Internal Use Only
非公開

TR－H－040

0054

Transformed Auditory Feedback: Effects of Fundamental Frequency Perturbation

Hideki KAWAHARA

1993.12. 8

ATR人間情報通信研究所
〒619-02 京都府相楽郡精華町光台2-2　☎07749-5-1011

ATR Human Information Processing Research Laboratories

2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
Telephone: +81-7749-5-1011
Facsimile: +81-7749-5-1008

Transformed Auditory Feedback: Effects of Fundamental Frequency Perturbation

Hideki KAWAHARA

Abstract

A rapid interaction between auditory perception and speech production has been found using fundamental frequency perturbation to transform auditory feedback. The technique of transformed auditory feedback has been developed to keep the disturbance of normal speech production processes to a minimum, while allowing interactions between speech production and auditory perception to be detected. In the first experiment, subjects were instructed to sustain the Japanese vowel /a/ and to keep the pitch constant. Subjects' speech was frequency modulated using sinusoids ranging from 2Hz to 7Hz and fed back diotically via a headphone. The modulation depth was 200 cents peak to peak. The results indicated that there was a phase-locking effect in the fundamental frequency of produced speech. In the second experiment, a correlation analysis using a pseudorandom signal as a modulation source revealed that the reaction to fundamental frequency perturbation is corrective and that its latency ranges from 100ms to 200ms. Its relation to the auditory-laryngeal reflex will also be discussed.

1 Introduction

Research on interactions between speech perception and speech production may provide an interesting clue to the internal representation of speech, if speech production processes involve speech perception as an indispensable ingredient. If we could extract and measure interactions quantitatively, we would have a direct probe for auditory information processing. One idea is to introduce modifications into one of the feedback paths, which account for the interactions, and observe its effects on produced speech. DAF (Delayed Auditory Feedback) is a well known example of such attemps[1, 2].

1.1 Background (DAF)

Delayed auditory feedback was a hot research topic in speech production and speech perception because its effects on speech production are very strong and striking, and it was thought to be evidence of interactions between speech production and speech perception. In a sense, speech production as a multiple servo-mechanism was a common view in the 1950s[1, 2].

The main effects of DAF resemble stuttering. The effects consist of 1) increased repetition, 2) slowing 3) increased mispronunciation and 4) increased intensity and fundamental frequency. DAF was intensively investigated as a technique of remediation for stuttering because many stutterers could speak fluently under DAF, suggesting that one possible cause of stuttering is a timing error in the feedback paths[3].

But the strong destructive effect of DAF itself made DAF an inappropriate tool to investigate interactions between speech perception and speech production under natural conditions. A recent review of DAF suggests that it is an abnormal reaction to an abnormal feedback condition: therefore it was concluded that the effects caused by DAF would reveal nothing about interactions under normal conditions. This view is supported by the findings of a DAF experiment using a speech-modulated square wave, which produced similar effects[4].

These points do not necessarily imply that there are no interactions between speech perception and speech production under normal conditions. There are several indirect evidence of such interactions to exist. The speech parametric shift caused by Lombard effect may be account for some effect of auditory/speech perception on production[5]. Short term audotory laryngeal reflex and effects of frequency moduleted tones on phonation also suggest possible physiological basis for interactions under normal conditions[6, 7]. However, these effects were reported very small and their role in speech production still remains unclear.

1.2 Foreground (TAF)

Transformed auditory feedback (TAF) has been developed to investigate interactions between speech perception and speech production under natural conditions using parametric transformation of feedback speech[8]. The basic idea of TAF is to keep the disturbance to speech production caused by the perturbation to a minimum, while allowing interactions between speech perception and speech production to be detected. This can be done by combining a small perturbation with correlation analysis of the input perturbation and the output speech parameter.

In order to apply TAF successfully, it is indispensable to design an appropriate perturbation signal and to decide what parameters should be transformed. One important requirement for the perturbation signal is orthogonality which makes it possible to increase detectability and makes interpretation and data processing easy. Sinusoids and pseudorandom signals are such candidates. It is important that the parameters to be transformed are reliable and sensitive since they directly determine detectability. It is suggested that prosodic parameters may have stronger effects than phonological parameters[4, 9, 10]. A preliminary test revealed that the fundamental frequency is several orders of magnitude more reliable than spectral parameters such as LPC and formants.

Fundamental frequency modulation was selected as the transformation for the first TAF experiments.

1.3 Pilot study

A series of measurements and experiments were performed to replicate results of DAF experiments and to establish standard procedures for TAF experiments and a control condition for reference. The effects of DAF mentioned in the previous section were replicated.

The preliminary TAF experiments used sinusoids as a modulator. The frequency of sinusoids ranged from 2Hz to 7Hz, and modulation depth was 200 cents from peak to peak. Several types of speech were produced under TAF conditions. The utterance types are

