TR - H - 120

Transformed Auditory Feedback: The Collection of Data from 1993.1 to 1994.12 by a New Set of Analysis Procedures

河原　英紀
Hideki Kawahara

1995. 1. 17

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　☎ 0774-95-1011

ATR Human Information Processing Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
Telephone: +81-774-95-1011
Facsimile: +81-774-95-1008

Transformed Auditory Feedback: The collection of data from 1993.1 to 1994.12 by a new set of analysis procedures

Hideki Kawahara
ATR Human Information Processing Research Laboratories
kawahara@hip.atr.co.jp

January 13, 1995

Abstract

Transformed Auditory Feedback (TAF) was earlier proposed to enable the quantitative measurement of interactions between speech perception and speech production. This technical report is a collection of TAF data acquired from 1993.1 to 1994.12. It is a first attempt at applying TAF to investigate interactions in fundamental frequency control. Experiments with TAF have revealed that there is a compensatory response to fundamental frequency perturbations. The typical latency of this response is around 150ms in terms of the peak to peak distance. The experiments listed in this technical report cover talker dependency, pitch dependency, hemispheric dependency, EMG measurement and timber distance dependency.

The data in this report is analyzed in a uniform manner. it includes
(1) Fundamental frequency trajectories
(2) Periodic average representations of fundamental frequencies of fed-back and produced speech (phonation).
(3) The coherency of each variation frequency component.
(4) The loop transfer function of fed-back-to-produced interactions with confidence intervals.
(5) The minimum AIC estimations of AR parameters of the fundamental frequency trajetories for natural feedback conditions without artificial manipulations.
(6) Decomposition of the estimated response into two dominant components.
Representations from (2), (3) and (4) are averaged over separate measurements of the same conditions and illustrated. In addition mathematical descriptions and calculation procedures of all statistical values are given in detail. This allows us to estimate impulse and step responses to auditory stimulations.

New findings using this set of procedures include (1) a strong but slow response around the 0.5Hz region and (2) the possible existence of the same response for natural speech. The first finding is due to a new decomposition algorithm. The validness of this decomposition is demonstrated by the fact that introducing about a 500ms delay into the artificial auditory feedback path makes the pitch contour very unstable.

Contents

Symbol	description
p_v	Voicing probability
w	Voicing indication (binary)
f_0	Fundamental frequency
T_p	Period of PN signal
s	Pseudo random sequence
N	Length of data
$\mathcal{F}$	Discrete Fourier transform
P	Fourier representation of differentiation
h	Impulse response of the system
y	System output
A	Transfer function of auditory system
S	Transfer function of production system
G	Loop transfer function
$h^{(open)}$	Open loop impulse response
$h^{(close)}$	Closed loop impulse response
r	Step response
ε	Residual
σ^2	Variance of residual
R^2	Power of residual
σ^2_{resp}	Variance associated with response obtained by cross correlation
σ	Square root of variance associated with response obtained by cross correlation
ξ	Confidential interval of response
γ	Coherency
x	Input signal
y	Output signal
n	Additive noise
Φ_{YY}	Power spectrum density of output
Φ_{XX}	Power spectrum density of input
Φ_{XY}	Cross spectrum between input and output
Φ_{NN}	Power spectrum density of noise
G	Transfer function
$\|G\|$	Transfer function (modulus)
ϕ	Transfer function (phase)

1 Introduction

This technical report is a collection of analyses for data collected in several sets of TAF (Transformed Auditory Feedback) experiments. The important point of this report is that all of the data is analyzed using the same analysis procedures. TDoing so fo cource is not always ideal but very informative as will be shown.

The data to be analyzed was collected from January 1993 through December 1994. under TAF conditions. The analysis presented in our previous report was incomplete due to the lack of a proper model for the f_0 control mechanism and the lack of elaborated statistical analyses.

2 Background: DAF and TAF

Humans are believed to use auditory feedback information to control their way of speaking. The effects of DAF (Delayed Auditory Feedback)[31, 7] and the Lombard Effect[36] are good examples. Under DAF conditions, introducing a several hundred ms delay (typically 200ms) into the auditory feedback path, usually disrupts normal speech and causes sounds like stuttering. The acoustic-laryngeal reflex gives another example of perception-production coupling[38, 39, 40, 44], where an abrupt strong sound increases the fundamental frequency in a short period of time (typically 30ms), and the presentation of an FM modulated sinusoid induces a synchronous variation of the voice fundamental frequency. However, quantitative analyses of these interactions are not well documented, and consequently, how speech perception is integrated into speech production is still an open question.

We developed a measuring technique called TAF (Transformed Auditory Feedback) to investigate interactions between speech perception and speech production[11, 13]. TAF enables the quantitative analysis of interactions mediated by various parametric representations of speech[14]-[27][45, 8].

TAF was introduced to facilitate the quantitative measurement of interactions between speech perception and speech production. The basic idea behind TAF's development was to introduce a parametric perturbation small enough so as not to disturb normal speech production but large enough to make the effects detectable. Two types of orthogonal functions are employed as perturbation signals to separate responses to perturbations from background fluctuations. One is a sinusoid and the other is pseudo random noise (PN) derived from an M-sequence[46, 33].

3 Analysis

This section describes the method used in the following analysis. The mathematical background of the procedures is also described.

The important points are summarized as follows.

- The reliability of the analyses is improved by taking advantage of the periodic nature of PN sequences.

- Responses to perturbations are represented in both the time domain and frequency domain.
- Responses to step perturbations are calculated based on estimated impulse responses.
- A method to decompose the response waveform into a set of second order responses is introduced.
- A new set of procedures makes these analyses applicable to natural speech which is intermittent in nature.

3.1 Periodic averaging

We utilize a set of pseudo random signals generated using an M-sequence. The M-sequence itself has an averaging function and is periodic in nature. Because impulse estimation using the M-sequence is a linear operation, additional averaging based on its periodicity further improves signal to noise ratio.

Fundamental frequency f_0 and voicing probability p_v are extracted from a speech sample. Let T_p be the period of the pseudo random sequence s. Then, the periodic average of f_0 is defined as follows.

$$\widetilde{f_0(n)} = \frac{\sum_{k=0}^{\lfloor N/T_p \rfloor} f_0(kT_p + n)w(kT_p + n)}{\sum_{k=0}^{\lfloor N/T_p \rfloor} w(kT_p + n)} \tag{1}$$

where

$$w(m) = \begin{cases} 1 & (m \leq N) \wedge (p_v(m) > \theta) \\ 0 & (m > N) \vee (p_v(m) \leq \theta) \end{cases} \tag{2}$$

The threshold value for voicing decision θ is usually set close to 1 (0.95 is used in our analysis). This scheme allows us to analyze intermittent (or fragmented) data. The voicing decision function w is sometimes replaced by smoothed decision $\bar{w}$, which eliminates very short segments and noisy data portions near the boundaries.

3.2 Periodic correlation

Let us extend the definition of $\tilde{f}_0$ to represent its periodic counterpart.

$$\tilde{f}_0(k) = \tilde{f}_0(k + nT_p) \quad \text{where} \quad (n = ..., -2, -1, 0, 1, 2, ...) \tag{3}$$

A similar extension is applied to pseudo random sequence s to produce $\tilde{s}$. An extended system impulse response $\tilde{h}$ is then calculated by periodic cross correlation with the periodic average and the pseudo random sequence $\tilde{s}$.

$$\tilde{h}(l) = \frac{1}{T_p} \sum_{k=1}^{T_p} \tilde{f}_0(k)\tilde{s}(k - l) \tag{4}$$

However, $\tilde{h}$ is not directly applicable for estimating the characteristics of auditory and speech production systems, because the f_0 trajectories result from a feedback system. This will be discussed later.

