Articulatory compensation for low-pass filtered formant-altered auditory feedback

Yasufumi Uezu,[1] Sadao Hiroya,[1] and Takemi Mochida[1]

NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, 3-1, Morinosato-Wakamiya, Atsugi-shi, Kanagawa, 243-0198, Japan[a)]

The following article appeared in
Yasufumi Uezu, Sadao Hiroya, Takemi Mochida; Articulatory compensation for low-pass filtered formant-altered auditory feedback. J. Acoust. Soc. Am. 1 July 2021; 150 (1): 64–73.
and may be found at https://doi.org/10.1121/10.0004775 .

Auditory feedback while speaking plays an important role in stably controlling speech articulation. Its importance has been verified in formant-altered auditory feedback (AAF) experiments where speakers utter while listening to speech with perturbed first (F1) and second (F2) formant frequencies. However, the contribution of frequency components higher than F2 to the articulatory control under the perturbations of F1 and F2 has not yet been investigated. In this study, we conducted a formant AAF experiment where a low-pass filter was applied to speech. The experimental results showed that the deviation in the compensatory response was significantly larger when a low-pass filter with a cutoff frequency of 3-kHz was used, compared to that of 4- and 8-kHz. We also found that the deviation in 3 kHz condition correlated with the fundamental frequency and spectral tilt of the produced speech. Additional simulation results using a neurocomputational model of speech production (SimpleDIVA model) and our experimental data showed that feedforward learning rate increased as the cutoff frequency decreased. These results suggest that high-frequency components of the auditory feedback would be involved in the determination of corrective motor commands from auditory errors.

---

[a] uezu8223@gmail.com

I. INTRODUCTION

Sensory feedback is essential when humans are learning and performing complex movements. In speech production, auditory and somatosensory feedback plays an important role in the coordination of speech organs, by which fluency in speaking is achieved (Perkell *et al.*, 1997). Especially, the absence of auditory feedback significantly disrupts language acquisition in infants and online correction of speech errors. The mechanism behind auditory-based speech motor control is considered to be related to an auditory prediction associated with the speech motor command to be executed. Optimal speech motor coordination can be performed by evaluating and compensating the error between the predicted and actual auditory consequences. This requires a function to transform auditory errors into corrective motor commands.

Speech motor control has been investigated by exploring how speech production changes when acoustic features of speech collected from a microphone are perturbed and returned through headphones to the speaker. Some altered auditory feedback (AAF) experiments have demonstrated compensatory responses to the perturbation of vowel formant frequencies, where speakers' vowel production changes in the direction of reducing the formant frequency error between the intended target and the actual feedback (Houde and Jordan, 1998; Purcell and Munhall, 2006; Villacorta *et al.*, 2007). These studies have provided crucial evidence that the brain is capable of estimating the correct amount of articulatory compensation to be made from the auditory error. We have also experimentally confirmed the importance of accurate formant manipulation for the correct compensatory response in AAF experiments

(Uezu *et al.*, 2020). In parallel with these experimental developments, a neural network model simulating brain function in speech production, called directions into velocities of articulators, or DIVA, has been proposed to investigate the neural mechanisms underlying auditory feedback control during speech production (Guenther *et al.*, 2006). The validity of the model has been verified by fitting of formant AAF experimental data (Villacorta *et al.*, 2007) and by functional brain imaging (Tourville *et al.*, 2008).

Speech sounds are composed of information in various frequency regions. In the study of vowel production, F1 and F2 are considered to be goals of vowel production (Perkell *et al.*, 1997). In fact, many AAF experiments have perturbed F1 and F2 and investigated the compensatory responses. However, few vowel-production studies have focused on high-frequency components, although the shape of the vocal tract determines acoustic features such as F1 and F2 as well as high-frequency components. On the other hand, studies on vowel perception have investigated information in speech signals at different frequencies (Fig. 1). In general, lower-order formants such as F1 and F2 are parameters for determining the phonological properties of vowel sounds. Studies on vowel perception and classification have also shown that intelligibility is improved by considering F3 as well (Hillenbrand and Gayvert, 1993; Miller, 1989; Schwartz and Escudier, 1989). Other studies have found that high-frequency components around F3 are associated with speaker individuality (Kitamura and Akagi, 1995). Therefore, it is considered that the high-frequency components around F3 include phoneme-specific and speaker-specific information on vowel perception.

Conversion of auditory errors into corrective motor commands, which is specific to each speaker, is essential for compensation to occur, but it has a one-to-many mapping (Atal *et al.*, 1978; Hiroya and Honda, 2004). Therefore, if a higher frequency component than F1 and F2 is not included in the speech in formant AAF experiments, it is considered difficult to accurately estimate the corrective motor commands from the error in F1 and F2 in narrowband auditory input. However, it is unclear whether the compensatory response to perturbations to F1 and F2 is affected by the presence or absence of frequency components higher than F1 and F2.

In this study, we examined how high-frequency components of speech affect the auditory-motor control of vowel production by combining formant AAF and a low-pass filter. Speakers were asked to produce a syllable containing the Japanese vowel /e/. F1 and F2 are simultaneously perturbed toward the vowel /a/. 3, 4 and 8 kHz were used as cutoff frequencies of the low-pass filter. We used a phase equalization-based autoregressive exogenous model (PEAR) (Oohashi *et al.*, 2015) for its high formant estimation accuracy because the accuracy in the low-frequency component is important. The analysis was based on the magnitude and deviation of the projection of the compensatory response to the perturbation in the F1-F2 plane (Daliri and Dittman, 2019; Niziolek and Guenther, 2013).

The purpose of this study is to investigate whether the presence or absence of frequency components higher than F1 and F2 affect the one-to-many mapping in determining corrective motor commands from auditory errors in F1 and F2 during vowel production, irrespective of phoneme-specific or speaker-specific properties on vowel perception. The effect should be observed as a difference in the magnitude and deviation of the compensatory response depending on the cutoff frequency. Furthermore, in order to investigate the neural mechanism, formant AAF experimental data was fitted by a simplified version of the DIVA model

(SimpleDIVA model) (Kearney *et al.*, 2020) and the feedforward and feedback parameters were quantified. We also examined what specific features in the high-frequency components contributed to the compensatory responses by calculating the correlation of the magnitude and deviation of response with F3, the fundamental frequency (f_0) and spectral tilt.

II. EXPERIMENTS

A. Experimental procedure

The participants were 29 native Japanese speakers (nineteen females and ten males; average age, 37.9 years, standard deviation, 9.1, age range, 20-55 years). None of the speakers reported hearing or speech difficulties. All gave informed consent to participate in the study, which was approved by the NTT Communication Science Laboratories Research Ethics Committee.

The experiment was performed in a soundproof room. Figure 2 shows a block diagram of the formant altered auditory feedback used in this study. The speakers sat 20-cm from the microphone (SONY ECM-678/9X) and wore headphones (SENNHEISER HD280 Pro). The speech signal from the microphone was amplified (M-Audio DMP3), low-pass-filtered (MTT MS2319) at a cutoff frequency (Fc) of 8 kHz, A/D converted at 16 kHz, and transmitted to a real-time formant transformation system (Texas Instruments C66x) to generate formant-shifted speech signals. The altered speech signals were D/A converted, low-pass-filtered at a cutoff frequency of 3, 4, or 8 kHz (NF P-86, -135 dB/oct rolloff), amplified (Audio-Technica AT-MA 55), and presented via headphones with a delay of 16.5 ms. The uttered and altered speech was recorded on a PC using a DAQ device (National Instruments USB-6210). Speakers were encouraged to utter at a natural rate and level with timing controlled by a prompt on a monitor and instructed to close their lips at the end of each trial. Each prompt lasted 2 s, and the inter-trial interval was approximately 3.5 s.

Speakers were asked to produce the Japanese syllable /he/, which was shown on the monitor in hiragana (Japanese character) (Mitsuya *et al.*, 2011). Figure 3 shows perturbation patterns of F1 and F2 in a block including 140 utterance trials. A block consisted of four phases: Baseline (trials 1-20), Ramp (trials 21-70), Hold (trials 71-90), and Return (trials 91-140). In the Baseline phase, speakers produced utterances with no altered feedback. In the Ramp phase, the perturbation to the formant frequency increased linearly until reaching the maximum level of perturbation. In the Hold phase, speakers uttered while receiving speech with formant frequencies altered at the maximum level of perturbation. In the Return phase, speakers produced utterances with normal feedback, which was the same as in the Baseline phase. The maximum level of perturbation for formant frequencies was (F1, F2) = (+150, −300 Hz) (Martin *et al.*, 2018). The cutoff frequency of 3, 4, or 8 kHz for the altered speech was fixed within one block. All subjects participated under all experimental conditions. The order of the three cutoff frequency conditions was counterbalanced among the subjects. Masking noise was not used. The participants performed 30 trials for training. During these trials, calibration was performed so that the headphone output of the speech was 72 dBA SPL (ACO TYPE6240 and TYPE2015).

Studies on speech production while speakers listen to speech whose frequency band is limited by the low-pass filter have been carried out in the past (Burzynski and Starr, 1985;

Garber and Moller, 1979; Garber *et al.*, 1980, 1981; Peters, 1955). In those experiments without the AAF system, when a low-pass filter with a cutoff frequency of less than 1 kHz was applied to the speech, the speech became clear and nasalization of speech subsided. However, the cutoff frequencies were at most 1.8 kHz, and F2 may not have been included for front vowels such as /i/ and /e/. Therefore, in this study, the cutoff frequency of the low-pass filter was set to at least 3 kHz to include F2. We set cutoff frequencies of 4 kHz including F3 for all speakers and of 8 kHz, which is close to the frequency band of normal speech.

Speech signals were pre-emphasized by a first-order high-pass filter. Then, a 16-ms Blackman window was applied, and LPC coefficients were obtained every 8 ms by using the PEAR method. To estimate a time-stable spectrum, TANDEM windows were used (Oohashi *et al.*, 2015). The numbers of LPC coefficients (13 to 17) and taps of the phase equalization filter (9 to 28) for each of the speakers were determined by calibration (Uezu *et al.*, 2020; Vallabha and Tuller, 2004).

AAF experiments require formants to be estimated in real-time. Although linear predictive coding (LPC) (Itakura and Saito, 1970) is widely used in such experiments (Purcell and Munhall, 2006; Villacorta *et al.*, 2007), the estimated formant frequency is prone to errors (Oohashi *et al.*, 2015). To improve estimation accuracy, we have proposed the PEAR method (Oohashi *et al.*, 2015). Our previous studies have shown that the compensatory response to perturbations in PEAR is greater than that in LPC (Uezu *et al.*, 2020) For this experiment, a system with a sampling frequency of 16 kHz was developed on the basis of the PEAR algorithm (Oohashi *et al.*, 2015). In the conventional system (Oohashi *et al.*, 2015), the electroglottography (EGG) electrode is attached to the neck to extract the pitch mark corresponding to the glottal closure from the speech signal, whereas in the present system, instead of EGG, the SEDREAMS algorithm was used to extract the pitch mark (Drugman *et al.*, 2012).

B. Analysis

The formant analysis of the speech was performed using the PEAR method offline, and the median value from 40 to 80% of the vowel interval was used as a representative value for each trial. Compensatory responses in formant frequencies were determined by subtracting the value in the Baseline phase from that in the Hold phase. The Baseline and Hold values of the formant frequencies in each block were set by computing the mean value from Baseline (11-20 trials) (Munhall *et al.*, 2009) and Hold (71-90 trials) formants, respectively.