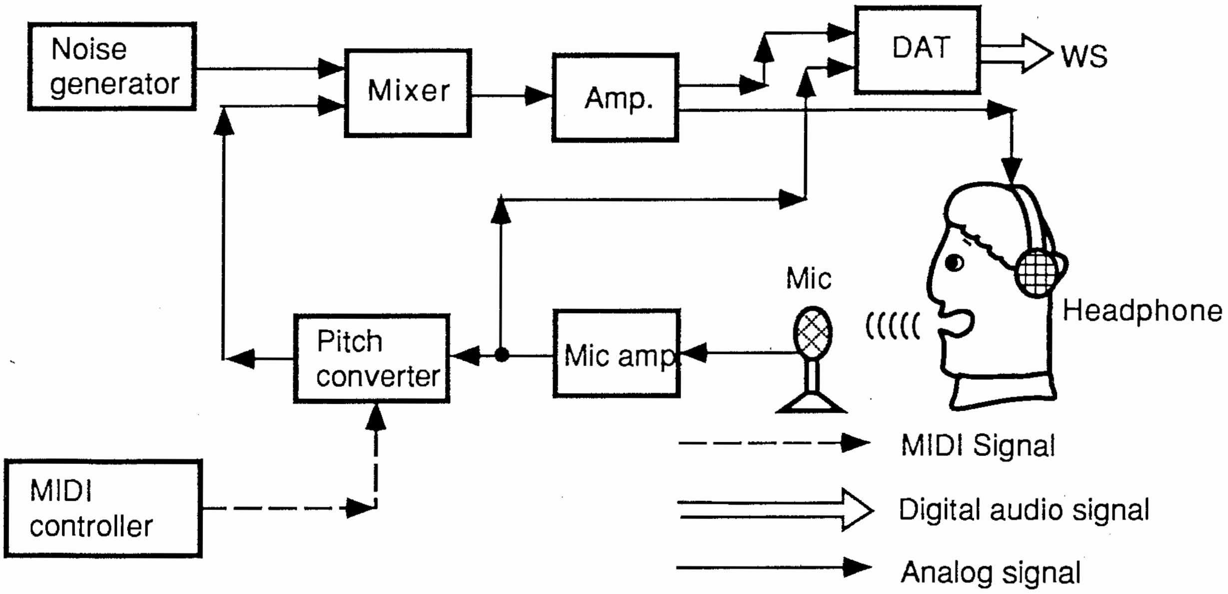

Figure 1: A schematic diagram of TAF experiment.

as follows: sustained vowel, a sentence consisting of all voiced sounds, ordinary sentences from a Japanese fairy tale, and a technical document.

The results indicated that interactions are detectable only in the sustained vowel case. The observed interaction is a phase–locking effect between the fundamental frequency variation of the sustained vowel and the perturbation signal. The other conditions consisted of large fundamental frequency variations caused by the prosodic components of the sentence utterance, which made it difficult to separate the effects of perturbation from prosodic variations in the fundamental frequencies. This does not mean that there are no interactions, but simply that the signal to noise ratio is too small to detect the effects. An attempt to cancel the prosodic components in the sentence utterance using DTW (Dynamic Time Warping) failed, probably due to imperfect reproducibility of prosodic contours.

2 Experiments

Two series of experiments were conducted to quantitatively investigate the interactions between speech production and perception. The first experiment examined the effects of sinusoidal perturbation, while second evaluated the effects of pseudo-random perturbation.

2.1 Equipment

A schematic block diagram is shown in Figure 1. An H3000S (Eventide) pitch converter was used to manipulate the fundamental frequency of fed-back speech. Control parameters were prepared as a MIDI control signal, and MIDI software (Performer) running on a Macintosh computer was used for sending the parameters to the H3000S pitch converter. The algorithm for the fundamental frequency transformation is essentially based on waveform expansion, contraction, and splicing.

This equipment is set up in an ordinary room environment without special sound attenuation. In order to reduce the effect of natural side tones and bone conduction, subjects wore circumaural headphones (Sennheiser HD250 linear II), through which their speech was fed back with the addition of 78dB(A) of pink noise. This treatment also reduced the effect of any environmental noise.

The produced speech sound and the fed-back sound were separately recorded on left and right channels of a dual channel DAT recorder then, via a DAT interface, were sent to a workstation and stored.

2.2 Perturbation

Two types of perturbation signals were used: a sinusoidal signal and a pseudo random sequence. A sinusoidal signal, because of its line sprctral representation, is useful in frequency domain analysis. Four sinusoidal signals were used: 2Hz, 3Hz, 5Hz and 7.13Hz with 200 cent, peak–to–peak deviations.

A pseudo-random signal is a deterministic binary sequence, which is useful for time domain analysis, because auto-correlation of the signal is the delta function; therefore, cross-correlation between the signal and the system's response to the signal gives the impulse response of the system. The pseudo-random signal M_n, a bias-adjusted M-sequence, was generated using the following recursive equation, and its bias was adjusted according to the literature[11].

$$X_n = X_{n-1} \bigoplus X_{n-4} \tag{1}$$

$$M_n = \begin{cases} 1 \pm \frac{2}{\sqrt{k+1}} & \text{if} \quad X_n = 1 \\ -1 & \text{if} \quad X_n = 0 \end{cases} \tag{2}$$

Here, $\bigoplus$ represents the "exclusive or" and k represents the period of the sequence (31 in this case).

The actual perturbation signal S_t that was fed to the pitch converter, was a version of M_n that was up sampled eight times. The samples of signal S_t were sent at intervals of 1/128 of a second. In this case the cycle of the pseudo-random signal was approximately 1.94s. We set the perturbation depth of the pseudo-random signal at a peak-to-peak value of 100 cents.

2.3 Speech Materials

Subjects for pseudo random perturbation totaled eight: two girls, four men, and two women, who were instructed to sustain the Japanese vowel /a/ with a constant pitch for about ten seconds. The subjects were asked to phonate at their normal pitch. Additionally, three of the men and the two women were asked to phonate both at high and low pitch as well as at their normal pitch. Three men were also tested using sinusoidal perturbation. Each subjects' speech production was recorded three times under various feedback conditions.