3.2.1 Normalization condition

The pseudo random sequence s has to satisfy the normalization condition.

$$\frac{1}{T_p}\sum_{k=1}^{T_p}\tilde{s}(k)\tilde{s}(k-l) = \begin{cases} 1 & (\text{if } l=0) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

An M-sequence with bias adjustment satisfies this condition.

An additional modification is introduced to avoid non-linear distortion in the modulation of the fundamental frequency. The original M-sequence has a period of 31 time units. This sequence is over-sampled 8 times. The method for oversampling and its problems are discussed in Appendix A. Applying an easy decision to the over sampling parameters makes our measurements using the PN signal unreliable in the frequence region over 8Hz. Note that a better PN signal is introduced in Appendix A.

3.2.2 Data alignment

It is necessary to align the time axis for correlations, because there is no absolute origin in time. Here, the origin is set to the point where the correlation between the perturbation and the actual modulation shows maximum. The actual modulation is computed from the difference between fed-back speech f_0 and produced speech f_0. The difference is equal to s when there is no noise. This definition of the time origin is used throughout this report.

$$n_p = \text{argmax}\{\tilde{h}^{\text{fb}}(n) - \tilde{h}^{\text{out}}(n)\} \tag{6}$$

The aligned response is then re-defined using the maximum position.

$$\tilde{h}_{(align)}(k) = \tilde{h}(k - n_p) \tag{7}$$

3.3 Intermittent data

The power spectrum of an f_0 contour is calculated by using an approximation. Normal speech consists of many disjoint fragments of sound energy. This makes it difficult to analyze frequency characteristics of the fundamental frequency contour. One possible way to work around this is to minimize the truncation effects introduced by intermittent data. In other words, it is equivalent to minimize the discontinuities at the boundaries.

The fragmental nature of speech introduces aliasing caused by truncation. A usual way of working around this kind of aliasing is to use smooth window functions which reduce the amount of interference. But this method is not relevant with varying lengths of fragments. Another method, the one we selected is as follows: First, the signal under examination is equalized before truncation. Second, a frequency analysis is performed on the truncated signal (based on voice/unvoice information). Then, the obtained spectral information is weighted by the inverse characteristics of the equalizer. These procedures are summarized in the following equation.

$$F_{\text{speechF0}} = P^{-1}\mathcal{F}(\bar{w}\acute{f}_0) \tag{8}$$

where $\acute{f}_0$ represents the differentiated version of f_0.

This is different from the following expression, which is a straight application of a short-term Fourier transform.

$$F^{(\mathrm{bad})}_{\mathrm{speechF0}} = \mathcal{F}(\bar{w} f_0) \tag{9}$$

The truncation introduced by w smears out the original spectrum components in this case, because f_0 has a high bias component.

3.4 Periodic averaging for intermittent speech

Periodic averaging also suffers from the intermittent nature of speech. The major component is base line variation due to prosodic components. The effects of truncation are reduced by using the same manipulation as the Fourier transformation case.

$$\widetilde{f_0(n)} = \frac{\sum_{m=0}^{n} \left(\sum_{k=0}^{\lfloor N/T_p \rfloor} \acute{f}_0(kT_p + m)\bar{w}(kT_p + m) - \mathrm{bias} \right)}{\sum_{k=0}^{\lfloor N/T_p \rfloor} \bar{w}(kT_p + n)} \tag{10}$$

where the bias is set to satisfy the periodicity condition: $\widetilde{f_0(T_p)} = \widetilde{f_0(0)}$. The simbol $\acute{f}_0$ represents the differentiated version of f_0.

It is also necessary to estimate the amount of residuals. These residuals consist of (1) a prosodic component, (2) irregularities in vocal fold vibrations, and (3) irregularities in neural control signals. It is reasonable for a first order approximation to assume that the residual calculated from $\acute{f}_0$ and periodic averaging is Gaussian white noise. The next assumption is that the perturbation signal has negligible power when compared with the innate variations. Because there is no reason to believe that a specific frequency component in the fundamental frequency variation will continue with a steady phase, the asymptotic value of the periodic component vanishes when the number of observations goes to infinity. Then, the residuals are approximated by the following method.

$$\begin{aligned} \varepsilon(n) &= \bar{w}(n) \left(\acute{f}_0(n) - \acute{\tilde{f}}_0(n - T_p \lceil \frac{n}{T_p} \rceil) \right) \\ &= \bar{w}(n) \left(\acute{f}_0(n) - \acute{\tilde{f}}_0(n) \right) \\ &\simeq \bar{w}(n) \acute{f}_0(n) \end{aligned} \tag{11}$$

Because we assume that the observation noise is stable and time invariant, only variance is important to assess the accuracy of measurements in the time domain. The effect of periodic averaging on variation of the residual signal can be simplified to the normalization by the number of cycles, based on this independence assumption. The variation of the residual after periodic averaging is represented by the next equation.

$$\tilde{\sigma}^2 = \frac{\sum_{k=1}^{N} \varepsilon^2(k)\bar{w}(k)}{\sum_{n=0}^{\lfloor N/T_p \rfloor} \sum_{k=1} N\bar{w}(k)} \tag{12}$$

The frequency representation of residuals can be calculated using a similar expression.

$$\widetilde{R^2} = \frac{\sum_{k=0}^{\lfloor N/T_p \rfloor} |\mathcal{F}(\varepsilon_{kT_p} \bar{w}_{kT_p})|^2}{\sum_{k=0}^{\lfloor N/T_p \rfloor} \sum_{n=0}^{T_p-1} \bar{w}(kT_p + n)} \tag{13}$$

Where (14)

$$\varepsilon_{kT_p} = \{\varepsilon(m + kT_p)\}_{m=0}^{T_p-1}$$

$$\bar{w}_{kT_p} = \{\bar{w}(m + kT_p)\}_{m=0}^{T_p-1}$$

3.4.1 Confidence interval of observation

Because a dilated set of PN signals is an orthonormal function, variations in white Gaussian observation noise are equally distributed to each component. Let N_{PN} be the order of PN signals. Then, the variance of the observation after manipulation by the PN signals is represented by the following equation.

$$\widetilde{\sigma^2_{PN}} = \frac{\tilde{\sigma^2}}{N_{PN}} \tag{15}$$

It is also necessary to compensate the effects of band limitation in the over-sampled PN signals. This band limitation was introduced to reduce interactions with pitch period variatins in the high frequency region (10Hz or more). Let C_f be the ratio of band limited energy to total noise energy. Then, the variance for a band limited PN signal with periodic averaging $\tilde{\sigma^2}_{BLPN}$ is represented by:

$$\tilde{\sigma^2}_{BLPN} = C_f \sigma^2_{PN} = C_f \frac{\sigma^2}{N_{PN}} \tag{16}$$

Here, C_f plays as a correction factor for band energy reduction by the band limitation; it is dependent on the noise spectrum and the shape of the band limitation filter.

Then, the confidence interval in periodic form is calculated as follows.

$$\tilde{\xi}_\beta = \lambda_\beta \tilde{\sigma}_{BLPN} \tag{17}$$

where λ_β is a coefficient defined by $\int_0^{\lambda_\beta} g(t)dt = \beta/2$, based on the probability distribution function $g(t)$ of normal distribution $N(0, 1)$.

3.4.2 Estimating observation noise

In short, the confidence interval is calculated from the estimated noise variance σ^2_{BLPN}. If we could make a better estimate of σ^2_{BLPN}, the corresponding estimate of the confidence interval would be improved. One feasible method is described here to estimate observation noise with modifications by periodic averaging and by correlation analysis with the PN signals.

First of all, let us assume that the effect of perturbation is much smaller than the observation noise. Then, the pitch deviation of the produced sound is a good approximation of the observation noise.

$$\eta(n) = \frac{1}{T_p} \sum_{l=1}^{n} \sum_{k=0}^{T_p-1} \tilde{w}(l) \acute{f}_0(l) s(n-k) \tag{18}$$

The same technique for reducing the truncation effect is also employed here. The first summation is the inverse function of differentiation, which is used to derive $\acute{f}_0$ from f_0. This procedure introduces a slowly varying component, i.e., an artifact.

