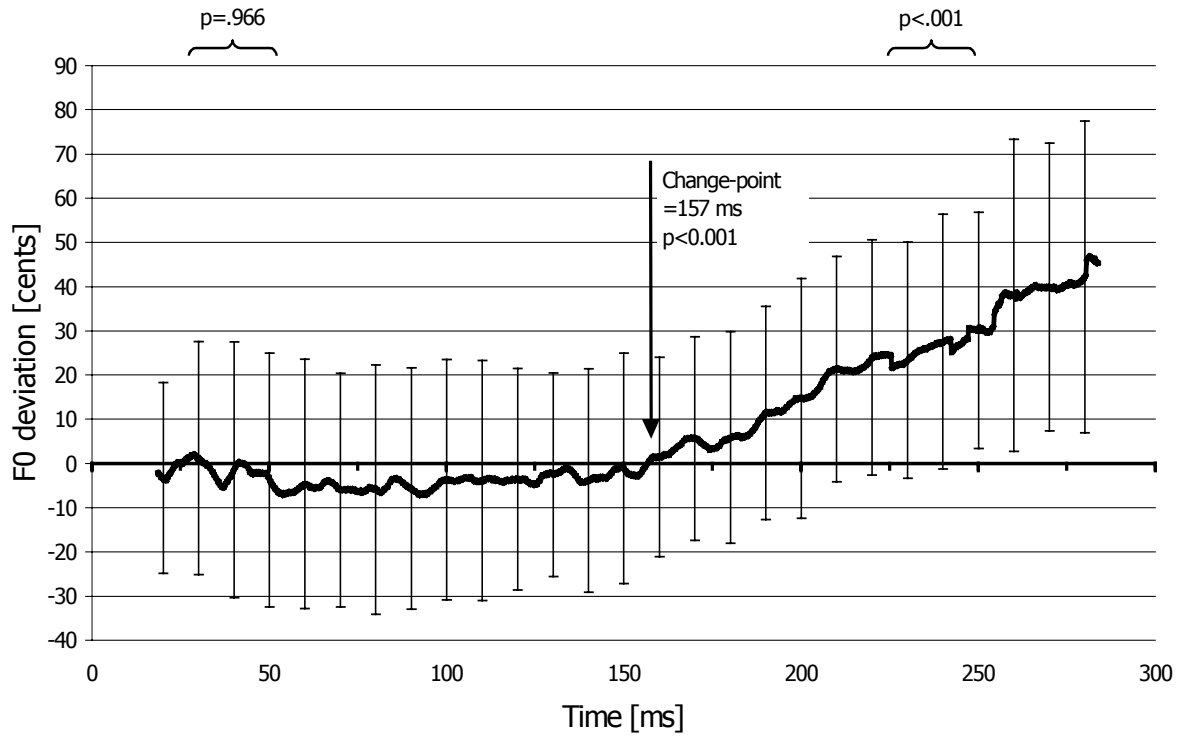


Figure 3

Donath, T. M.