2.4 Speech Parameter Extraction

Since preliminary studies predicted that the effect of TAF would be less than 1Hz when converted to the rms value of the fundamental frequency variation, it was necessary to extract them precisely. In the preliminary experiments, we employed a fundamental frequency extraction method using a lag window in the autocorrelation domain[14] to find an initial estimate of the pitch period. Then parabolic interpolation was applied to three

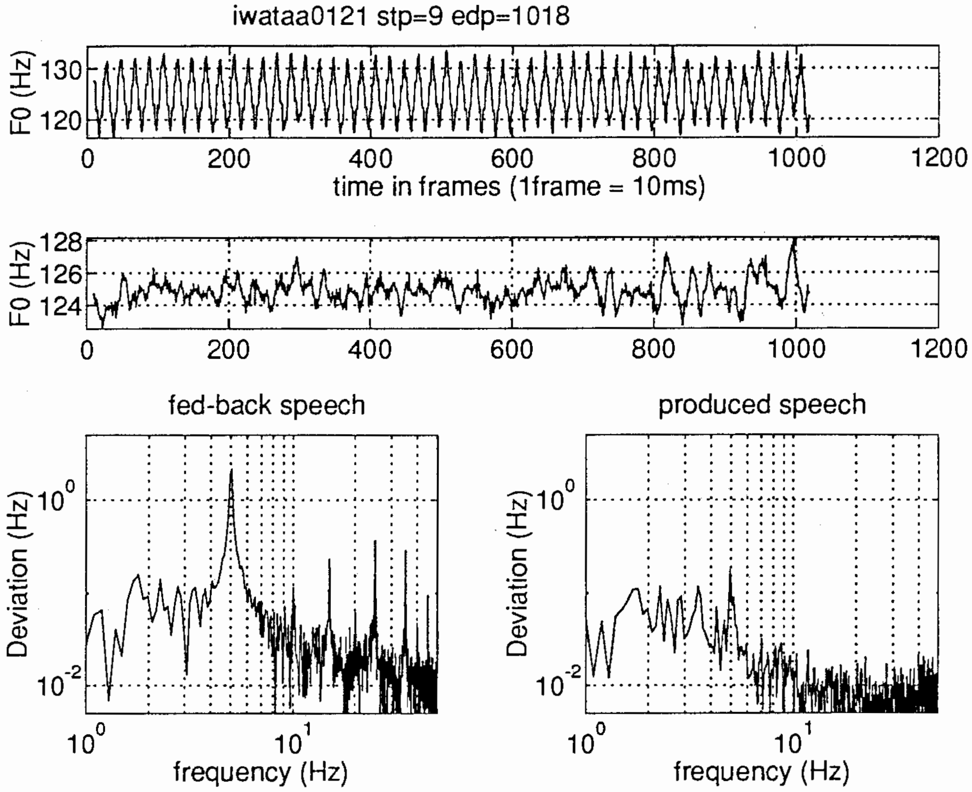

Figure 2: An example of response to a sinusoidal perturbation. The top plot shows the fundamental frequency of the fed-back speech. The middle plot shows that of produced speech. The lower left plot shows the power spectrum of the fundamental frequency variation of the fed-back speech. The lower right plot shows that of the produced speech. The frequency of the perturbation is 5Hz.

autocorrelation values around the initial estimate to find the final precise estimate. These values are recorded as the reference data to evaluate the other faster algorithms. The "formant" program in ESPS speech signal processing software system [12, 15] was used throughout the experiments described below, because it showed reasonable (within 0.5Hz) agreement with the reference data and was several times faster than the reference method.

The original speech waveform data were sampled at 48KHz and stored. Then, their sampling rate was converted to 16KHz to make later processing efficient. Fundamental frequencies were calculated every 10 ms using the down-sampled data. A semi-automatic interactive segmentation method was used to select the data segment to be analyzed.

3 Experimental Results

3.1 Sinusoidal Perturbation

One example of the experimental results with the sinusoidal perturbation signals is shown in Fig. 2 and Fig. 3.

Fig. 2 shows the extracted fundamental frequencies and their power spectrums. It is