Let us derive the noise component by removing bias locally.

$$v(n) = \sqrt{\frac{T_p}{N}} \left(\eta(n) - \frac{\sum_{k=0}^{T_p-1} \eta(n+k)}{T_p} \right) \tag{19}$$

The normalization factor here represents noise reduction by periodic averaging. The variance of band limited, correlated periodic averaged noise is then estimated as follows.

$$\tilde{\sigma_{BLPN}}^2 = \frac{\sum_{k=1}^{N} v^2(k)}{\sum_{k=1}^{N} w} \tag{20}$$

3.4.3 Error estimates in the frequency domain

Similar error estimates are possible under the same independent Gaussian assumption. The probability distribution of a frequency component is also Gaussian under this assumption. Even though the same residuals are added to the input observation and output observation, only the output observation is assumed to consist of noise.

The coherency γ is calculated from the power spectrum distribution of input, output and transfer function G.

$$\begin{aligned} \gamma^2 &= \frac{|G|^2 \Phi_{XX}}{\Phi_{YY}} \\ &= \frac{|\Phi_{XY}|^2}{\Phi_{XX}(|G|^2 \Phi_{XX} + \Phi_{NN})} \end{aligned} \tag{21}$$

It is then possible to calculate the confidence interval using these estimates. The difference of this estimation from usual cases is that the input and the output are represented as periodic averages and correlated with the PN signal. These values are calculated from observed values in the following manner. Let x, $y, and\ n$ be the input, the output and the noise, respectively, and let $\tilde{}$ denote a periodic representation. Then, we get the foloowing.

$$\Phi_{XX} = \mathcal{F}\left(\sum_{k=0}^{T_p-1} \tilde{x}(n)\tilde{x}(n+k) \right) \tag{22}$$

$$\Phi_{YY} = \mathcal{F}\left(\sum_{k=0}^{Tp-1} \tilde{y}(n)\tilde{y}(n+k) \right) \tag{23}$$

$$\Phi_{XY} = \mathcal{F}\left(\sum_{k=0}^{Tp-1} \tilde{x}(n)\tilde{y}(n+k)\right) \tag{24}$$

$$\Phi_{NN} = \mathcal{F}\left(\sum_{k=0}^{Tp-1} \tilde{n}(n)\tilde{n}(n+k)\right)$$
$$\simeq R^2 \tag{25}$$

The confidence interval in the frequency domain is represented using coherence γ and transfer function G. First, let us represent the estimate for the transfer function in terms of an absolute value and phase.

$$\hat{G}(j\omega) = |G(\hat{j}\omega)| \exp(j\phi(\hat{j}\omega)) \tag{26}$$

Then, the estimates of variance for the absolute value and phase have the following relations.

$$\sigma^2_{|G|} \simeq |G|^2\sigma^2_\phi \tag{27}$$
$$\simeq \frac{|G|^2}{W_eT}\frac{1-\gamma^2}{\gamma^2}$$

The denominator W_eT is 1, because a rectangular window and normalization by the window length are assumed in this case. W_e represents the correction factor of the effective window length and T is the nominal window length.

If the system consists of a feedback loop, the estimates should be normalized by the power spectrum of feedback gain $|1-G|^2$. The estimates of variance yield the following.

$$\sigma^2_{|G|fb} \simeq \frac{|1-G|^2}{W_eT}\frac{|G|^2(1-\gamma^2)}{\gamma^2} \tag{28}$$

$$\sigma^2_{\phi fb} \simeq \frac{|1-G|^2}{W_eT|G|^2}\frac{1-\gamma^2}{\gamma^2} \tag{29}$$

The confidence interval of the gain and phase are calculated using these variances and the coefficient λ_β. That is,

$$\tilde{\xi}_{|G|fb}(j\omega) = \frac{\lambda_\beta\tilde{\sigma}_{|G|fb}(j\omega)}{\sqrt{N_e}} \tag{30}$$

$$\tilde{\xi}_{\phi fb}(j\omega) = \frac{\lambda_\beta\tilde{\sigma}_{\phi fb}(j\omega)}{\sqrt{N_e}} \tag{31}$$

where λ_β is the coefficient defined by $\int_0^{\lambda_\beta} p(t)dt = \beta/2$, based on the probability distribution function $p(t)$ of normal distribution $N(0,1)$, and $N_e = \sum_{k=1}^{Ndata}\bar{w}(k)$ is the effective data count.

3.4.4 Integrating data from separate measurements

It is sometimes necessary to provide a procedure to integrate data from separate measurements. Periodic averaging makes this easy. Let $\tilde{x_p}$, $\tilde{y_p}, and\ \tilde{n_p}$ represent the periodic averaged input, output and noise data from the p-th measurement. In addition, let N_{ep} represent the effective length (count) of data in the p-th measurement, and M represent the total number of separate measurements. Then, the averaged counterparts of these signals are calculated taking advantage of the fact that the temporal axes are time-aligned.

$$\bar{\tilde{x}}(n) = \frac{\sum_{p=1}^{M} \tilde{x_p}(n) N_{ep}}{\sum_{p=1}^{M} N_{ep}} \tag{32}$$

$$\bar{\tilde{y}}(n) = \frac{\sum_{p=1}^{M} \tilde{y_p}(n) N_{ep}}{\sum_{p=1}^{M} N_{ep}} \tag{33}$$

$$\bar{\tilde{n}}(m) = \frac{\sum_{p=1}^{M} \tilde{n_p}(m) N_{ep}}{\sum_{p=1}^{M} N_{ep}} \tag{34}$$

where

$$(n = 1, ..., T_p)$$
$$(m = 1, ..., N_p)$$
$$N_{ep} = \sum_{n=1}^{N_p} \bar{w}_p(n) \tag{35}$$

The coherency for the whole measurements is calculated using these averaged values. First, let us calculate corresponding power specturum and cross spectrum.

$$\overline{\Phi_{XX}} = \mathcal{F}\left(\sum_{k=0}^{T_p-1} \bar{\tilde{x}}(n) \bar{\tilde{x}}(n+k) \right) \tag{36}$$

$$\overline{\Phi_{YY}} = \mathcal{F}\left(\sum_{k=0}^{T_p-1} \bar{\tilde{y}}(n) \bar{\tilde{y}}(n+k) \right) \tag{37}$$

$$\overline{\Phi_{XY}} = \mathcal{F}\left(\sum_{k=0}^{T_p-1} \bar{\tilde{x}}(n) \bar{\tilde{y}}(n+k) \right) \tag{38}$$

$$\overline{\Phi_{NN}} = \mathcal{F}\left(\sum_{k=0}^{T_p-1} \bar{\tilde{n}}(n) \bar{\tilde{n}}(n+k) \right)$$
$$\simeq \frac{\sum_{p=1}^{M} R_p^2 N_{ep}^2}{(\sum_{p=1}^{M} N_{ep})^2} \tag{39}$$

Then, using these averaged values, we get the estimate of coherency $\bar{\gamma}^2$ as follows.

$$\gamma^2 = \frac{\overline{|\Phi_{XY}|^2}}{\overline{\Phi_{XX}}(\overline{|G|^2 \Phi_{XX}} + \overline{\Phi_{NN}})} \tag{40}$$

In short, the integration of M separate measurements reduces the effective noise variance $1/M$, while maintaining the other components as approximately the same.