We not only evaluated the magnitude of the compensatory response of F1 and F2 independently, but also used the projection of the compensatory response to the perturbation in the F1-F2 plane (Daliri and Dittman, 2019; Niziolek and Guenther, 2013). Figure 4 shows a schematic diagram of the concept. The origin in the F1-F2 formant space is the formant frequencies of the subject at the baseline. The vector from the baseline to the mean value of the formant frequencies of the subject in the Hold phase is defined as the formant response vector $\overrightarrow{F_R}$. A perturbation vector from the origin to the maximum perturbation (+150, −300 Hz) is considered, and its inverse vector is defined as an “ideal” compensatory response vector $\overrightarrow{F_I}$. Here, the magnitude M was defined as the projection component from the response vector to the perturbation-compensation line, and the deviation response D

(Daliri and Dittman, 2019) was defined as the value of the perpendicular component:

$$M[Hz] = \frac{\overrightarrow{F_I} \cdot \overrightarrow{F_R}}{|\overrightarrow{F_I}|} \tag{1}$$

$$D[Hz] = \sqrt{|\overrightarrow{F_R}|^2 - M^2}. \tag{2}$$

A positive magnitude M means that compensation has been made for the perturbation, and the larger the magnitude, the greater the compensation. The deviation D must be greater than or equal to zero, and a larger deviation means that the compensation for the perturbation diverges from the ideal compensatory response vector. If the value of $F1/F2$ in $\overrightarrow{F_I}$ matches that in $\overrightarrow{F_R}$, then M is $|\overrightarrow{F_R}|$ and D is 0. Therefore, M and D are related through the ratio of the change in F1 and F2.

An opposite direction of the effect, i.e., a following response to the AAF, has been reported in recent studies (Vaughn and Nasir, 2015). While the compensatory response helps to reduce the acoustic error between the intended and the actual speech, the mechanism underlying the following response is not well understood. To statistically evaluate the following response to the perturbation, we calculated the ratio A of the magnitude of the compensatory response to the absolute value of the formant response (Fig. 4):

$$A = \frac{M}{|\overrightarrow{F_R}|}\ (-1 \leq A \leq 1). \tag{3}$$

In the cosine formula, $\cos^{-1}(A)$ is the angle in radians. If A is positive, the response is considered compensatory; and if it is negative, the response is considered following. In addition, if A is close to 1, the response is a more ideal compensatory response. This index is expected to correctly evaluate the ratio of compensation and following responses, independently of the magnitude of the compensatory response.

Although the DIVA model (Guenther *et al.*, 2006) can simulate the data of formant AAF experiments (Villacorta *et al.*, 2007), it has many parameters. Recently, a simplified model with only three parameters, SimpleDIVA (Kearney *et al.*, 2020), has been proposed. SimpleDIVA makes it possible to evaluate the reliance on auditory feedback by fitting experimental data. The parameters are auditory feedback gain α_A, somatosensory feedback gain α_S, and feedforward learning rate λ_{FF}, and the model is

$$\begin{aligned} y_{prod}(n) &= y_{FF}(n) + \Delta y_{FB}(n) \\ \Delta y_{FB}(n) &= \alpha_A \times (y_T(n) - y_{AF}(n)) \\ &\quad + \alpha_S \times (y_T(n) - y_{SF}(n)) \\ y_{FF}(n+1) &= y_{FF}(n) + \lambda_{FF} \times \Delta y_{FB}(n), \end{aligned}$$

where y is formant frequency and n is trial number. FF stands for feedforward, FB for feedback, *prod* for production, *AF* for auditory feedback, *SF* for somatosensory feedback, and T for target. The parameters were estimated from the time-series subject-averaged formant data of the AAF experiment for each cutoff frequency. Thus, data in the Baseline, Ramp and Return phases were also considered.

III. RESULTS

A. Response to perturbation

Table I shows the mean and standard deviation of F1, F2, f_0, and vowel duration at the baseline. An ANOVA showed that there was no significant difference between the cutoff frequencies for all of the feature values. This suggests that the cutoff frequency of the low-pass filter used in this study does not affect speech production if the auditory feedback is not perturbed.

Figure 5 shows the changes in the baseline of F1 and F2 in each trial, where the low-pass filter cutoff frequency was 3, 4, or 8 kHz. Note that corrections have been made so that the baseline mean is zero. In all cases, compensation for the perturbation in Fig. 3 appears in the Ramp and Hold phases, and it returns to zero in the Return phase (131-140 trials) in all conditions for 3 kHz ($t(28) = 0.98,\ p = 0.33$ for F1 and $t(28) = 1.70,\ p = 0.09$ for F2), 4 kHz ($t(28) = 0.02,\ p = 0.98$ for F1 and $t(28) = 0.24,\ p = 0.80$ for F2) and 8 kHz ($t(28) = -0.38,\ p = 0.70$ for F1 and $t(28) = 1.51,\ p = 0.14$ for F2).

Figure 6 shows response vectors $\overrightarrow{F_R}$ for each participant for the cutoff frequency condition Fc = (3, 4, 8) kHz on the F1-F2 plane. If a participant produces formant compensations in the ideal direction for the given perturbations, we would expect to see a response vector in the upper left direction along the perturbation-compensation line, and many response vectors indeed occurred in the upper left direction as expected. However, the magnitude and direction of the response vector were affected by individual differences. In addition, it was found that the directions of the response vectors in the 3-kHz condition tended to be more scattered compared with other cutoff frequency conditions.

Figures 7 and 8 show the compensatory responses for F1 and F2 for the cutoff frequency conditions. A one-sample t-test revealed that the absolute value of the compensatory response to baseline was significantly greater than 0 in all conditions for F1 ($t(28) = -4.33$ for 3 kHz, -5.45 for 4 kHz, -6.10 for 8 kHz, $p < 0.01$) and F2 ($t(28) = 5.71$ for 3 kHz, 7.05 for 4 kHz, 7.56 for 8 kHz, $p < 0.01$). An ANOVA showed that there was no significant difference between the cutoff frequencies in F1 ($F(2, 84) = 0.24,\ p = 0.78$) and F2 ($F(2, 84) = 0.07,\ p = 0.92$). Note that the dependent variable is the compensatory response of F1 or F2, the independent variable is the cutoff frequency, and repeated measure ANOVA was not used.

There was no difference between the cutoff frequencies when the compensatory responses of F1 and F2 were evaluated independently. We also examined the magnitude and deviation of the compensatory response in the F1-F2 plane. Figures 9 and 10 show the magnitude and the deviation of compensatory response for the cutoff frequency conditions, respectively. Note that a Shapiro-Wilk test showed that the distribution of the deviation in the compensatory response did not satisfy normality. A Kruskal-Wallis test revealed that there was no significant difference in the magnitude of the compensatory response between the conditions ($\chi^2(2) = 0.69,\ p = 0.70$). However, the deviation in the compensatory response tended to decrease when the cutoff frequency was high. Another Kruskal-Wallis test revealed that there was a significant difference in the deviation in the compensatory response between the conditions ($\chi^2(2) = 7.86,\ p = 0.01$). A two-tailed Mann-Whitney U test with Holm correction revealed that there were significant differences between the 3- and 4-kHz conditions (effect size (r) $= 0.46,\ p < 0.05$) and between the 3- and 8-kHz conditions ($r = 0.43,\ p < 0.05$), but not between the 4- and 8-kHz conditions ($r = 0.14,\ p = 0.43$). This indicates that the deviation in the compensatory response to the perturbation increased when the vowel was uttered while the participants listened to the speech through a low-pass filter having a cutoff frequency of 3 kHz. Note that, in all cutoff conditions, the magnitude and deviation of the compensatory response did not show significant differences between genders, except for a minor difference in the magnitude at 4 kHz ($p < 0.05$).

Figure 11 shows that the largest value of ratio A is at 8 kHz. A Shapiro-Wilk test showed that the distribution of A did not satisfy normality. A Kruskal-Wallis test showed that there was a difference in A between cutoff frequencies ($\chi^2(2) = 7.30,\ p = 0.02$). A two-tailed Mann-Whitney U test with Holm correction revealed that there was a significant difference between the 3 and 8 kHz conditions ($r = 0.47,\ p < 0.05$), but not between 4 and 8 kHz ($r = 0.24,\ p = 0.19$) and between 3 and 4 kHz ($r = 0.30,\ p = 0.19$), although the mean value was smaller for 4 kHz.

B. SimpleDIVA simulation

Table II shows the results estimated using SimpleDIVA (Version 1.3). In the SimpleDIVA study (Kearney *et al.*, 2020), λ_{FF}, which was primarily affected by data in the Ramp phase, ranged between 0.11 and 0.15, which are reasonable values for 8 kHz, but the values for 3 and 4 kHz were extremely large. Moreover, the magnitudes of the feedback gains α_A and α_S for 3 and 4 kHz were smaller than those for 8 kHz. The variation in the ratio of α_A and α_S, which determines the maximum amount of compensation in the Hold phase (Kearney *et al.*, 2020), between the cutoff frequencies was smaller than that of λ_{FF}.

C. Correlation with F3, f_0 and spectral tilt

High-frequency component of speech at 3 kHz or higher include not only the third and fourth higher formant frequencies derived from the vocal tract but also the harmonic components of the fundamental frequency and the spectral tilt characteristics derived from the glottal source. Therefore, eliminating the high-frequency components of the speech with a low-pass filter means that these pieces of source information are lost. We examined the correlation of the magnitude and deviation of the compensatory response at the cutoff frequency of 3 kHz with the values of F3, f_0, and the spectral tilt. The spectral tilt, which represents the slope of the source, was obtained from linear prediction coefficients of the first order (Wakita, 1973): the larger the coefficient, the steeper the tilt. These values were obtained from speech at a sampling frequency of 16 kHz during calibration.

Figure 12 shows that the magnitude in the compensatory response at the cutoff frequency of 3 kHz was not significantly correlated with F3 (correlation coefficient $(R) = 0.08,\ p = 0.67$) or with f_0 ($R = -0.15,\ p = 0.46$), but it was significantly correlated with spectral tilt ($R = -0.44\ p < 0.05$). Moreover, the figure shows that the deviation at 3 kHz was not significantly correlated with F3 ($R = 0.00,\ p = 0.99$), but it was marginally correlated with f_0 ($R = 0.33,\ p = 0.07$) and significantly correlated with the spectral tilt ($R = 0.47,\ p < 0.01$). Note that the magnitude and deviation at 4 and 8 kHz were not correlated with these values, except for a correlation between the magnitude at 4 kHz and F3 ($R = 0.37,\ p < 0.05$), and the magnitude and deviation in all cutoff conditions were not correlated with age, and there were significant differences in F3 and f_0 between genders ($p < 0.01$), but not in spectral tilt ($p = 0.07$).

IV. DISCUSSION

We investigated the effect of using a low-pass filter to cut the high-frequency components of speech on the compensatory response of formant AAF. The perturbations in this experiment were given to the F1 and F2 values, which were less than 3 kHz for all subjects. When low-pass filters with cutoff frequencies of 3, 4, and 8 kHz were used, although the perturbations for F1 and F2 were the same between conditions, the frequency components higher than F1 and F2 were not included in the 3 and 4 kHz conditions. The results of the experiment indicated that the deviation in the compensatory response at 3 kHz was significantly larger than that of 4 and 8 kHz, but that there was no significant difference between 4 and 8 kHz, although the magnitude of the compensatory response did not differ among the cutoff frequencies. The fact that the magnitude of the compensatory response was almost the same, but the deviation differed between conditions suggests that the same magnitude of compensatory response can be generated from different corrective motor commands. This corresponds to a redundancy in the acoustic-to-articulatory mapping (Atal *et al.*, 1978). Therefore, the absence of frequency components higher than 3 kHz increased the redundancy in the determination of corrective motor commands from the auditory errors of F1 and F2, resulting in an increase in the deviation of compensatory response. These findings suggest that the corrective motor commands for the magnitude of the auditory errors can be determined precisely from only the errors contained in the auditory feedback regardless of the presence or absence of high-frequency components and that the difference in deviation between conditions results from the redundancy in the determination of corrective motor commands. In other words, compensation for the magnitude of perturbations is a task for auditory-motor control of vowel production.