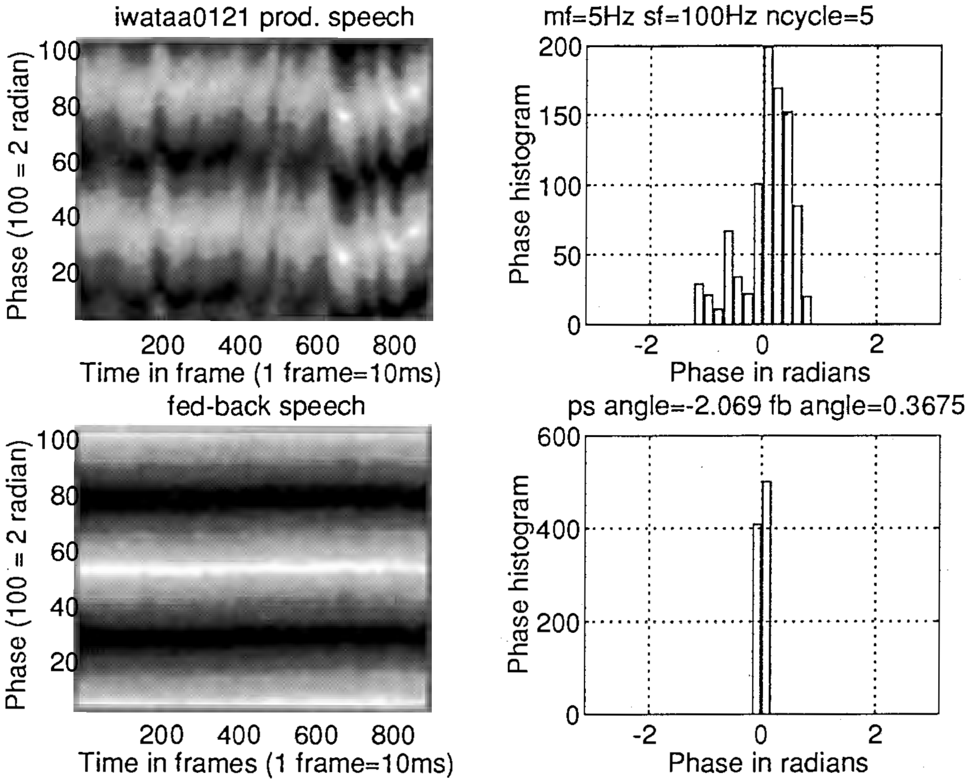

Figure 3: An example of phasic relation to perturbation. The top left image represents the relative phase of the fundamental frequency of the produced speech. A stable phasic relation is shown as horizontal bars. The bottom left image represents that of fed-back speech. The top right plot shows the histogram of the phase difference between the perturbation and the fundamental frequency of the produced speech. The bottom right shows that of the fed-back speech.

clearly shown that the perturbation to the fed-back speech produced a sinusoidal variation in the fundamental frequency of the fed-back speech. The power spectrum of the fundamental frequency variation of the produced speech shows a small peak at the perturbation frequency, indicating that there might be some interactions. However, the peak is not high enough to be statistically significant.

Fig. 3 shows the result of a correlation analysis for the same data. A cross correlation with a complex sinusoid ($e^{-j\omega t}$) of the perturbation frequency was computed every 10ms with a 1-s correlation window. The complex correlation obtained was calibrated using the phase of the perturbation signal. The vertical axis of the images on the left in Fig. 3 represents phasic rotation. The density of the image represents the real part of the complex correlation. If the phasic relation to the perturbation signal stays stable with time, bold horizontal bars will appear in the image.

The definite dark horizontal bars in the lower left image indicate that the sinusoidal perturbation to the fundamental frequency of the fed-back speech was successful. The top left image shows that the fundamental frequency variation of the produced speech was generally phase-locked to the perturbation signal.

This observation is validated by the following analysis using the right-hand plots in Fig. 3. The right plots represent histograms of phasic relations. If the phasic relation is constant, a sharp distribution will result. However, if there are no stable phasic relations, a uniform distribution will result. The Kolmogorov-Smirnov test of the histogram revealed that the histogram for the produced speech was not uniform, indicating that the phase-locking effect is not an artifact.

Almost all the subjects showed the phase-locking effect under 5Hz perturbation. The effect sometimes was not clear under 2Hz and 3Hz perturbation. Detailed results are given in the literature[17].

3.2 Pseudorandom Perturbation

It is now clear that there is an interaction between speech production and perception. The next step is to investigate the characteristics of the interaction quantitatively. A series of experiments using the pseudorandom perturbation was conducted to extract temporal response.

3.2.1 General Features

One example of the experimental results with pseudorandom perturbation is shown in Fig. 4. The vertical axis of Fig. 4 shows the normalized value of correlation between signal S_t and the extracted fundamental frequencies, normalized to show the same value as a fundamental frequency variation in Hz. The abscissa shows the pitch analysis frame numbers, and one frame corresponds to 10 ms.

The middle plot in the figure represents the correlation between the fundamental frequency of the fed-back speech and the perturbation. The bottom plot in the figure represents the correlation between the fundamental frequency of the produced speech and the perturbation. The interval between the positive salient peaks corresponds to the cycle of the pseudo-random signal. The bottom plot shows the correlation between the fundamental frequency of the produced speech and the perturbation, and demonstrates inverted polarity with reference to the middle plot.

This means that when the fundamental frequency varies during the process of phonation, there is a mechanism to compensate for fundamental frequency variance. The effect of compensation in produced speech in this example is less than 20% of the change in the fundamental frequency of the fed-back speech, if its peak values are used as the measure. However, when the area of the peaks is compared, the amount of compensation is more than 50%. This indicates that the response has low-pass characteristics.

Here we defined the distance between the positive peak of the perturbation and the corresponding negative peak of the response as the latency of the perturbation effect on the auditory feedback loop. In the example in Fig. 4, the latency defined above is 154 ms. In order to have a more precise measure of latency than the frame size permits, a parabolic interpolation was applied as was done in fundamental frequency extraction.

3.2.2 Statistical Analysis

The results of ANOVA with two variables: subjects and pitch production conditions (normal pitch, high pitch, low pitch); and a dependent variable of latency were: subject $[F(4,30) = 22.87, p < 0.0000]$, condition $[F(2,30) = 19.32, p < 0.0000]$, and interaction $[F(8,30) = 5.12, p < 0.0005]$ were significant.