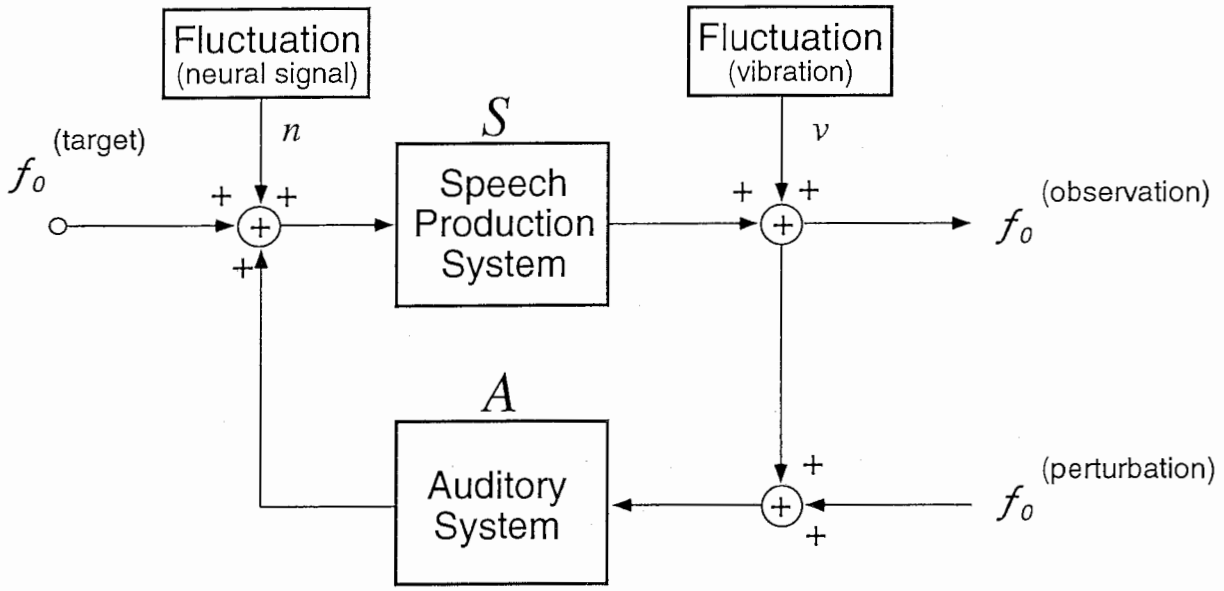

Figure 1: A functional diagram of how measurements are made with Transformed Auditory Feedback.

3.5 Pitch control model

A functional model of how measurements are made with TAF experiments is given in Figure 1[24]. While this model is specific to *Pitch* perturbation, it can easily be adopted to other parametric perturbations. The model is based on a linear system approximation for small perturbations and is formulated in a discrete time system. $A(z)$ represents the characteristics of an auditory system, and $S(z)$ represents the characteristics of a speech production system. Additionally, two types of flirtations, fluctuations in neural commands $n(z)$ and irregularities in vocal cord vibrations $v(z)$, are added to the signals. The variable (z) is not explicitly denoted hereafter.

Let $f^{(o)}$, $f^{(p)}, and$ $f^{(t)}$ be the observed, perturbated and target fundamental frequency trajectories, respectively. $f^{(o)}$ yields the following equation:

$$f^{(o)} = \frac{SA}{1-SA} f^{(p)} + \frac{S}{1-SA}(f^{(t)} + n) + \frac{1}{1-SA} v \quad . \tag{41}$$

The coefficient for $f^{(p)}$ in the first term on the right hand side of this equation, is the value measured in TAF experiments. The correlation between the other terms introduces the estimation error. When there is no perturbation, only the second and third terms remain. Because the denominators of all of the terms are the same, TAF results and non-TAF results share the same pole.

3.5.1 Loop transfer function and feedback response

It is possible to estimate the loop transfer function from speech perception to speech production when contributions from the second and the third term of Eq. 41 are negligible.

In fact, it is represented as an inseparable combination of these systems, $G = AS$. Here, G is the loop transfer function. The loop transfer function will be estimated from two responses.

$$G = \frac{\mathcal{F}(\tilde{h}^{\text{out}})}{\mathcal{F}(\tilde{h}^{\text{fb}})}\ \zeta \tag{42}$$

Here, $\tilde{h}^{\text{out}}$ is the output response and $\tilde{h}^{\text{fb}}$ is the fed-back response. The symbol ζ represents weight based on coherence γ between the input and the output. For this series of specific experiments the weight ζ is defined as follows.

$$\zeta(\omega) = \begin{cases} 1 & (\omega \leq 2\pi(7/f_s)) \\ 0 & (\omega > 2\pi(7/f_s)) \end{cases} \tag{43}$$

The upper limit frequency of 7Hz is selected, because the effective bandwidth of the perturbation signal is 6.25Hz.

The impulse response to pitch perturbation under an open loop condition is directly calculated by the inverse Fourier transform of G.

$$h^{(open)} = \mathcal{F}^{-1}(G) \tag{44}$$

The impulse response to pitch perturbation under a closed loop condition is calculated by the inverse Fourier transform of the feedback transfer function which is represented in terms of G.

$$h^{(close)} = \mathcal{F}^{-1}\left(\frac{G}{1-G}\right) \tag{45}$$

This indicates that the response measured by TAF method is an estimate of this closed loop response.

The response to a step perturbation can be calculated using these responses. For example, the closed loop response to a step input is estimated as follows.

$$r(n) = \sum_{k=0}^{n}(h^{(close)}(n) + \frac{\text{target}}{T_p}) \tag{46}$$

In this case, the target is an appropriately assumed resting value, because the TAF experiments will not give any information about steady state. If the composite system of speech perception and speech production can totally compensate perturbations, the target should be set to -1.

3.5.2 Spectrum estimation by MAICE with AR model

In a series of TAF experiments, the observed response to a perturbation looked like a single negative response. This type of response can be modeled by an auto regressive (AR) model followed by a shaping filter.

If the effects of the shaping filter are equalized, the AR model will represent the power spectrum of the fundamental frequency trajectory. If the major source of correlation in the fluctuation is of auditory origin, there should be a pole (spectral peak) in the frequency region where the phase component of the estimated transfer function crosses

the zero phase. If there are other feedback mechanisms, there will be other poles and zeros. Parameter estimation of AR models is a common technique in speech analysis [10], but it is note directly applicable to analyze fundamental frequency deviations, because there are no a-priori knowledge about the order of AR modelling of natural fluctuations.

A measure called AIC (Akaike's Information Criteria) can be used to determine the opitimum analysis order [1, 2]. The rest of this section summarizes the method to apply AIC for optimum determination of the order of AR models [37].

Let the logarithmic likelihood function of a probabilistic model $f(x|\theta)$ be $L(\theta|x)$, where x represents the observation and θ represents the model parameters. Then, AIC is associated with the maximum likelihood estimate of parameters $\hat{\theta}$ by the following equation.

$$AIC = -2\log\left(L(\hat{\theta}|x)\right) + 2p \tag{47}$$

where p represents the number of parameters of the model. Minimum AIC Estimation (MAICE) is a method to estimate model parameters using the order determined as the one minimizing AIC. MAICE gives the best possible model within the given framework.

Because the pitch control model suggests feedback control, an auto regressive (AR) model is analyzed first. A general AR model is defined by a set of predictor coefficients and a noise source, namely $\theta = \{\alpha_1, \alpha_2, ..., \alpha_m, \sigma^2\}$. The maximum likelihood function of these parameters has $\hat{}$ on it. The log likelihood function of the AR model is approximated using the following equation.

$$\log\left(L(\hat{\theta}|x)\right) = -\frac{n}{2}\log 2\pi\hat{\sigma}^2 - \frac{n}{2} \tag{48}$$

where

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}\left(x_t - \sum_{k=1}^{m}\hat{\alpha}_k x_{t-k}\right)^2 \tag{49}$$

Substituting this value into the definition of AIC, we get the following equation.

$$AIC_{\mathrm{AR}}(m) = n(\log 2\pi + 1) + n\log\hat{\sigma}^2 + 2(m+1) \tag{50}$$

The constant term in Eq. 48 can be ignored because it is independent of the order of the model. The effective amount of data n is determined as $n = N - m$, and the number of parameters p is defined as $p = m + 1$ for AR model estimation.