JASA

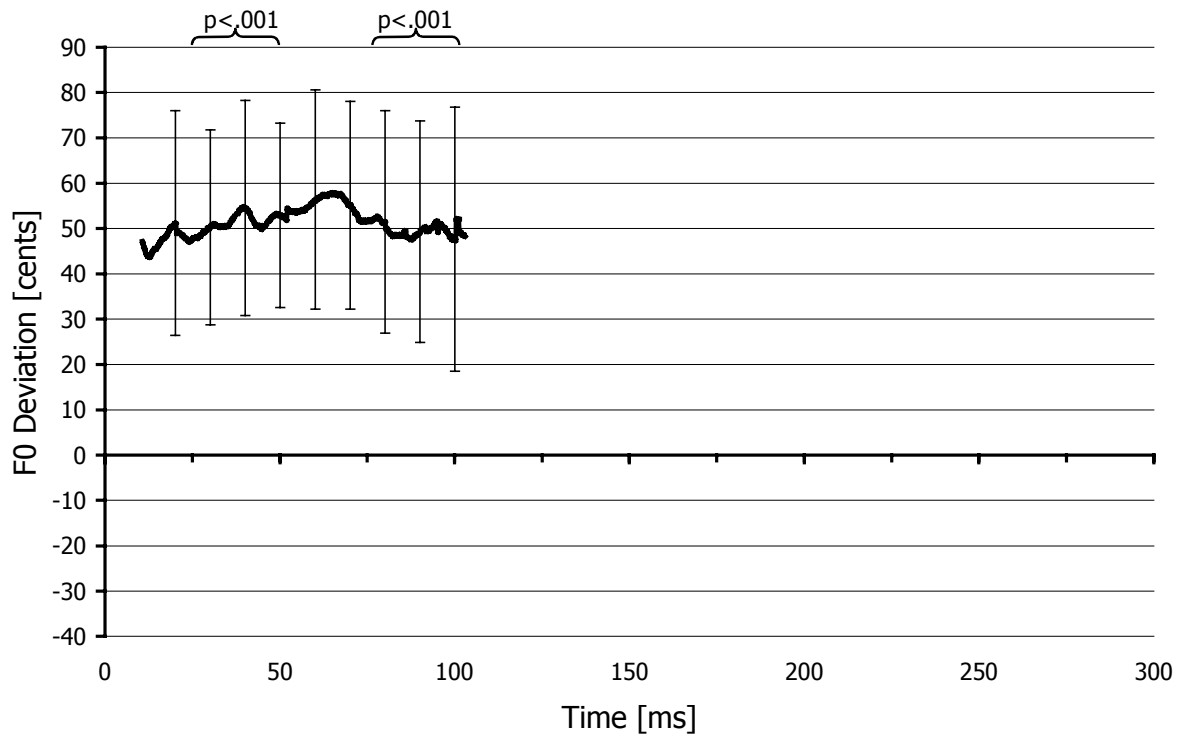


Figure 4

Donath, T. M.