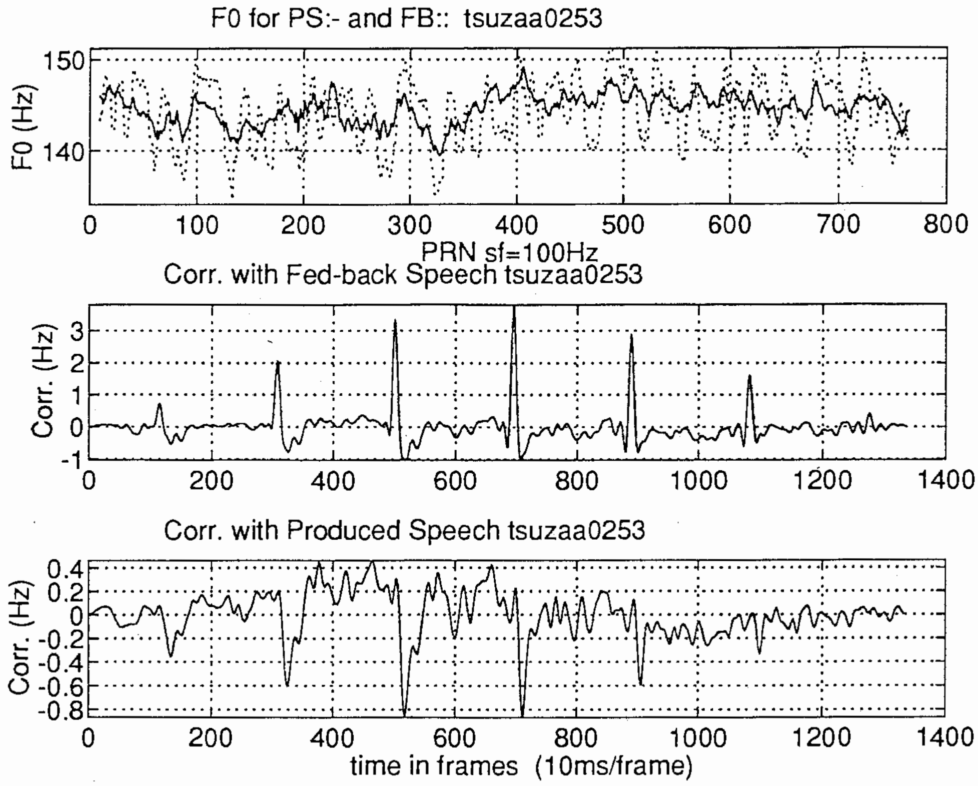

Figure 4: An example of fundamental frequency correlations. The top plot shows the fundamental frequencies of the fed-back speech and the produced speech with pseudo random signal. The solid line represents produced speech, and the dashed line represents fed-back speech. The middle plot shows the correlation between the pseudorandom perturbation and the fundamental frequency of the fed-back speech. The bottom plot shows the similar correlation of the produced speech.

Fig. 5 shows the variation in latency for each subject. In the figure, a diamond represents the mean and 95% confidence interval and a box shows the percentile values. Latency varied from 100–200 ms and was distributed around 150 ms except for one subject who showed especially short latencies.

Fig. 6 shows the variation in latency in each pitch production condition for each subject. The latency decreased as the fundamental frequency increased except for one subject. The subject showing a different tendency and the subject showing especially short latencies in Fig. 5 being identical.

3.3 Relations between two conditions

Relations between these two perturbation conditions are investigated to find out whether they represent the same response or not. In the previous experiments it was observed that the major component of the response to the pseudorandom perturbation is the corrective response. If this is indeed the case, the phasic difference under sinusoidal perturbation can be predicted using the observed latency and the perturbation frequency.

Fig. 7 shows the predicted phase and the observations. The prediction shows good agree-

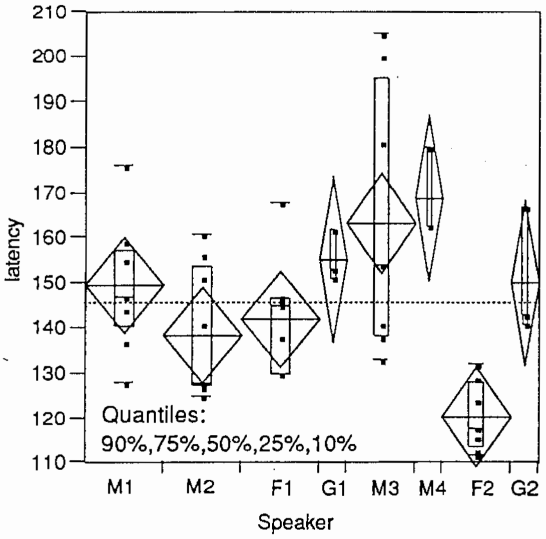

Figure 5: Reaction time to pitch perturbation for individual subjects. Diamonds represent the mean and confidence interval. Percentiles are also represented. (M: male, F: female, G: girl)

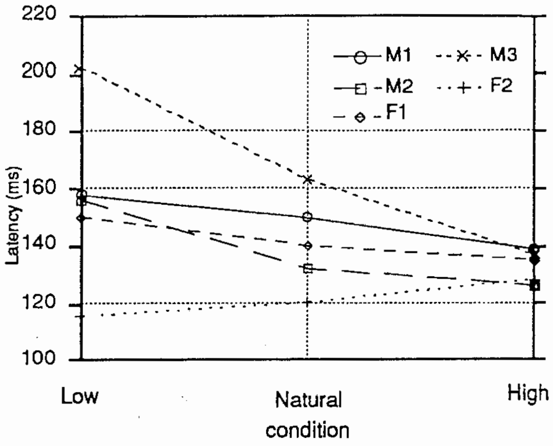

Figure 6: Reaction time dependency to voicing condition for each subject. Each line represents a different subject.