SimpleDIVA modeling results showed that the feedforward learning rate increased with decreasing cutoff frequencies. This suggests that the change of feedforward control caused by low-pass filtering may affect the redundancy in the determination of corrective motor commands. Daliri and Dittman (2019) examined the effect of the reliance on auditory feedback on the compensatory response by directly comparing compensatory responses to perturbations of F1 and F2 for task-relevant errors under the formant shift condition with task-irrelevant errors under the formant clamp condition. Our study is similar to Daliri and Dittman (2019) in that it varied the reliance on auditory feedback, but it differs from that study (Daliri and Dittman, 2019) in that the magnitude and deviation became smaller when the reliance on auditory feedback was low. This is due to differences in the experimental methods: While Daliri and Dittman (2019) varied the reliance on auditory feedback depending on whether F1 and F2 was controlled, we changed it depending on whether or not there was a high-frequency component. We speculated that the difference in reliance on auditory feedback between the formant shift condition and the formant clamp condition of Daliri and Dittman (2019) was larger than that between our cutoff frequencies. Therefore, since the reliance on auditory feedback was too low in Daliri and Dittman (2019), the dependence on feedforward control would increase, and there would be little compensatory response.

These results indicate that the absence of high-frequency components in the feedback speech changed not only the redundancy in determining corrective motor commands but also the reliance on auditory feedback. However, the relationship between the changes remains unclear. This is because SimpleDIVA, unlike full DIVA (Guenther *et al.*, 2006), does not require F3 input and does not convert auditory errors into corrective motor commands, which would be necessary to quantify the amount of redundancy. Also, the results indicating that the deviation was different were obtained from the compensatory response in the Hold phase, while the results of SimpleDIVA were obtained from the formant data of all 140 trials. We speculate that an increase in redundancy caused a decrease in the reliance on auditory feedback, but further investigation will be needed to confirm it.

From the experimental results, we speculated that there is a spectral feature of speech that changes the deviation in the compensatory response between 3 and 4 kHz. Although F3, which is considered to be related to speech intelligibility (Hillenbrand and Gayvert, 1993; Miller, 1989; Schwartz and Escudier, 1989) and speaker individuality (Kitamura and Akagi, 1995) on vowel perception, may be involved, it is unlikely to be responsible for the deviation, because about half of the subjects had F3 values less than 3 kHz and there was no significant correlation with F3. Therefore, this suggests that the increase in deviation at the 3-kHz cutoff frequency is not related to the deterioration of vowel identification or loss of speaker individuality on vowel perception due to the removal of F3, and that the deviation in the compensatory response is affected by the presence or absence of F4 or higher features in the frequency band from 3 to 4 kHz. It has been reported that F4 is related to the hypopharyngeal cavities of the speaker and that inter-speaker variation of the cavity is large, but intra-speaker variation is small (Kitamura *et al.*, 2005). This suggests that speaker-specific information on vowel production may affect the reliance on auditory feedback, but clarifying whether this is the case would entail experiments involving, e.g., direct manipulation of speaker-specific information (Toyomura and Omori, 2005; Zheng *et al.*, 2011).

The results in Sec. III C showed that the magnitude and deviation in the compensatory response at 3 kHz was correlated with f_0 and spectral tilt. This indicates that the source characteristics of the high-frequency components, which have been rarely considered in previous studies, may play an important role in auditory feedback during speech production.

Perceptual experiments have shown that the vowel formant frequency discrimination threshold is reduced when the fundamental frequency is low (Kewley-Port *et al.*, 1996) and that speech intelligibility increases in noisy environments as the spectral tilt becomes flatter (Simantiraki *et al.*, 2020). This may be because lowering f_0 increases the number of harmonics of f_0 contained in the low-pass filtered speech and because decreasing the spectral tilt increases the amplitude of the formant frequency of the high-frequency components. Thus, those findings may be related to our results, but it will be necessary to directly investigate the effect of source characteristics on auditory feedback because the harmonics of f_0 and the spectral tilt are also included below 3 kHz.

The results on the ratio A of the magnitude of the compensatory response to the absolute value of the formant response suggest that the lower the cutoff frequency is, the more likely it is that the following response will occur. We speculate that lowering the reliance on auditory feedback may cause the following response. Recent f_0 perturbation experiments claimed that the participants' feeling of being "externally driven" is the cause of the following response (Franken *et al.*, 2018, 2019b), but it remains to be seen whether they would feel that way when the cutoff frequency is lowered.

SimpleDIVA has limitations making it hard to use it to model data from unexpected perturbations (Kearney *et al.*, 2020) and following responses. Although the gradual perturbation was used as the AAF in this experiment, it is known that the feedforward command is unable to adapt when an unexpected perturbation is given (Franken *et al.*, 2019a). The effect of a cutoff frequency in the case of an unexpected perturbation remains to be studied.

Subjects may receive bone-conducted auditory feedback even if a low-pass filter was applied. However, if they were strongly affected by such an effect, they should have shown the same compensatory responses across cutoff frequency conditions. Also, as with the 4 kHz and 8 kHz conditions, there should be no correlation between deviation and spectral tilt under the 3 kHz condition. However, the experimental results in 3 kHz condition were different from those in 4 and 8 kHz conditions. Therefore, we believe that the effect of bone-conducted auditory feedback was not so large as to change the experimental results.

As in this experiment, Daliri and Dittman (2019) changed the reliance on auditory feedback in AAF experiments, but did not measure reliance itself directly. This is because the reliance on auditory feedback cannot be easily measured as intelligibility in speech perception. In other words, the reliance on auditory feedback is difficult to measure simply by having participants passively listen to low-pass filtered sound. Note that passive listening is a passive input of speech to the auditory system. On the other hand, active listening means monitoring one's own voice while speaking. These differences are related to the presence or absence of prediction when listening to speech, and the reliance on auditory feedback can be evaluated by detecting the difference between predicted speech and actual feedback. The mechanisms of passive and active listening are known to differ, and they are distinguished by differences in brain activity in speaking-induced auditory suppression (Curio *et al.*, 2000) and from classification of AAF data using convolutional neural networks to verify the importance of speech prediction (Taguchi *et al.*, 2020). Therefore, quantifying reliance on auditory feedback requires experiments with active listening, such as rating reliance during AAF experiments and measuring physiological indices. However, attention should be paid to the possibility that doing so imposes a burden on the subject and may affect the formant AAF experiment itself due to a dual task.

It is known that the elderly tend to lose the ability to hear high-frequency sounds (Cruickshanks *et al.*, 1998). In light of our results, it would be interesting to examine the relationship between speech production and hearing in the elderly. Studies have shown that auditory acuity can predict the compensatory response of AAF (Martin *et al.*, 2018; Villacorta *et al.*, 2007), but they did not measure acuity under active listening conditions. A better understanding could be gained if the difficulties in conducting perceptual experiments in active listening can be overcome in the future.

V. CONCLUSION

We conducted a formant AAF experiment in which the high-frequency components of speech were removed by low-pass filtering. The experimental results showed that the deviation in the compensatory response was significantly larger than that for 4 and 8 kHz when a low-pass filter with a cutoff frequency of 3 kHz was used, but the magnitude did not differ between the cutoff frequencies. A simulation using a SimpleDIVA model found that the feedforward control became dominant when the cutoff frequency decreased. These results indicate that increased redundancy in the determination of corrective motor commands from auditory errors and decreased reliance on auditory feedback due to the absence of high-frequency components affected the compensatory response of F1 and F2 in the AAF experiment. Further analysis suggested that the presence or absence of F4 or higher in the 3 to 4 kHz frequency band, f_0, and the spectral tilt of the glottal source signal are responsible for the increasing deviations. In the future, it will be necessary to examine a method of measuring the reliance on auditory feedback simultaneously during formant AAF experiments.

ACKNOWLEDGMENTS

We thank Dr. Shigeto Furukawa for his valuable comments on the manuscript and Jeffrey K. Hara of Stanford University for the development of the DSP system.

Atal, B., Chang, J., Mathews, M., and Tukey, J. (**1978**). “Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique.,” The Journal of the Acoustical Society of America **63**(5), 1535–1553.

Burzynski, C. M., and Starr, C. D. (**1985**). “Effects of feedback filtering on nasalization and self-perception of nasality,” Journal of Speech, Language, and Hearing Research **28**(4), 585–588.

Cruickshanks, K. J., Wiley, T. L., Tweed, T. S., Klein, B. E., Klein, R., Mares-Perlman, J. A., and Nondahl, D. M. (**1998**). “Prevalence of Hearing Loss in Older Adults in Beaver

Dam, Wisconsin: The Epidemiology of Hearing Loss Study," American Journal of Epidemiology **148**(9), 879–886.

Curio, G., Neuloh, G., Numminen, J., Jousmäki, V., and Hari, R. (**2000**). "Speaking modifies voice-evoked activity in the human auditory cortex," Human Brain Mapping **9**(4), 183–191.

Daliri, A., and Dittman, J. (**2019**). "Successful auditory motor adaptation requires task-relevant auditory errors," Journal of Neurophysiology **122**, doi: `10.1152/jn.00662.2018`.

Drugman, T., Thomas, M., Gudnason, J., Naylor, P., and Dutoit, T. (**2012**). "Detection of glottal closure instants from speech signals: A quantitative review," IEEE Transactions on Audio, Speech, and Language Processing **20**(3), 994–1006.

Franken, M. K., Acheson, D. J., McQueen, J. M., Hagoort, P., and Eisner, F. (**2018**). "Opposing and following responses in sensorimotor speech control: Why responses go both ways," Psychonomic bulletin & review **25**(4), 1458–1467.

Franken, M. K., Acheson, D. J., McQueen, J. M., Hagoort, P., and Eisner, F. (**2019**a). "Consistency influences altered auditory feedback processing," Quarterly Journal of Experimental Psychology **72**(10), 2371–2379, doi: `10.1177/1747021819838939`.

Franken, M. K., Hartsuiker, R. J., Johansson, P., Hall, L., Wartenberg, T., and Lind, A. (**2019**b). "Does passive sound attenuation affect responses to pitch-shifted auditory feedback?," The Journal of the Acoustical Society of America **146**(6), 4108–4121, doi: `10.1121/1.5134449`.

Garber, S. R., and Moller, K. T. (**1979**). "The effects of feedback filtering on nasalization in normal and hypernasal speakers," Journal of Speech, Language, and Hearing Research

22(2), 321–333.

Garber, S. R., Siegel, G. M., and Pick Jr, H. L. (**1980**). “The effects of feedback filtering on speaker intelligibility,” Journal of communication disorders **13**(4), 289–294.

Garber, S. R., Siegel, G. M., and Pick Jr, H. L. (**1981**). “Regulation of vocal intensity in the presence of feedback filtering and amplification,” Journal of Speech, Language, and Hearing Research **24**(1), 104–108.