An example of AR model analysis of a fundamental frequency trajectory is shown in Figure 2. The top left figure shows the power spectrum of the f_0 trajectory. The top right figure shows AIC variation versus the order of the AR model. The bottom left figure shows the original power spectrum and the MAICE power spectrum. The bottom figure panel shows the extracted poles.

3.6 Practical considerations in pitch extraction

It is neccessary to introduce a post processing of fundamental frequency trajectories, because sometimes pitch extraction algorithm produces a half or a double pitch. Such jumps introduce sharp discontinuities in the pitch contour. The pitch extraction program used in

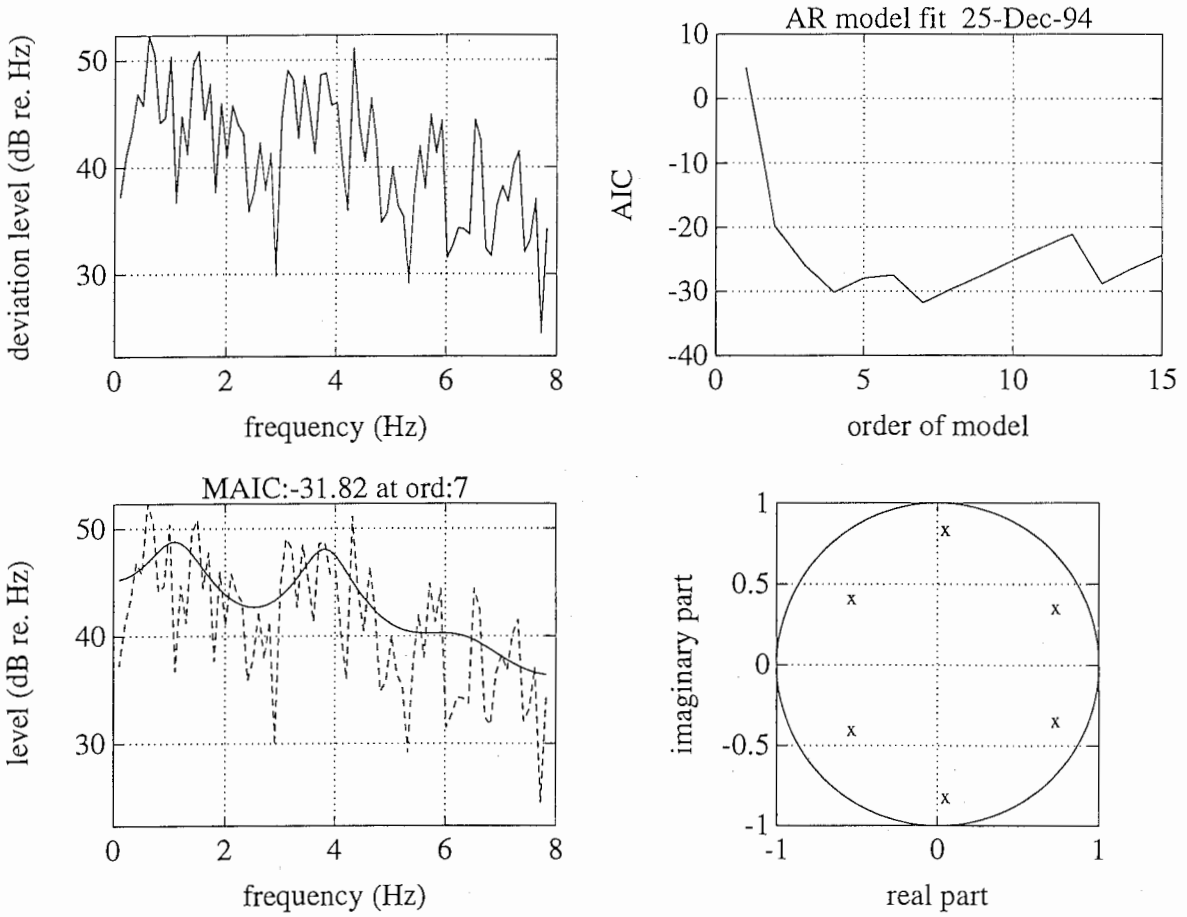

Figure 2: An example of AR model analysis of a fundamental frequency trajectory. The lower left plot shows the observed and the estimated spectrum by MAICE.

our experiments was 'formant' and 'get-f0' procedures in ESPS system by Entropic. These procedures consists of an integrated post processing of pitch trajectories [41]. However, there still remained a double and a half pitch problems.

For sustained vowels, target pitches are constant. Using this constraint, the following post processing can be introduced.

$$f_0^{\mathrm{P}}(n) = \begin{cases} f_0(n)/2 & (f_0(n) > 1.4\bar{f}_0) \\ f_0(n) & (0.7\bar{f}_0 \geq f_0(n) \leq 1.4\bar{f}_0) \\ 2f_0(n) & (f_0(n) < 0.7\bar{f}_0) \end{cases} \tag{51}$$

After this post processing, f_0^{P} is differentiated so that other pitch extraction errors can be detected. The indicator of voicing portion w is reset when the differentiated signal shows a sharp jump.

$$w(n) = \begin{cases} 0 & (|\acute{f}_0(n)| > 0.1\bar{f}_0) \\ w(n) & \text{otherwise} \end{cases} \tag{52}$$

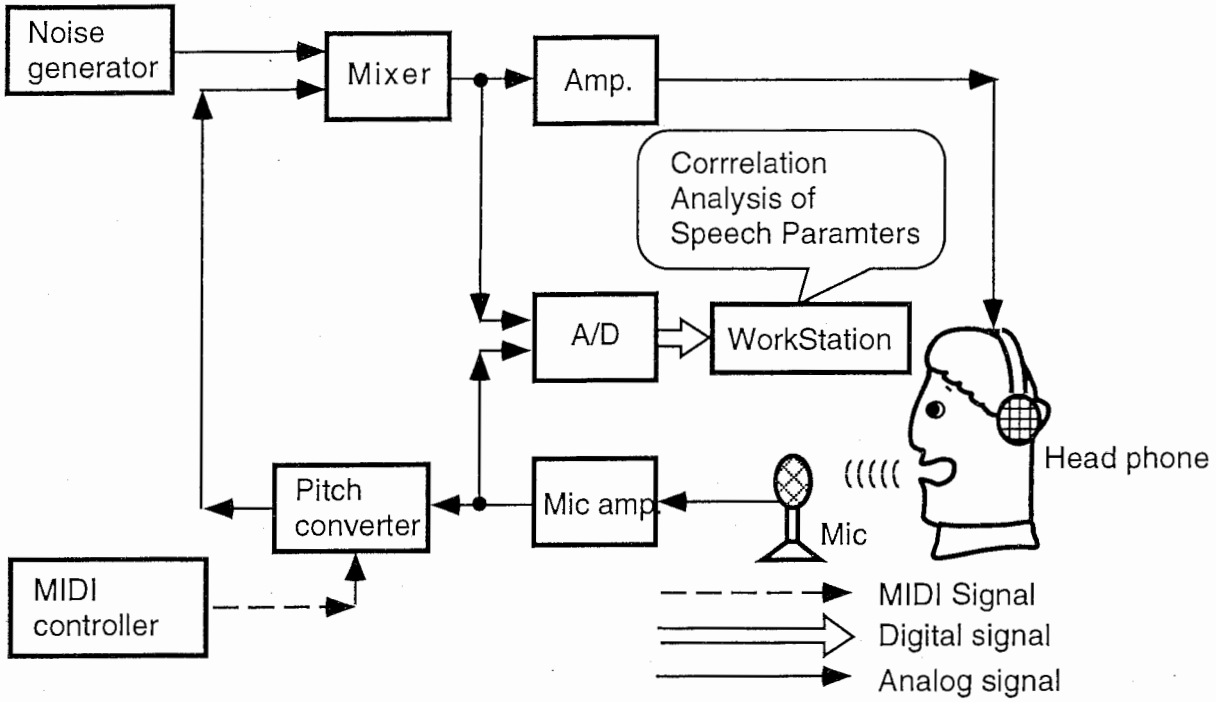

Figure 3: An schematic diagram of how measurements are made with Transformed Auditory Feedback.