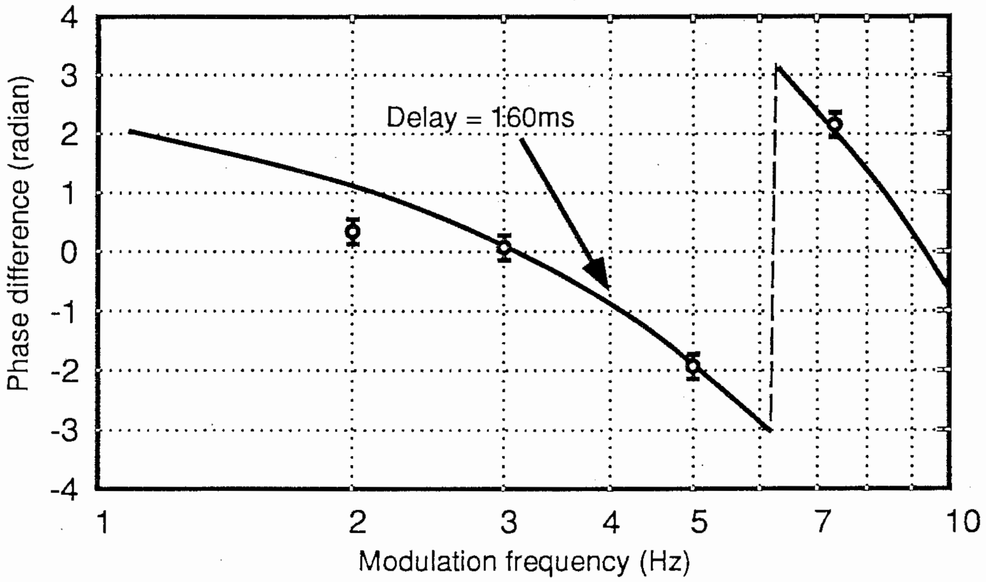

Figure 7: Relations between predicted phasic difference and observed phase for one subject. The solid line shows the predicted phase, assuming a 160-ms corrective response. The open circles are the average values of the observed phasic difference.

ment with the observations, suggesting that the reactions to these two types of perturbation signals are produced by the same mechanism. The apparent difference at 2Hz perturbation may reflect an oversimplification of the response, because the low-pass characteristics of the response imply that approximation using a negative impulse introduces a considerable error in the low frequency region.

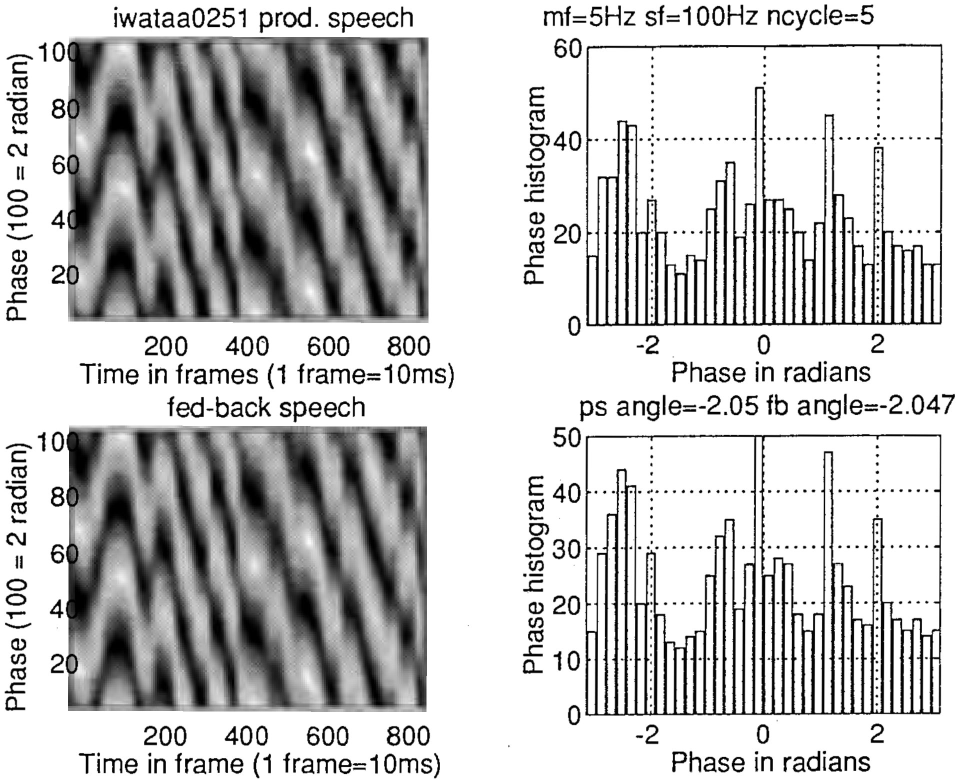

Figure 8: An example of the phasic relation analysis for the control data. The top and bottom panels are basically identical, because there is no perturbation. The left image shows the phasic relation and the right plot shows the phase histogram. No stable phasic relation is visible.

Further discussions and detailed results are given in the literature[8, 17].

3.4 Experimental Controls

In the control experiment, where the feedback speech was not manipulated, the fundamental frequency data were analyzed using the same procedures. The difference between the fundamental frequency extracted from the feedback speech mixed with pink noise and that extracted directly from the produced speech, was less than 0.1Hz and negligible.

In the first analysis, using a sinusoidal reference signal indicated no phase-locking for these control data. An example of a phasic image is given in Fig. 8, and clearly shows phase drift. The difference between phase distribution in the right plot and the uniform distribution, is not statistically significant.

This suggests that the phase-locking effect observed in the analysis of the correlation between the fundamental frequency of the speech produced under TAF conditions using the fundamental frequency perturbation by the sinusoidal signals was not an experimental artifact but an actual phenomenon.