Guenther, F. H., Ghosh, S. S., and Tourville, J. A. (**2006**). “Neural modeling and imaging of the cortical interactions underlying syllable production,” Brain and language **96**(3), 280–301.

Hillenbrand, J., and Gayvert, R. (**1993**). “Vowel classification based on fundamental frequency and formant frequencies,” Journal of speech and hearing research **36**, 694–700, doi: `10.1044/jshr.3604.694`.

Hiroya, S., and Honda, M. (**2004**). “Estimation of articulatory movements from speech acoustics using an hmm-based speech production model,” IEEE Transactions on Speech and Audio Processing **12**(2), 175–185, doi: `10.1109/TSA.2003.822636`.

Houde, J. F., and Jordan, M. I. (**1998**). “Sensorimotor adaptation in speech production,” Science **279**(5354), 1213–1216.

Itakura, F., and Saito, S. (**1970**). “A statistical method for estimation of speech spectral density and formant frequencies,” Electronics and Communications in Japan, A **53**(1), 36–43.

Kearney, E., Nieto-Castañón, A., Weerathunge, H. R., Falsini, R., Daliri, A., Abur, D., Ballard, K. J., Chang, S.-E., Chao, S.-C., Heller Murray, E. S. *et al.* (**2020**). “A simple

3-parameter model for examining adaptation in speech and voice production," Frontiers in Psychology **10**, 2995.

Kewley-Port, D., Li, X., Zheng, Y., and Neel, A. T. (**1996**). "Fundamental frequency effects on thresholds for vowel formant discrimination," The Journal of the Acoustical Society of America **100**(4), 2462–2470.

Kitamura, T., and Akagi, M. (**1995**). "Speaker individualities in speech spectral envelopes," Journal of the Acoustical Society of Japan (E) **16**(5), 283–289, doi: 10.1250/ast.16.283.

Kitamura, T., Honda, K., and Takemoto, H. (**2005**). "Individual variation of the hypopharyngeal cavities and its acoustic effects," Acoustical Science and Technology **26**(1), 16–26, doi: 10.1250/ast.26.16.

Martin, C. D., Niziolek, C. A., Duñabeitia, J. A., Perez, A., Hernandez, D., Carreiras, M., and Houde, J. F. (**2018**). "Online Adaptation to Altered Auditory Feedback Is Predicted by Auditory Acuity and Not by Domain-General Executive Control Resources," Frontiers in Human Neuroscience **12**.

Miller, J. (**1989**). "Auditory-perceptual interpretation of the vowel," The Journal of the Acoustical Society of America **85**, 2114–34, doi: 10.1121/1.2024119.

Mitsuya, T., Macdonald, E., Purcell, D., and Munhall, K. (**2011**). "A cross-language study of compensation in response to real-time formant perturbation," The Journal of the Acoustical Society of America **130**, 2978–86, doi: 10.1121/1.3643826.

Munhall, K., MacDonald, E., Byrne, S., and Johnsrude, I. (**2009**). "Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate," The Journal of the Acoustical Society of America **125**, 384–90, doi: 10.

1121/1.3035829.

Niziolek, C. A., and Guenther, F. H. (**2013**). “Vowel Category Boundaries Enhance Cortical and Behavioral Responses to Speech Feedback Alterations,” Journal of Neuroscience **33**(29), 12090–12098.

Oohashi, H., Hiroya, S., and Mochida, T. (**2015**). “Real-time robust formant estimation system using a phase equalization-based autoregressive exogenous model,” Acoustical Science and Technology **36**(6), 478–488.

Perkell, J., Matthies, M., Lane, H., Guenther, F., Wilhelms-Tricarico, R., Wozniak, J., and Guiod, P. (**1997**). “Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models,” Speech communication **22**(2-3), 227–250.

Peters, R. W. (**1955**). “The effect of filtering of side-tone upon speaker intelligibility,” Journal of Speech and Hearing Disorders **20**(4), 371–375.

Purcell, D. W., and Munhall, K. G. (**2006**). “Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation,” The Journal of the Acoustical Society of America **120**(2), 966–977.

Schwartz, J. L., and Escudier, P. (**1989**). “A strong evidence for the existence of a large-scale integrated spectral representation in vowel perception,” Speech communication **8**(3), 235–259.

Simantiraki, O., Cooke, M., and Pantazis, Y. (**2020**). “Effects of spectral tilt on listeners’ preferences and intelligibility,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6254–6258.

Taguchi, F., Hiroya, S., Uezu, Y., and Mochida, T. (**2020**). “Classification of formant transformed auditory feedback speech using convolutional neural networks,” Acoustical Science and Technology **41**(5), 800–803, doi: 10.1250/ast.41.800.

Tourville, J. A., Reilly, K., and Guenther, F. H. (**2008**). “Neural mechanisms underlying auditory feedback control of speech,” NeuroImage **39**(3), 1429–1443.

Toyomura, A., and Omori, T. (**2005**). “Auditory feedback control during a sentence-reading task: Effect of other’s voice,” Acoustical science and technology **26**(4), 358–361.

Uezu, Y., Hiroya, S., and Mochida, T. (**2020**). “Vocal-tract spectrum estimation method affects the articulatory compensation in formant transformed auditory feedback,” Acoustical science and technology **41**(5), 720–728, doi: 10.1250/ast.41.720.

Vallabha, G., and Tuller, B. (**2004**). “Choice of filter order in lpc analysis of vowels,” in *From Sound to Sense*, pp. 203–208.

Vaughn, C., and Nasir, S. M. (**2015**). “Precise feedback control underlies sensorimotor learning in speech,” Journal of Neurophysiology **113**(3), 950–955.

Villacorta, V. M., Perkell, J. S., and Guenther, F. H. (**2007**). “Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception,” The Journal of the Acoustical Society of America **122**(4), 2306–2319.

Wakita, H. (**1973**). “Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms,” IEEE Transactions on Audio and Electroacoustics **21**(5), 417–427.

Zheng, Z. Z., MacDonald, E. N., Munhall, K. G., and Johnsrude, I. S. (**2011**). “Perceiving a stranger’s voice as being one’s own: A ‘rubber voice’ illusion?,” PLOS ONE **6**(4), 1–8, doi: 10.1371/journal.pone.0018655.

Fc (kHz)	F1 (Hz)	F2 (Hz)	f_0 (Hz)	Duration (ms)
3	571 (107)	2239 (260)	206 (51)	446 (275)
4	573 (99)	2235 (240)	208 (49)	482 (296)
8	574 (99)	2249 (257)	207 (49)	455 (289)
F(2,84)	0.00	0.02	0.00	0.12
p	0.99	0.97	0.99	0.88

TABLE I. Mean, standard deviation, F-value and p-value of F1, F2, f_0, and vowel duration at baseline.

Fc (kHz)	α_A	α_S	λ_{FF}	α_A/α_S	r
3	0.04	0.13	0.84	0.30	0.95
4	0.05	0.12	0.51	0.41	0.96
8	0.13	0.36	0.12	0.36	0.96

TABLE II. Results of parameter fitting to simple DIVA model. α_A, α_S, λ_{FF} and r are auditory feedback gain, somatosensory feedback gain, feedforward learning rate, and Pearson's correlation coefficient, respectively.

FIG. 1. Spectrogram of vowel /e/ uttered by a female native Japanese speaker.

FIG. 2. Block diagram of altered auditory feedback in this study.

FIG. 3. Perturbation patterns of first and second formants in an experimental block. One block contains four phases: Baseline (trials 1-20), Ramp (trials 21-70), Hold (trials 71-90), and Return (trials 91-140).

FIG. 4. Formant response vector and its magnitude and error in F1-F2 plane.

FIG. 5. Patterns of formant frequency change for the baseline at each cutoff frequency condition. (Top) F2. (Bottom) F1. Shaded regions denote the standard error.

FIG. 6. Formant response vectors for each participant on F1-F2 plane. The number written beside each response vector indicates the ID of the participant, and the color of each response vector indicates the condition of the cutoff frequency: red, Fc = 3 kHz; yellow, Fc = 4 kHz; blue, Fc = 8 kHz. The dotted line is a straight line through the origin and the maximum perturbation (+150, −300 Hz). The origin represents the baseline.

FIG. 7. Box-plot of compensatory responses for the first format frequency (F1) for cutoff frequency conditions Fc = (3, 4, 8) kHz. Lower and upper error lines indicate minimum (Q1−1.5*IQR) and maximum (Q3+1.5*IQR), respectively. Each circle above or below those top lines stands for an outlier.

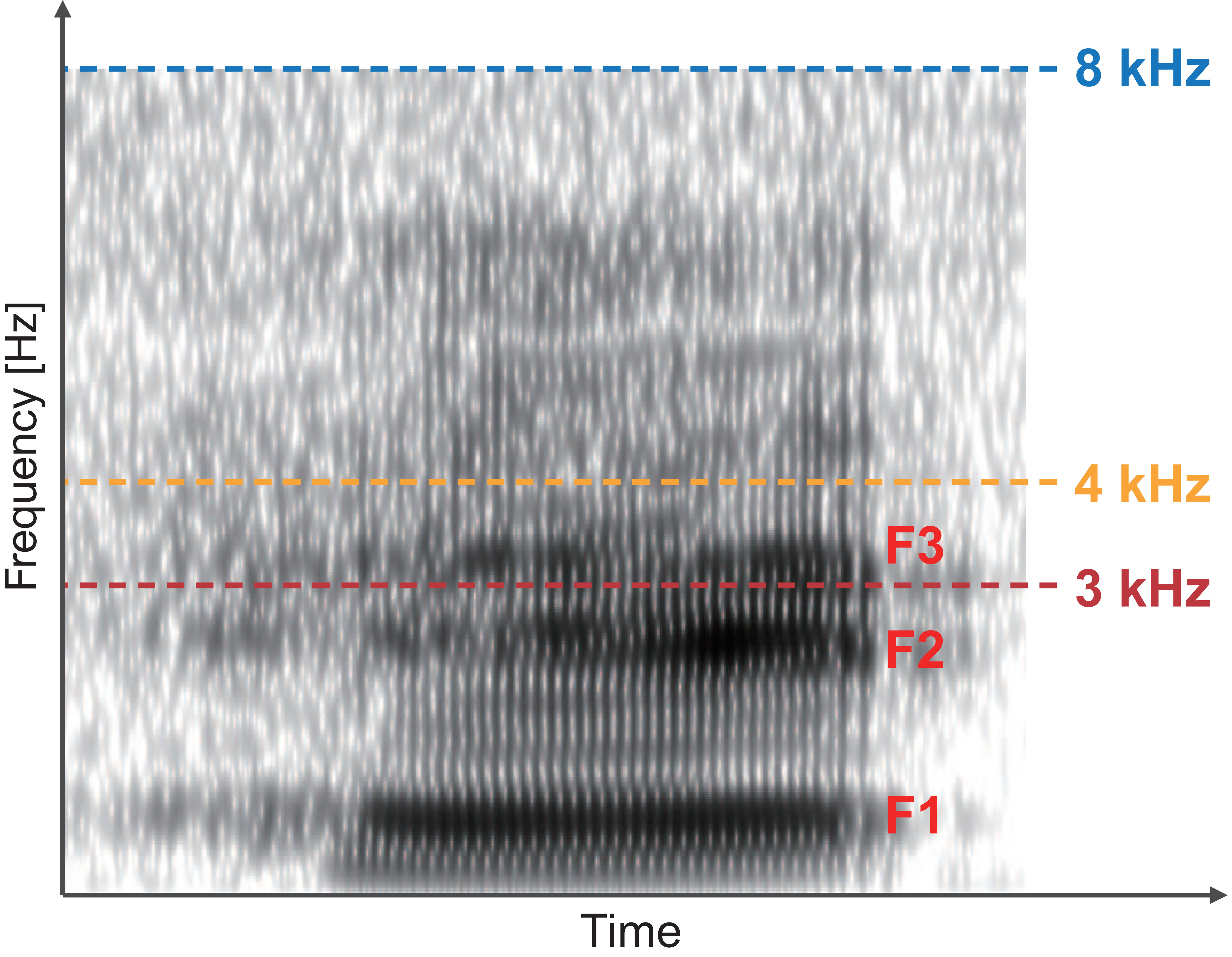
FIG. 8. Box-plot of compensatory responses for the second format frequency (F2) for for cutoff frequency conditions Fc = (3, 4, 8) kHz. Lower and upper error lines indicate minimum (Q1−1.5*IQR) and maximum (Q3+1.5*IQR), respectively. Each circle above or below those top lines stands for an outlier.