4 Experiments and data

This section describes a series of experiments and its focus and presents results analyzed by new procedures.

4.1 General description of experimental conditions

The initial set of TAF is designed to test interactions in fundamental frequency control. Figure 3 shows the block diagram. Table 1 lists the equipment used in these experiments.

The key component is the H3000S Harmonizer utilized as a programmable pitch converter. The pitch conversion algorithm used in the equipment was not open, but used a waveform-based method to perform conversion. The delay introduced by this conversion did not exceed 10ms on average.

Throughout the experiments, pseudo random sequences were used as the perturbations. Specific perturbation signals were generated by the procedure given in Appendix A.

When the gain of the artificial acoustic feedback path was set about 15dB to 20dB higher than the normal side tones, pink noise was added to produce the masking noise of about 80dB(A) with circumaural headphones to prevent interference by the natural side tones.

device	type
A/D, D/A converter	Ariel Pro-port
Microphone	Sony EMC-959 DT
Headphones	Sennheizer HD-250 linear II
Pitch converter	Eventide H3000S Harmonizer
DAT recorder	Sony DTC-1000ES
Noise source	B&K type 1049
Mixer	Sony MX-1000ESX
MIDI controller	Mark of Unicorn Performer
MIDI keyboard	YAMAHA DX-7
Amplifier	Sony TA-E901

Table 1: List of equipment used in the TAF experiments.

4.1.1 Description of figure format

The results are represented in two standard forms. The first form is the integrated representation of the transfer function analysis and the power spectrum information. This representation is the main one used throughout this report. The second form is the power spectrum representation of the fundamental frequency trajectory. The second representation is used when no perturbation exists. It is also used when a higher resolution is necessary for the spectral display.

An example of an integrated transfer function display is shown in Figure 4. The figure consists of eight parts. The very top figure represents the power spectrum of fundamental frequency trajectories. The frequency resolution was set to 100/193.75 Hz. The solid line represents the data of the produced speech. The dash-and-dot line represents the data of the fed-back speech. The break line represents components in the produced speech. The components mentioned above had linear dependencies on the input perturbation.

The left figure in the second row shows the time aligned cross correlation between the perturbation signal and the fundamental frequency trajectory of the fed-back speech. This figure shows how well the perturbation is actually applied to the subject's auditory system. The vertical axis is normalized to represent the instantaneous maximum deviation caused by the perturbation.

The middle left plot shows the time aligned cross correlation between the perturbation signal and the fundamental frequency trajectory of the produced speech. This figure shows the response to the perturbation shown above. The solid line represents the calculated value. The dash-and-dot lines represent the 90% confidence interval for the estimation of the response. 90% of the true response is expected to be within these boundaries.

The right figure in the second row shows the coherency γ^2. This plot gives a rough idea of the reliability of the measurement in the frequency domain.

The right figure in the third row shows the gain component of the loop transfer function $|G|$. The solid line represents the estimated gain. The region between dash-and-dot lines represents the 90% confidence interval for the estimated gain. The lower bound may be

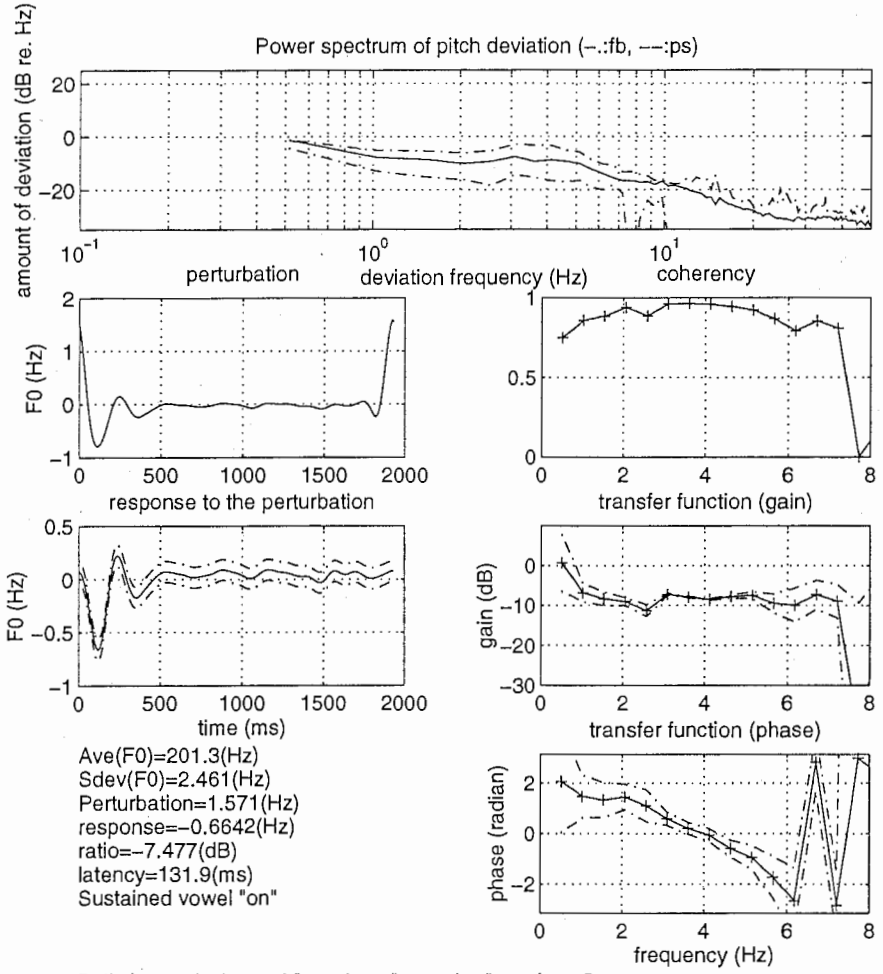

Figure 4: An example of an integrated display of the new analysis procedures. Details are given in the text. This example is from the latest long vowel experiment.

misleading, because the lower bound is drawn at a position symmetric to the upper bound relative to the solid line.

The bottom figure shows the phase component of the loop transfer function ϕ. The solid line represents the estimated phase. The region between dash-and-dot lines represents the 90% confidence interval for the estimated phase. The closed loop transfer function can be derived from these estimated gain and phase values; it is not represented here.

The bottom left part gives statistical information. They are
(1) the average f_0,

(2) the standard deviation of f_0,
(3) the maximum of the averaged perturbations,
(4) the minimum of the averaged responses,
(5) the ratio of (4) to (3) and
(6) the latency of the fast response.
"Sustained vowel 'on' " on the next line indicates when the post processing of a double and a half pitch was performed.

Bellow all of these are the file and directory information and the analysis date information. The effective data length is also shown here. The subject names are encrypted in the plots in Appendix for privacy.

In general, power spectrum plots consist of three parts. Figure 5 shows an example. The data description part is at the bottom of the figure; the names of the subjects and similar information are encrypted to prevent possible leak. This part consists of the data path name history, data file name history, average fundamental frequency, its standard deviation, the analysis status and the date of analysis.