In the analysis between the extracted pitch of the normally produced speech and the correlation with the pseudorandom signal used for perturbation, there were no clear peaks. An example of correlation is given in Fig. 9. Many small peaks due to natural variation of the fundamental frequency are seen, but no clear peaks. This example also gives an estimate

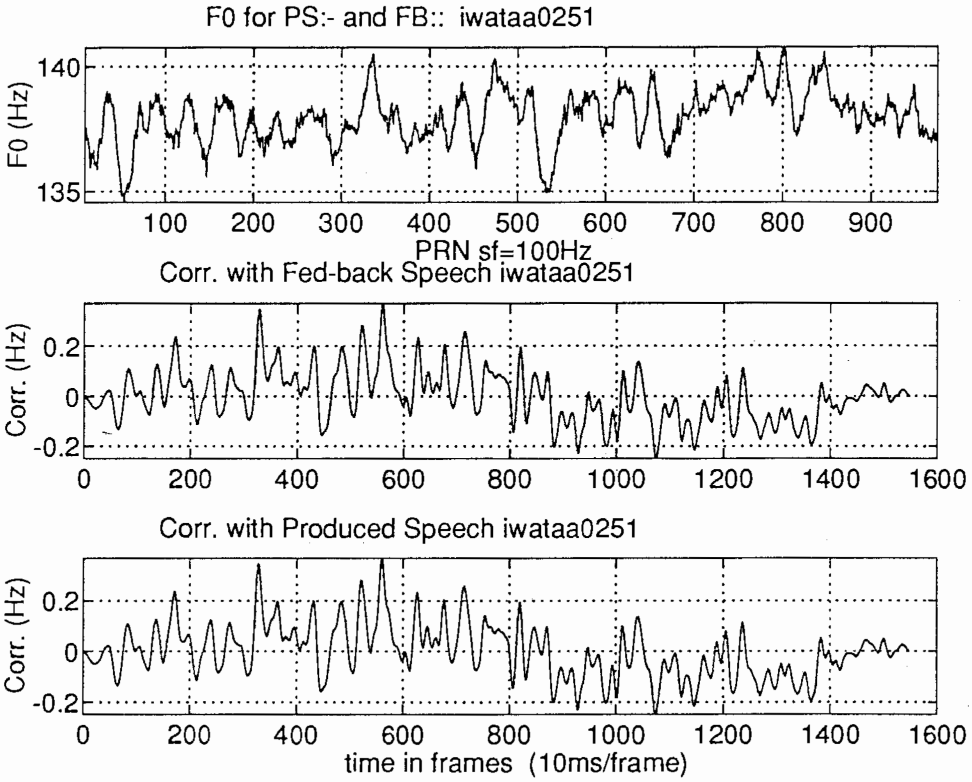

Figure 9: An example of correlation analysis using pseudorandom perturbation for control data. The top plot shows the fundamental frequencies of fed-back speech and produced speech. The middle and bottom plots are correlations between these fundamental frequencies and the pseudorandom signal. They are basically identical, because there was no perturbation. No definite peaks are visible.

of the error involved in the previous analysis using the pseudorandom perturbation. It suggests that the negative responses observed under TAF condition are three times bigger than these erroneous peaks.

This also suggests that the periodic peaks observed in the analysis of the correlation between the fundamental frequency of the speech produced under TAF conditions using the fundamental frequency perturbation by the pseudorandom signal was not an experimental artifact but an actual phenomenon.

4 Discussion

The TAF experiments using fundamental frequency perturbation as the transformation revealed that there is an interaction between speech perception and production, which compensates for variations in the fundamental frequency of one's own speech during phonation. The average reaction latency to the perturbation is around 150ms. This may provide some clues to the identity of the neural processes involved in this response. This value may be too short for a voluntary control to be the source of this response, because the results of a study on pitch tracking by voice [16] suggest that a delay of more than 300ms will be introduced in responding to perturbation, if the response is intentional.

There are several competing hypotheses on the origin of this corrective response: one is that this response is one of the variants of the "auditory laryngeal reflex"[13, 18, 21, 6], other assumes involvement of trans-cortical connections[7]. If the response is found to be specialized to a specific parametric dimension and insensitive to the other dimensions, the response may be mediated by higher sophisticated processes. And if the similar compensating response can be found in the other frequency control tasks, it may also support the latter hypothesis. Obviously, it is still too early to make a definite statement. But several pilot studies not reported here in detail [19, 20] suggest the possibility that the latter hypothesis holds, which may lead to an interesting discussion about the relation to the general principle of motor learning scheme[22].

There is a minor point to be mentioned. In this study we defined the temporal distance between the positive peak of the solid line and the negative peak of the dashed line in cross-covariant analysis as the latency of the feedback loop. However, there is no assurance that the detection of the perturbation depends on the perception of the absolute value of the perceived pitch, there being some possibility that subjects used pitch change as the cue. Although inferences drawn on the latency processes in the paragraph above are basically not affected by the perceptual mode used to detect pitch change, it is necessary to clarify the actual cue to get precise response characteristics.

The interesting characteristic of the TAF method is that the perturbation and the response have the same physical dimensions. This characteristics allow us to use TAF as a non-invasive analysis technique of our auditory functions in various parametric representations. We are planning to take advantage of this characteristic in our future research on the role of the perceptual mechanism in speech production.