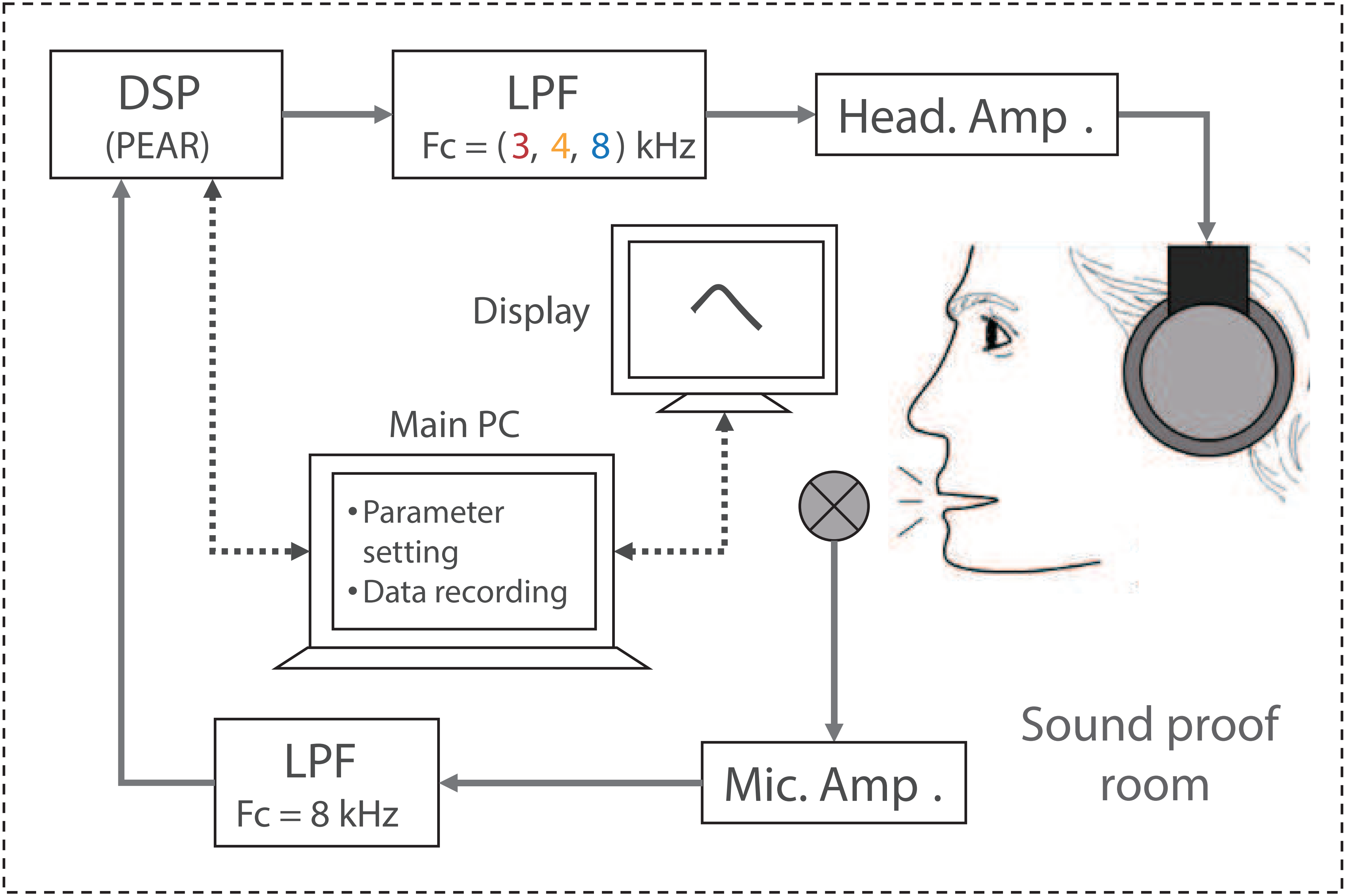
FIG. 9. Box-plot of magnitude M in compensatory response for cutoff frequency conditions Fc = (3, 4, 8) kHz. Lower and upper error lines indicate minimum (Q1−1.5*IQR) and maximum (Q3+1.5*IQR), respectively. Each circle above or below those top lines stands for an outlier.

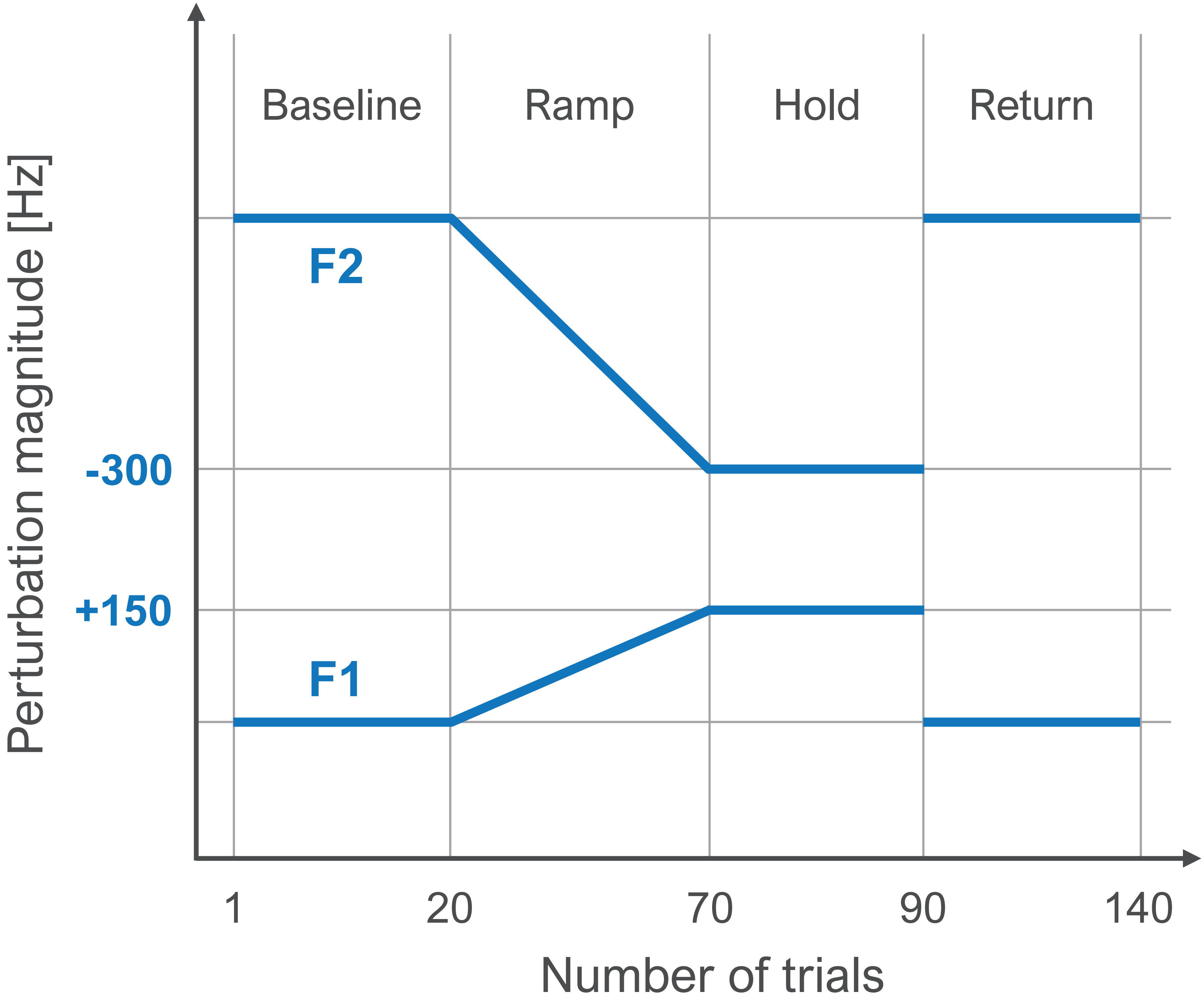
FIG. 10. Box-plot of deviation D in compensatory response for cutoff frequency conditions Fc = (3, 4, 8) kHz. Lower and upper error lines indicate minimum (Q1−1.5*IQR) and maximum (Q3+1.5*IQR), respectively. Each circle above or below those top lines stands for an outlier.

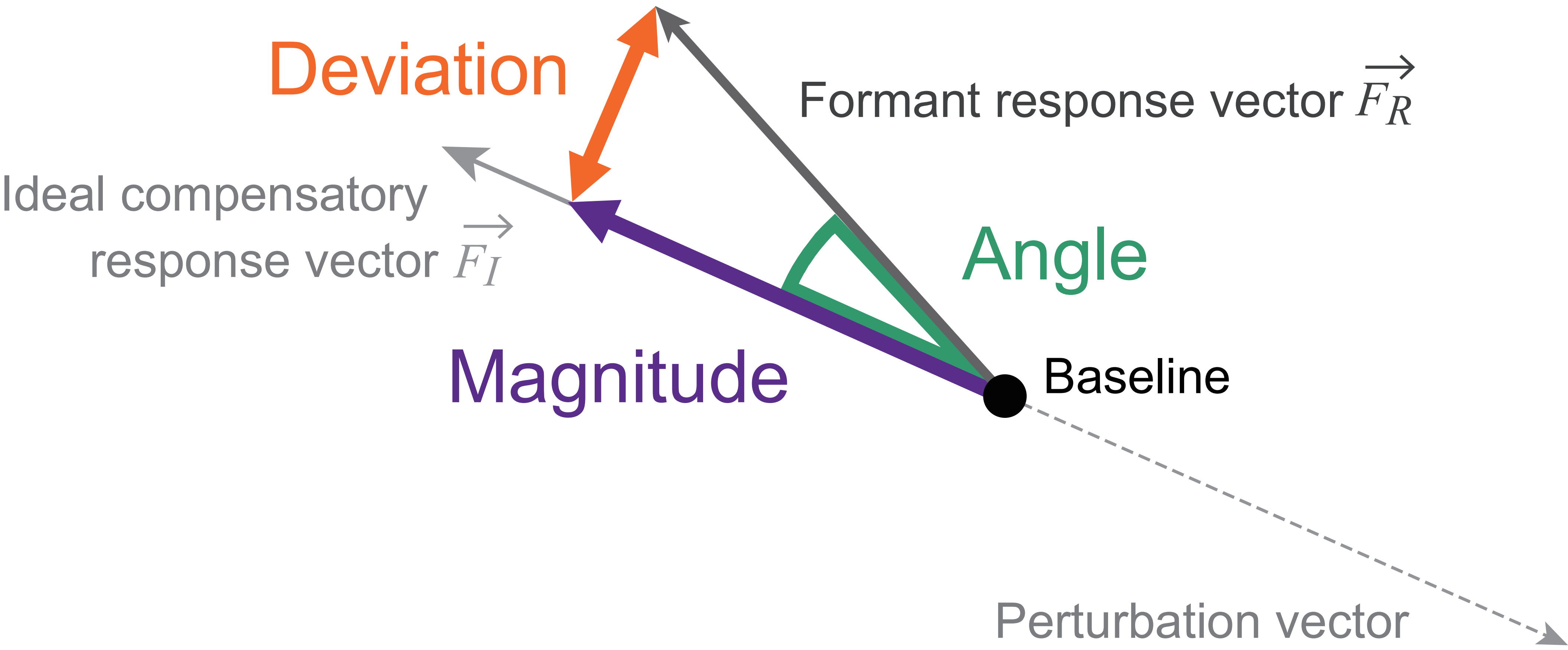
FIG. 11. Box-plot of ratio A of the magnitude of the compensatory response to the absolute value of the formant response for cutoff frequency conditions Fc = (3, 4, 8) kHz. Lower and upper error lines indicate minimum (Q1−1.5*IQR) and maximum (Q3+1.5*IQR), respectively. Each circle above or below those top lines stands for an outlier.