The very top figure represents the power spectrum of the fundamental frequency trajectory. The frequency resolution was set to 100/193.75 Hz. The solid line represents the data of the produced speech. The dash-and-dot line represents the data of the fed-back speech. The break line represents components in the produced speech.

The middle figure shows the ratio of the input to the output power spectrum. In this case, the frequency resolution was also set to 100/193.75 Hz.

4.2 Experiments from 1993.1 through 1993.3

This series of experiments was performed by Mr. Iwatani of Toyohashi Institute of Technology during his internship under my supervision [11, 13, 14, 16, 17, 18]. The basic part of TAF methods was developed in this period. The experiments consisted of standard DAF (Delayed Auditory Feedback) conditions and sentence materials and involved TAF experiments with sustained vowels.

In this section, only data from TAF experiments is analyzed. The conditions are listed in Table 2.

4.2.1 Review of the results

Our previous reports published the following.
(1) There is a compensatory response with about 150ms of latency to a perturbation on the fundamental frequency.
(2) The latency depends on the subject and the pitch.
(3) The transfer gain measured by sinusoidal perturbation signals suggests that these responses are weak (around -20dB) around 5Hz.

The results of the first series of experiments are illustrated in Figures 6 and 7. The figures are summary plots. The result of one subject from an EMG experiment (to be discussed next) is replicated. Integrated displays for all subjects are given in Appendix. The fast compensatory response reported in our previous publications is clearly shown in

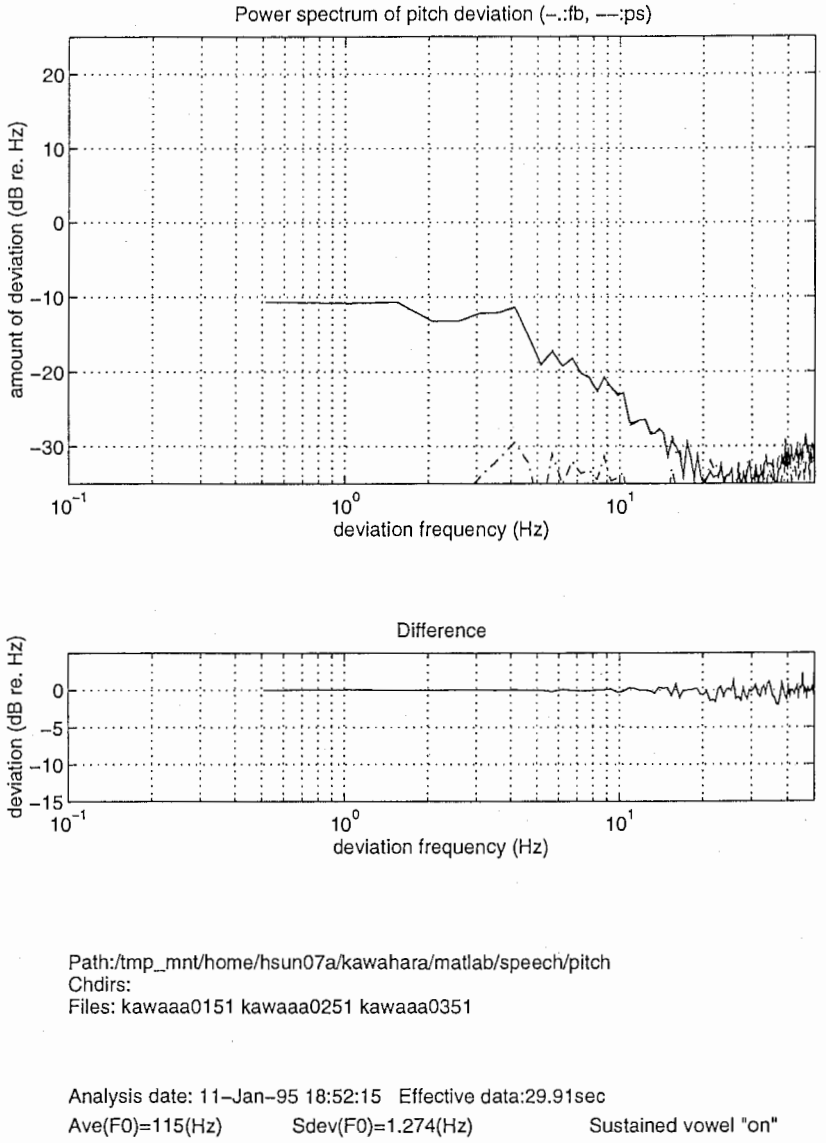

Figure 5: An example of the power spectrum display. This figure shows the power spectrum of fundamental frequency deviations under no-TAF conditions.

these individual plots. Therefore, our previous description "compensatory response with about 150ms latency" seems valid, because the estimated phase characteristics are almost linear in the 2Hz to 6Hz region.

One important point that can not be found in the previous reports is the slow and band limited response. Some boost in gain at the lower frequency region is often observed in those plots. Additionally, a second negative response to the perturbation sometimes appears. These suggest that there is a feedback process that is mediated by a cognitive process. The specific time constant associated with the response may be around 400ms.

factor	level	factor ID
Subject	Kawahara Hideki	kawa
	Obara Kazuaki	obar
	Iwatani Satoru	iwat
	Kawakami Rie	kawk
	Yamada Reiko	yama
Subject (extra)	Tsuzaki Minoru	tsuz
	Kawahara Makiko	maki
	Kawahara Yurika	yuri
Pitch	high	55
	natural	53
	low	54
	without headphones	51
Repetition	1st	01
	2nd	02
	3rd	03

Table 2: Experiment conditions for the first series of TAF experiments. A data file name is a concatenation of factor IDs: `<Subject><Repetition><Pitch>`. The extra subjects were tested only under the natural pitch condition.

The second finding in our previous reports was also found to be valid. Dependencies on subjects and pitch consistently appear. Generally, the phase characteristics around 2Hz to 6Hz are steeper when the pitch is low. The steepness of the phase and the shape of the response waveform vary from subject to subject.

The third finding in our previous reports was too rough to describe the new results. But it is not inconsistent with the new data.

Figure 7 shows estimated latencies from poles of an AR model for voicing under natural feedback conditions (without TAF). The estimated latencies reasonably agree with TAF results. This may suggest that the same control system consisting of an auditory system operates under normal conditions as well as under TAF conditions. This validates the claim that TAF procedures provide a method for extracting parameters of interactions between speech perception and speech production without disturbance.

4.3 Experiments of EMG

The next series of experiments was performed to measure EMG activities under TAF conditions [15]. This series was done during 1993.5 to 1993.7 in collaboration with Dr. Honda, Dr. Kusakawa and Mr. Hirai. Ms. Williams of Ohio State University also participated in this project. Unfortunately, due to a difference in skin characteristics, which prevented any female from producing reliable recordings of EMG, only male subjects were used in this series of experiments.