5 Conclusion

In this paper, a method of transformed auditory feedback was proposed, in which perturbation to speech parameters small enough not to disrupt normal speech, is fed-back auditorily in order to investigate its effect on the extracted parameters of the produced speech. Using this method, the effect of fundamental frequency perturbation on produced speech was investigated, it was discovered that, in the fundamental frequency of their produced speech, subjects compensated for the perturbation effect with a 100-200 ms latency when the fundamental frequency of the feedback speech was perturbed. Considering that the speed of change in this perturbation does not allow for voluntary control and that the latencies observed were too short to allow the interaction of conscious control, the results obtained in this study suggest the existence of a reflex pitch control loop from auditory processing to speech production.

Acknowledgement

Ms. J. C. Williams of Ohio-State University has made a great contribution in preparing this manuscript and made valuable comments on various aspects of auditory feedback phenomenon. Prof. Shimon Sapir of Northwestern University pointed out an important relation to auditory laryngeal reflex and his pioneering work. Mr. Satoru Iwatani of Toyohashi Institute of Technology performed the preliminary TAF experiments that led to this research. Dr. Kiyoshi Honda of ATR suggested several competing hypothesis on the physiological basis of TAF effects. The other members of ATR Human Information Processing Laboratories also contributed their stimulating discussions and participated as subjects. The author would like to express sincere appreciation to them.

References

[1] B. S. Lee: "Effects of Delayed Speech Feedback," JASA, 22(6), pp.824-826, (1950).

[2] G. Fairbanks: "Selective Vocal Effects of Delayed Auditory Feedback," J. of Speech and Hearing Disorders, 20(4), pp.333-346, (1955).

[3] B. A. Timmons: "Physiological Factors Related to Delayed Auditory Feedback and Stuttering: A Review," Perception and Mortor Skills, 55, pp.1179-1189, (1982).

[4] P. Howell and A. Archer: "Susceptibility to the Effects of Delayed Auditory Feedback," Perception & Psychophysics, 36(3), pp.296-302, (1984).

[5] D. Pisoni, et al. : "Some Acoustic Phonetic Correlates of Speech Production in Noise," Proc. IEEE ICASSP, pp.1581-1584, (1985).

[6] S. Sapir, M. D. McClean and C. R. Larson:"Human Laryngeal Response to Auditory Stimulation," J. Acoust. Soc. Am., 73, 1, pp.315-321, (1983).

[7] S. Sapir, M. D. McClean and E. S. Luschei:"Effects of Frequency-modulated Auditory Tones on the Voice Fundamental Frequency in Humans," J. Acoust. Soc. Am., 73, 3, pp.1070-1073, (1983).

[8] H. Kawahara:"On Interactions Between Speech Production and Perception using Transformed Auditory Feedback – Effects of Pitch Perturbation by Pseudo–random Signal –," Technical Report of ASJ, H–93–24, (May 1993).

[9] J. F. Elman: "Effects of Frequency-shifted Feedback on the Pitch of Vocal Productions, " JASA, 70(1), pp.45-50, (1981).

[10] S. Sapir, E. DeRosier, A. M. Simonson and A. Wohlert:"Effects of Frequency Modulation Tones and Vowel Formants on Perioral Muscle Activity During Isometric Lip Rounding," Journal of Voice, 4, 2, pp.152-158, (1990).

[11] H. Miyakawa: "Random Signal and Estimation of Dynamic System Parameters," Corona Co., pp.220-226, (1983). [in Japanese]

[12] B. G. Secrest and G. R. Doddington: "An Integrated Pitch Tracking Algorithm for Speech Systems," Proc. IEEE ICASSP, pp.1352-1355, (1983).

[13] K. Honda:"Physiological Background of Speech Production," Journal of the Japanese Acoustical Society, 48, pp.9-14, (1992). [in Japanese]

[14] S. Sagayama and S. Furui: "Pitch Extraction with Lag-Windowing," IEICEJ annual conference, 5-26, (1978). [in Japanese]

[15] J. Shore: "Interactive Signal Processing with UNIX ," Speech Technology, Mar./Apr., (1988).

[16] H. B. Nudelman and K. E. Herbrich: "A Model of the Phonatory Response Time of Stutterers and Fluent Speakers to Frequency-Modulated Tones," JASA, 92(4), pp.1882-1888, (1992).

[17] H. Kawahara and J. C. Williams:"Analysis of Pitch Perturbation Effects by Transformed Auditory Feedback," Technical Report of IEICEJ, SP93-38, (July 1993). [In Japanese]

[18] J. Udaka, H. Kanetaka and Y. Koike: "Response of the Human Larynx to Auditory Stimulation," in M. Hirano, J. A. Kirchner and D. M. Bless eds. Neurolaryngology, Collage-Hill Pub., pp.184-198, (1987).

[19] H. Kawahara, T. Hirai and K. Honda:"Laryngeal Musclar Control under Transformed Auditory Feedback with Pitch Perturbation," Technical Report of IEICEJ, SP93-39, (July 1993). [In Japanese]

[20] H. Kawahara:"Implications of Transformed Auditory Feedback in Hearing Research," ITE Technical Report, VAI, 17, 53, pp.1-6, (Sept. 1993). [In Japanese]

[21] S. Sapir:"Acoustic and Electromyographic Analyses of Human Laryngeal Responses to Auditory Stimulation," Doctoral Dissertation, Department of Speech and Hearing Science, University of Washington, (1982).

[22] M. Kawato and H. Gomi:"Feedback–Error–Learning as a Supervised Motor Learning Scheme," Proc. IEEE International Workshop on Neuro-Fuzzy Control, pp.207-214, (1993).