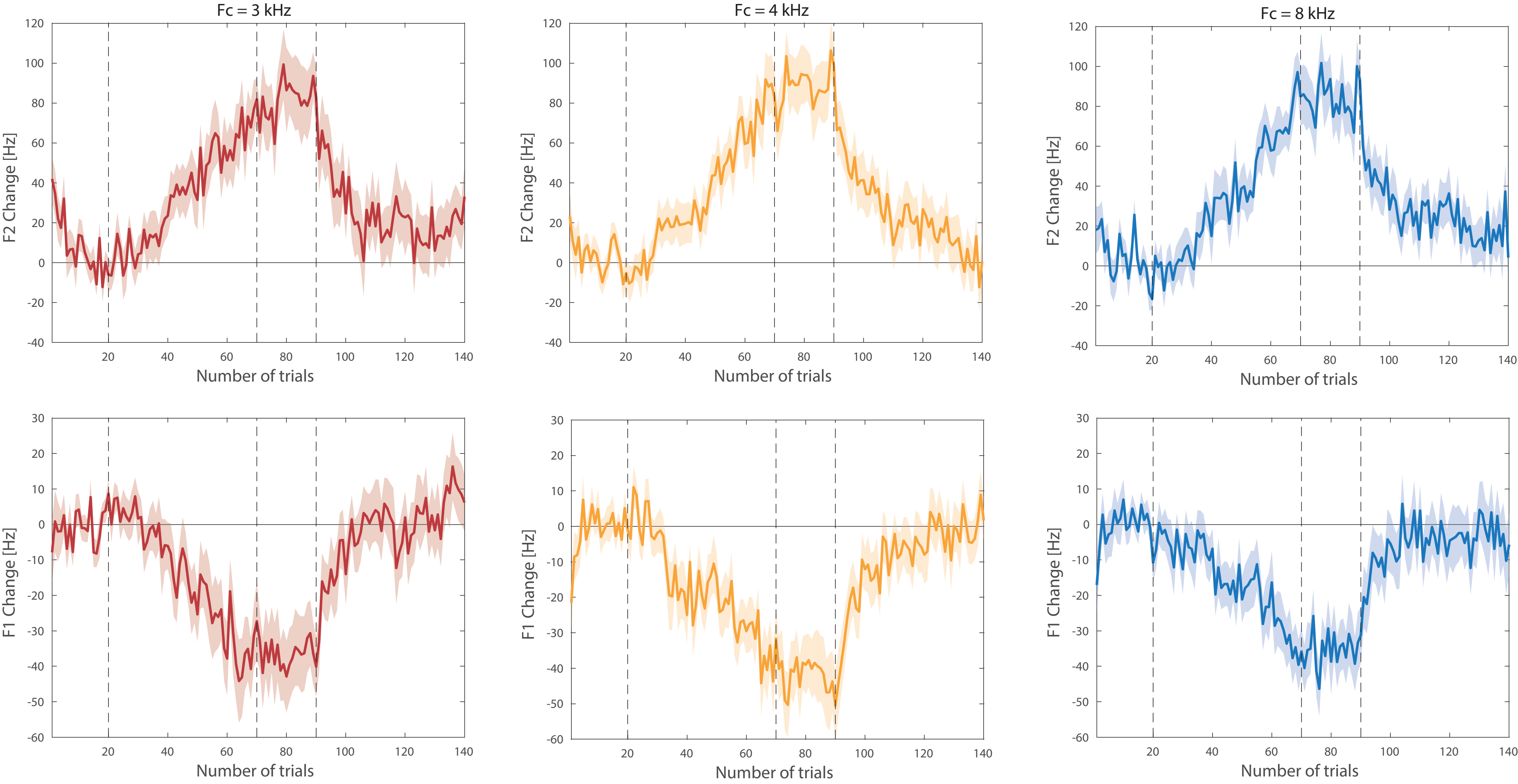
FIG. 12. Correlation of magnitude and deviation in compensatory response at the cutoff frequency of 3 kHz with F3, f_0, and spectral tilt.