JASA

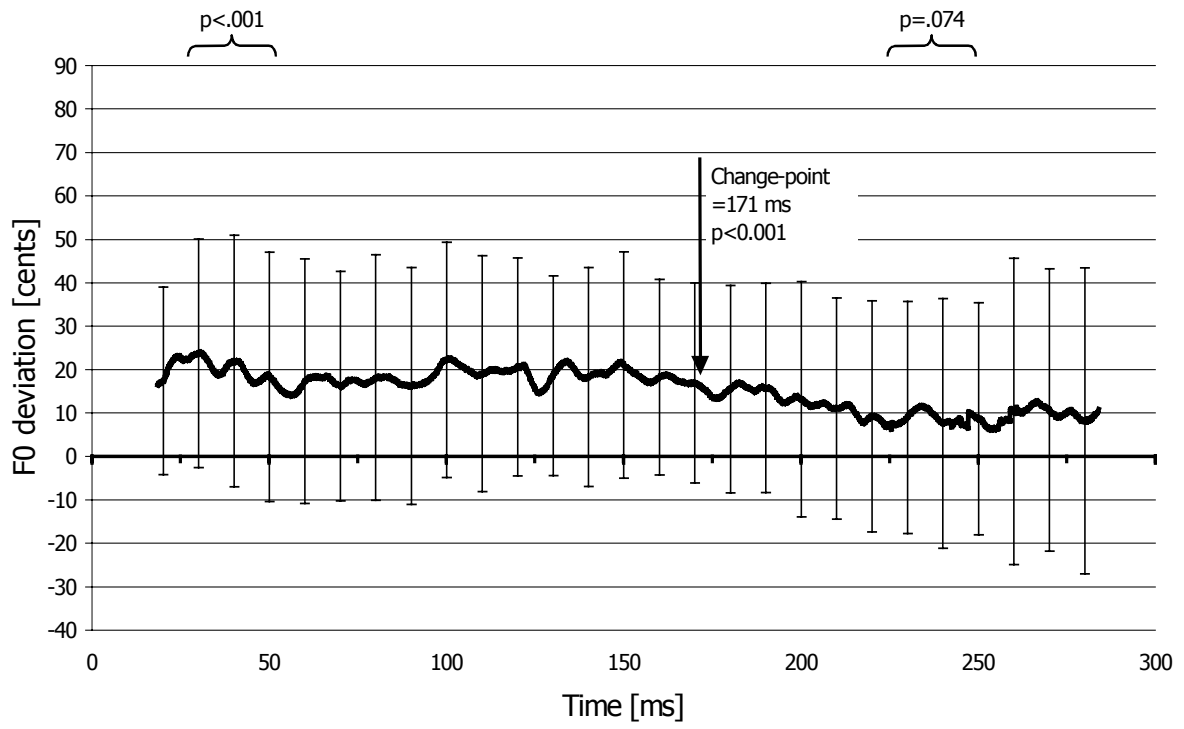


Figure 5

Donath, T. M.

JASA

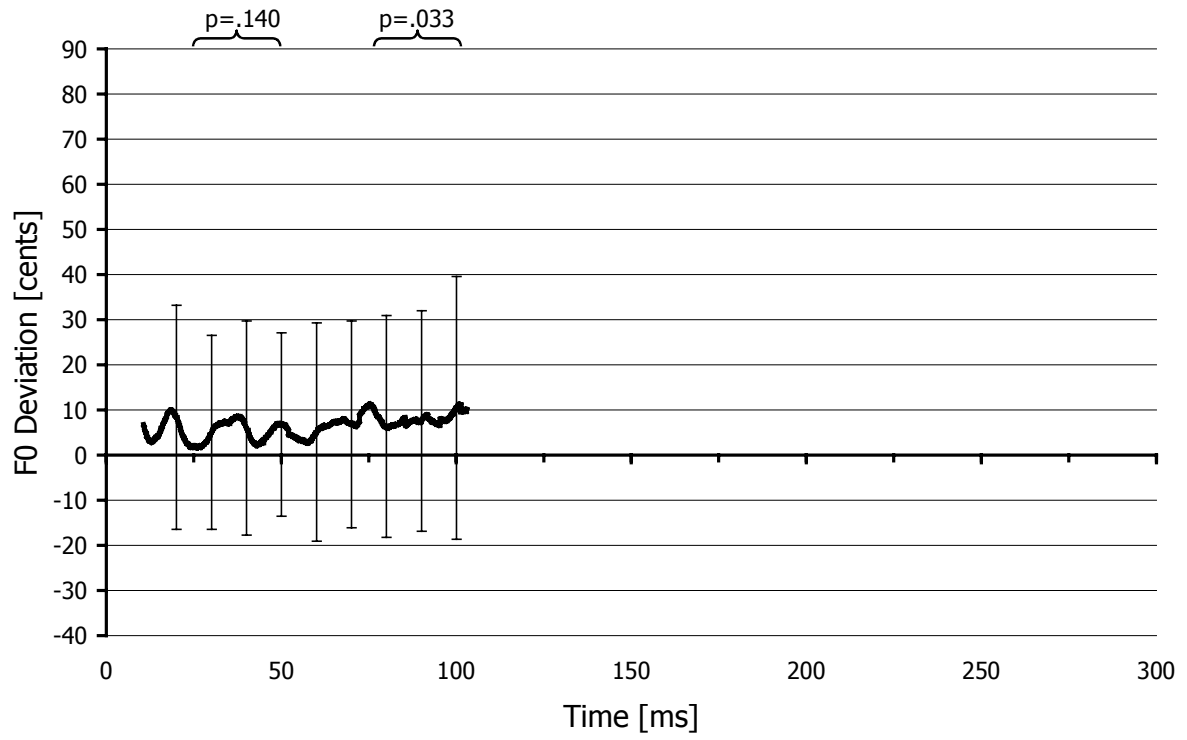


Figure 6

Donath, T. M.

JASA