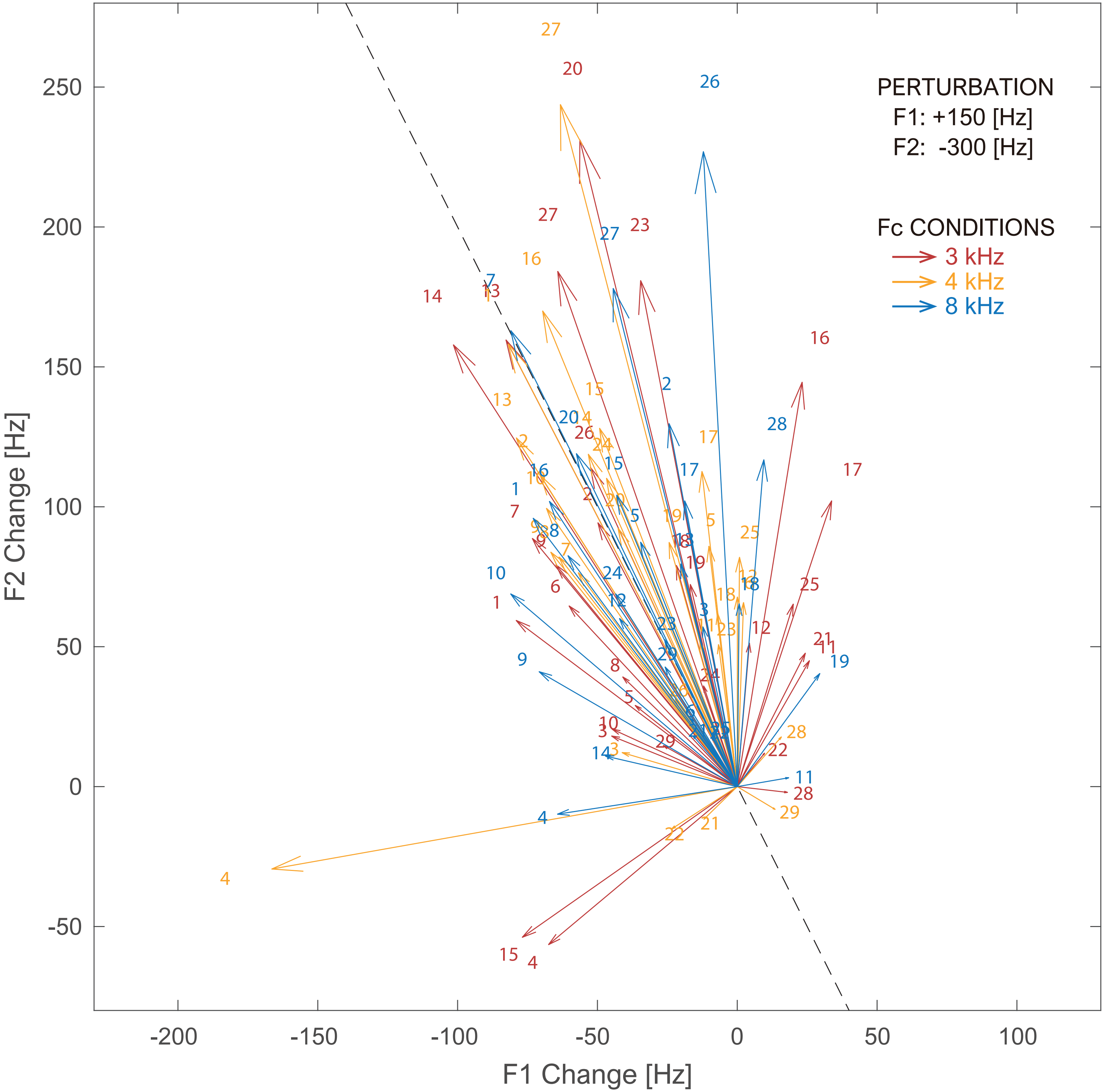

8 kHz
4 kHz
F3
3 kHz
F2
F1
Frequency [Hz]
Time

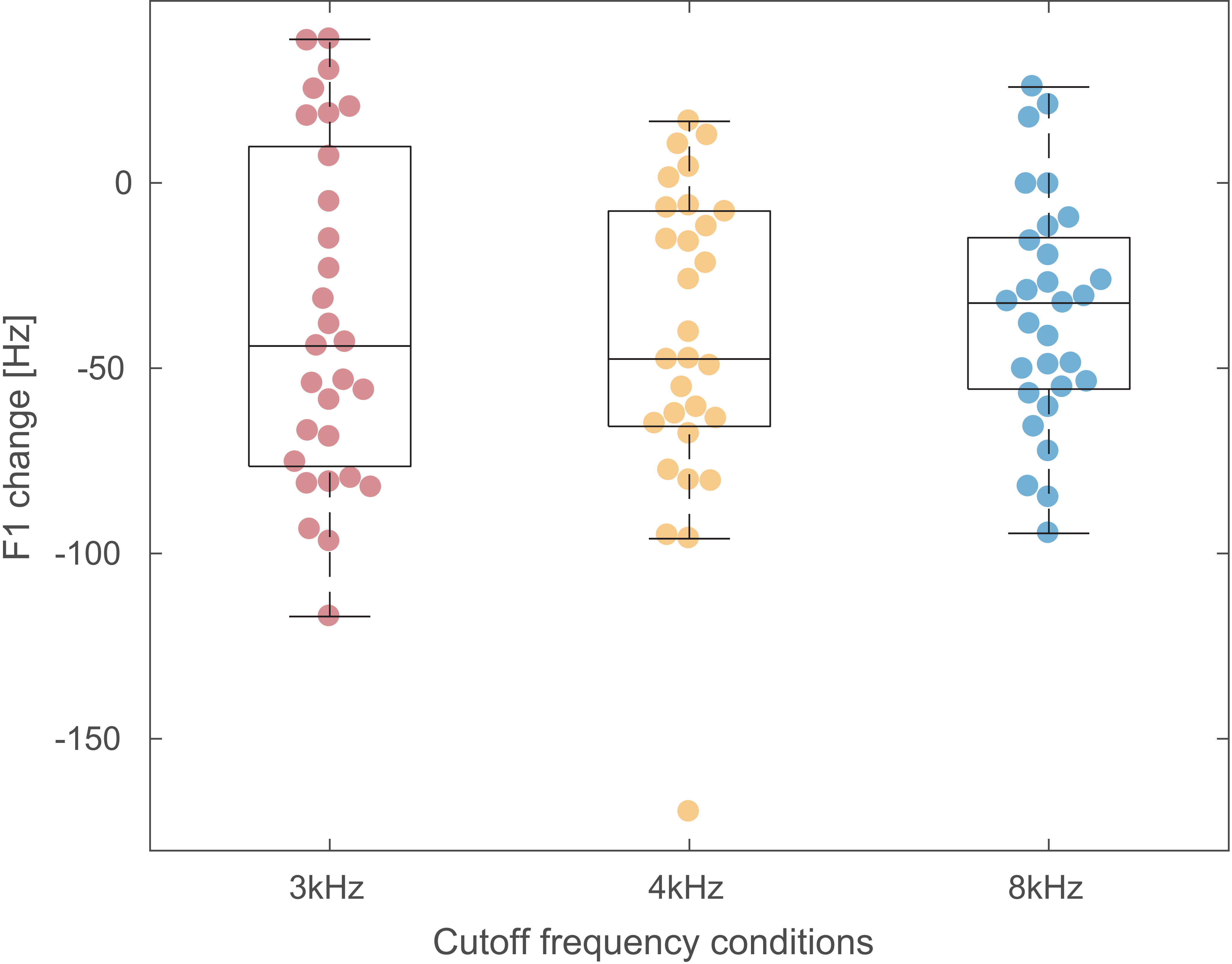

DSP
(PEAR)
LPF
Fc = (3, 4, 8) kHz
Head. Amp .
Display
Main PC
• Parameter setting
• Data recording
Sound proof room
LPF
Fc = 8 kHz
Mic. Amp .

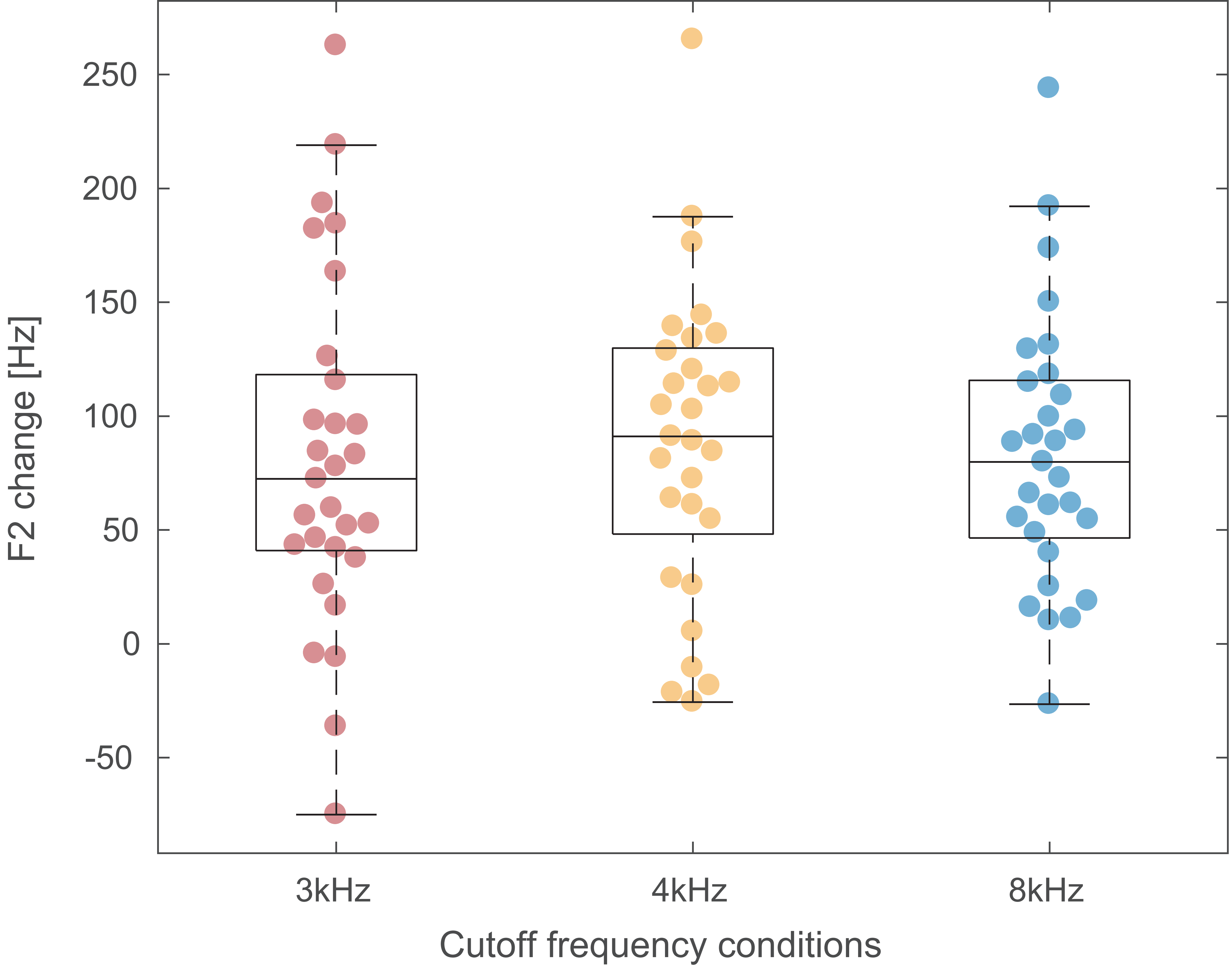

Perturbation magnitude [Hz]
Baseline
Ramp
Hold
Return
F2
-300
+150
F1
1
20
70
90
140
Number of trials


Deviation
Formant response vector $\overrightarrow{F_R}$
Ideal compensatory response vector $\overrightarrow{F_I}$
Angle
Magnitude
Baseline
Perturbation vector

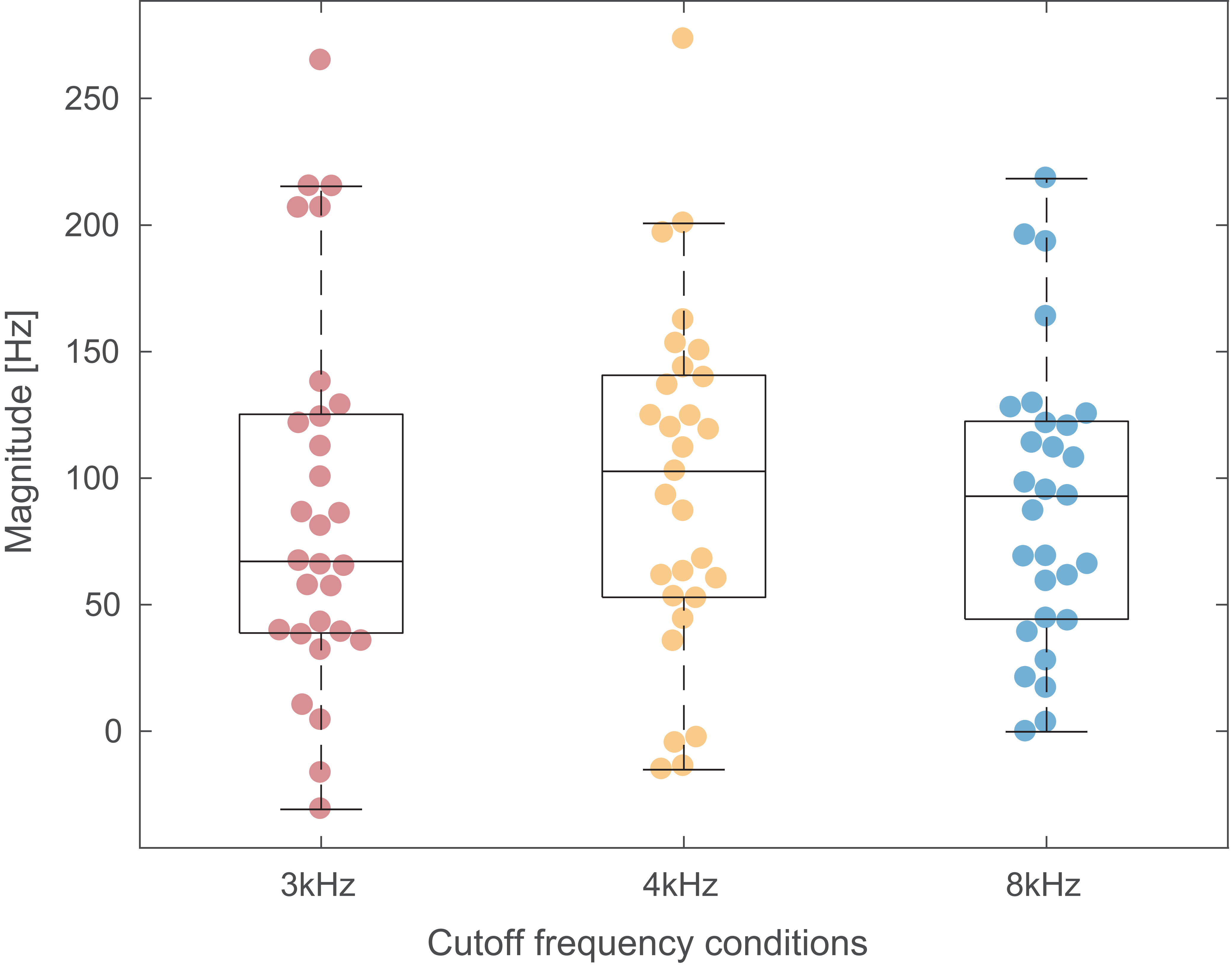

Fc = 3 kHz
Fc = 4 kHz
Fc = 8 kHz
F2 Change [Hz]
F1 Change [Hz]
120
100
80
60
40
20
0
-20
-40
30
20
10
0
-10
-20
-30
-40
-50
-60
20
40
60
80
100
120
140
Number of trials

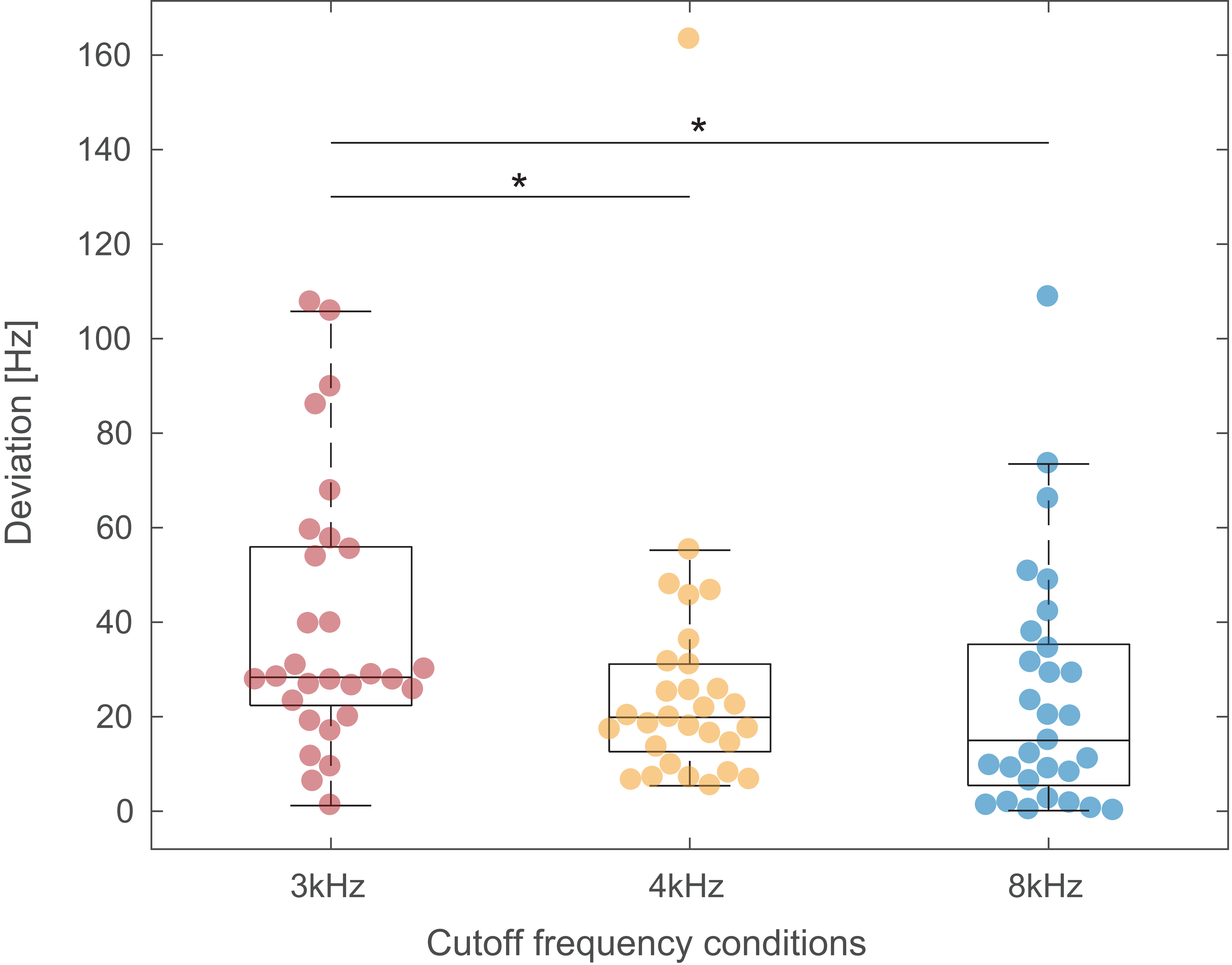

PERTURBATION
F1: +150 [Hz]
F2: -300 [Hz]
Fc CONDITIONS
3 kHz
4 kHz
8 kHz
F2 Change [Hz]
F1 Change [Hz]
250
200
150
100
50
0
-50
-200
-150
-100
-50
0
50
100

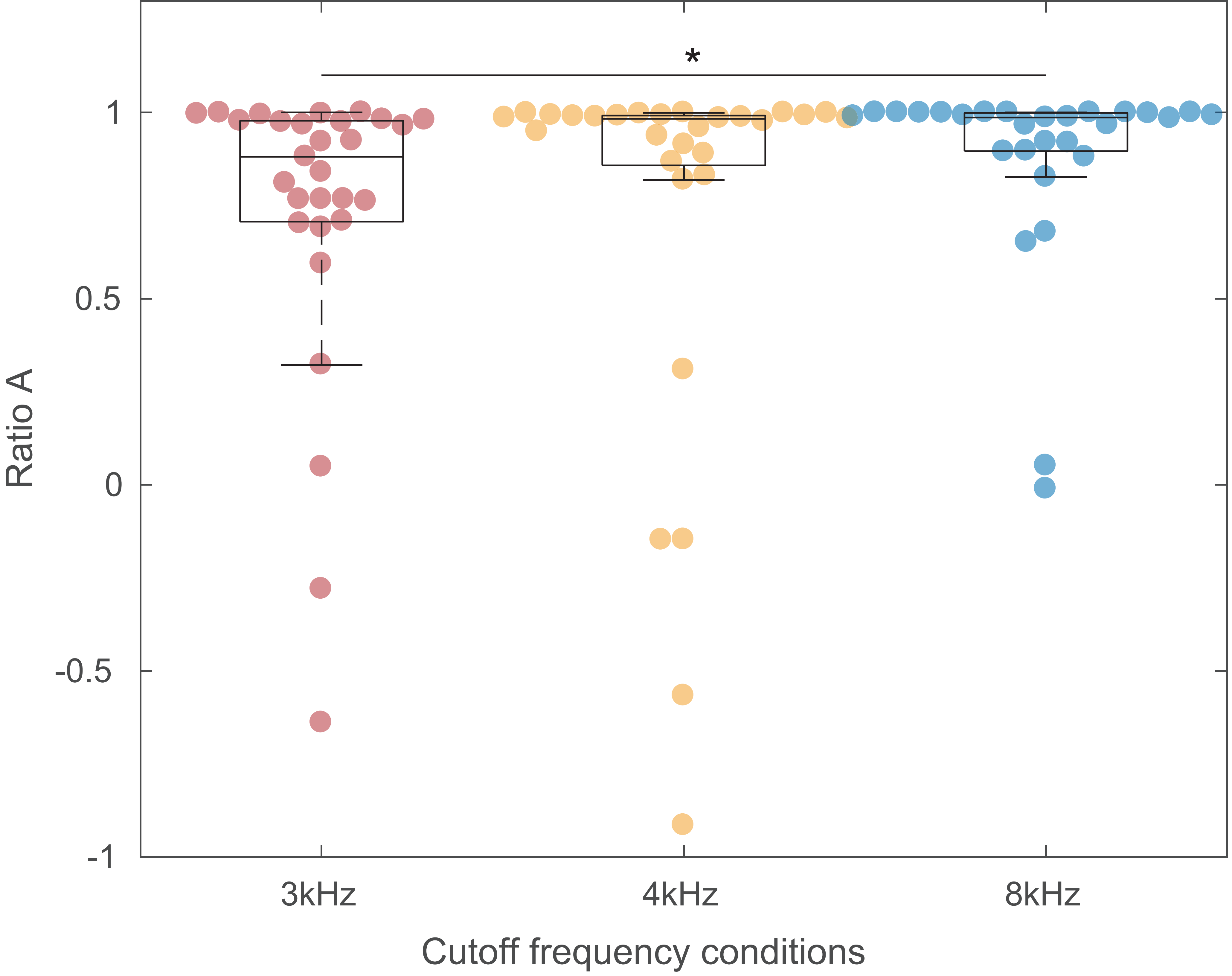

F1 change [Hz]
0
-50
-100
-150
3kHz
4kHz
8kHz
Cutoff frequency conditions

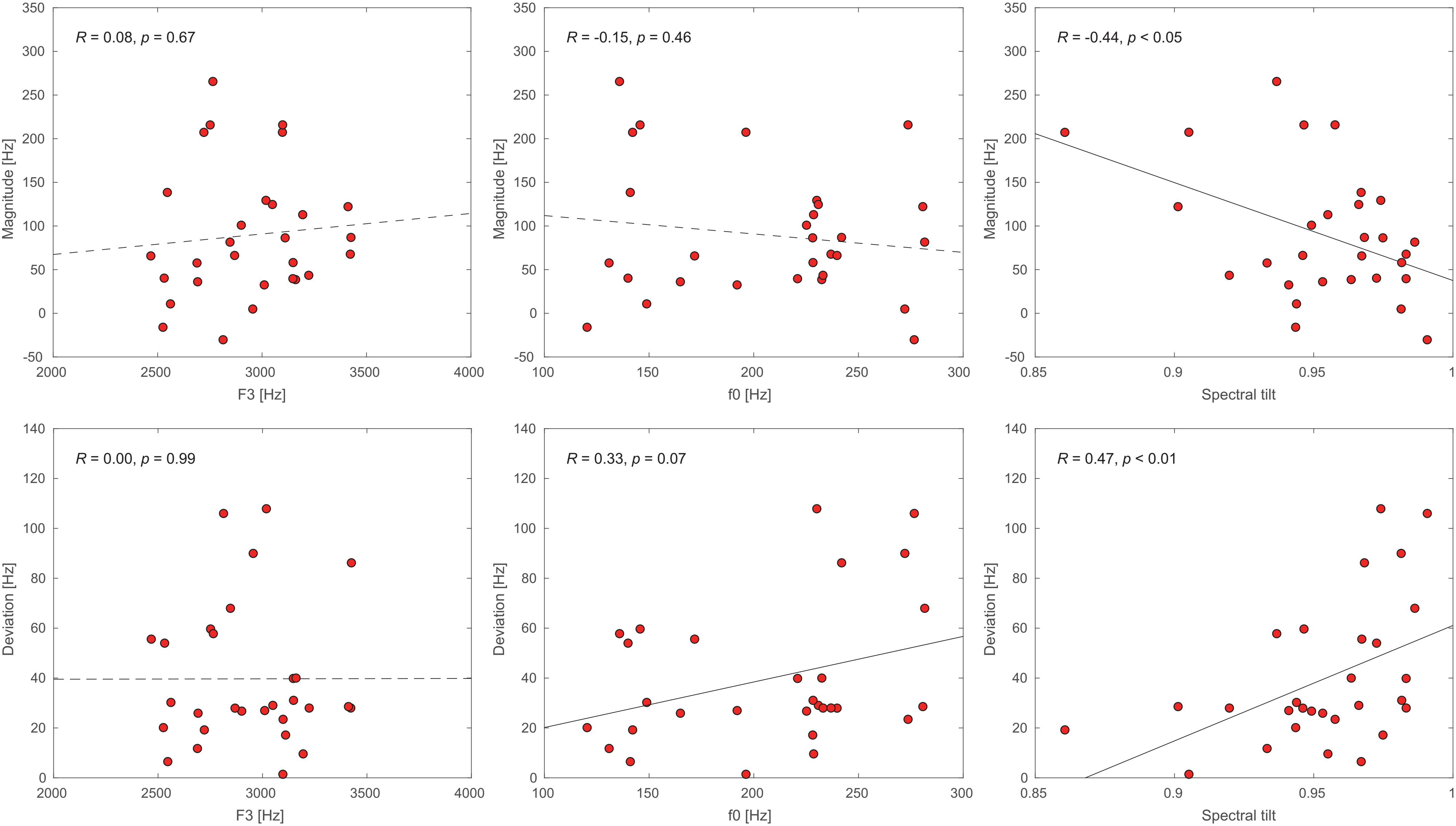


250
200
150
100
50
0
-50
F2 change [Hz]
3kHz
4kHz
8kHz
Cutoff frequency conditions


250
200
150
100
50
0
Magnitude [Hz]
3kHz
4kHz
8kHz
Cutoff frequency conditions


Deviation [Hz]
160
140
120
100
80
60
40
20
0
*
*
3kHz
4kHz
8kHz
Cutoff frequency conditions


*
1
0.5
0
-0.5
-1
Ratio A
3kHz
4kHz
8kHz
Cutoff frequency conditions



R = 0.08, p = 0.67
R = -0.15, p = 0.46
R = -0.44, p < 0.05
R = 0.00, p = 0.99
R = 0.33, p = 0.07
R = 0.47, p < 0.01
Magnitude [Hz]
Deviation [Hz]
F3 [Hz]
f0 [Hz]
Spectral tilt
350
300
250
200
150
100
50
0
-50
140
120
100
80
60
40
20
0
2000
2500
3000
3500
4000
100
150
200
250
300
0.85
0.9
0.95
1