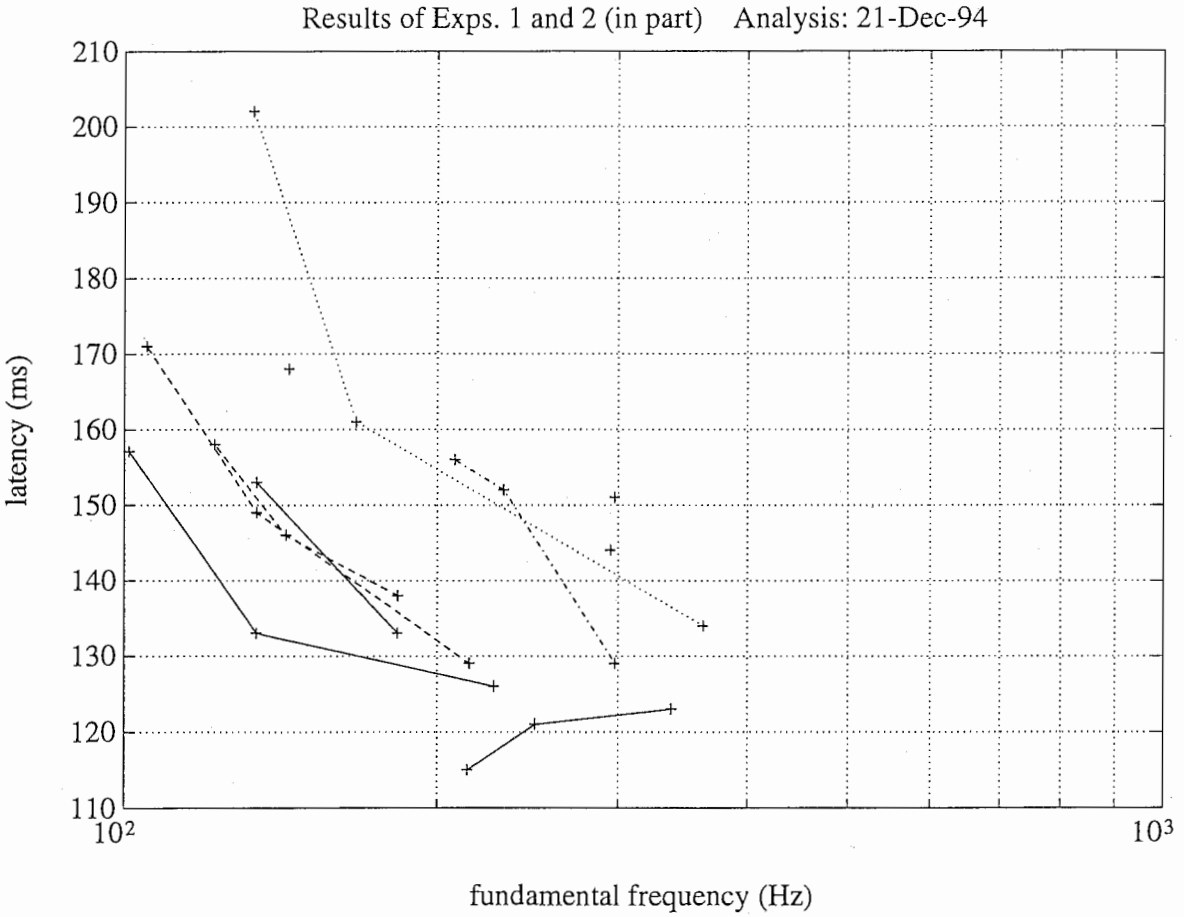

Figure 6: Relation between fundamental frequency and response latency in the first experiment.

Multi channel EMG signals, fed-back speech and produced speech were recorded. Three surface electrodes were placed on each subject's neck and one reference electrode was put on the subject's forehead. The last one differs from common practice, but was used because the typical ear reference point suffers interference from the magnetic field of headphones. EMG signals were converted to an average level signal after low pass filtering and rectification. The window length for the averaging was 10ms. The experimental conditions are listed in Table 3.

4.3.1 Review of the results

Figures in Appendix illustrates a re-analysis of the original data. The effect of pushing the surface electrode does not seem to be very big, because the difference in response latencies between this measurement and the previous measurement for one common subject is less than 20ms.

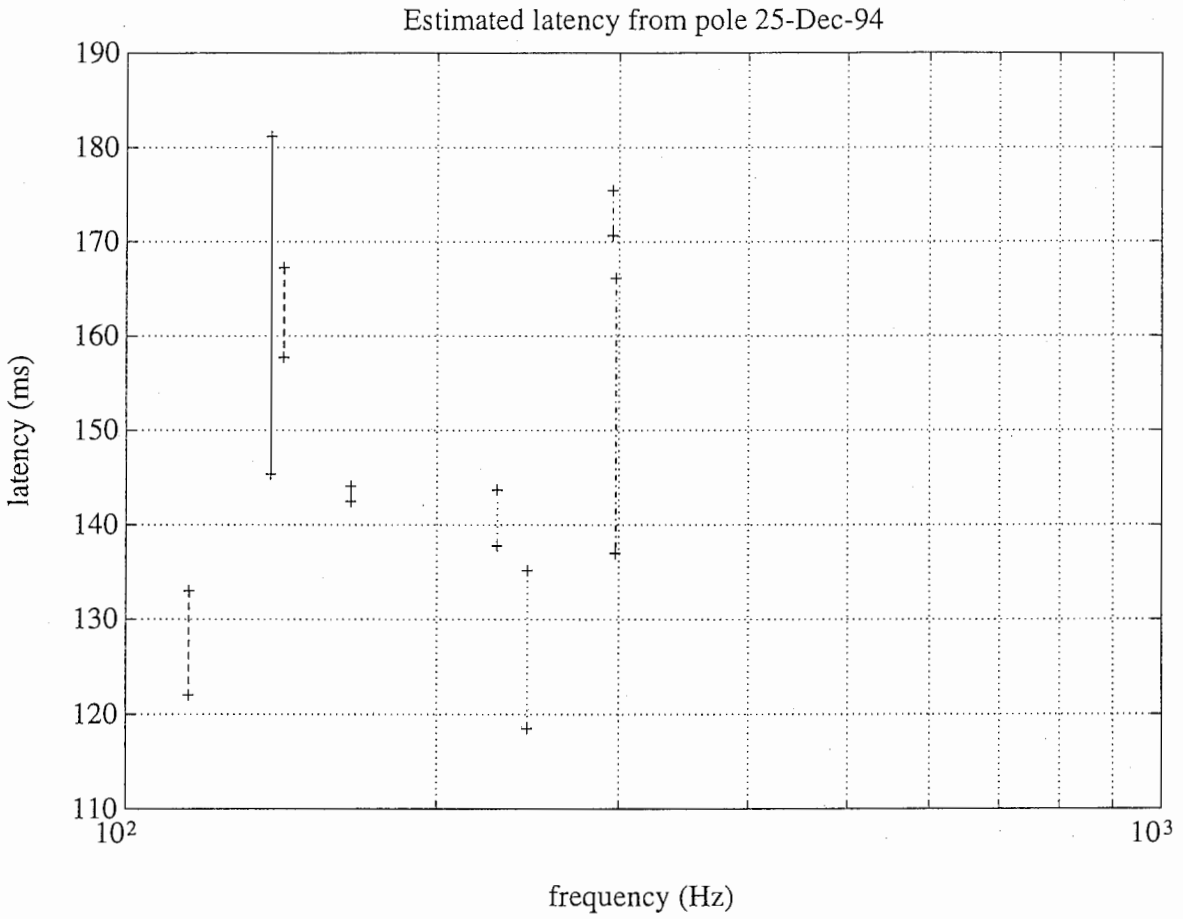

Figure 7: Relation between fundamental frequency and estimated latency from pole frequencies by MAICE analysis using an AR model for voicing under natural feedback conditions (without TAF).

The re-analysis was performed using the same periodic averaging procedure. In this series of experiments, output signals analyzed were EMG signals and f_0. One subject out of four could not be analyzed reliably. The others produced bi-phasic responses to the perturbation. The estimated phases showed a bias of approximately π radians.

A summary of these analyses is illustated in Figure 8. The top left represents the relation between frequency response latency and pitch frequency. The top right represents the relation between response strength and pitch frequency. The bottom left shows the relationship between EMG response latency and pitch frequency. The bottom right plot represents the relation between response latency and response strength.

The definition of EMG latency is the same as that of pitch latency, from the maximum position of the purterbation to the minimum position of the response. This definition may be misleading, because the bi-phasic response and the phasic bias in the estimated transfer function may suggest that our auditory system makes use of the fundamental

factor	level	factor ID
Subject	Kawahara	kawahara
	Hirai	hirai
	Honda	honda
	Komori	komori
Pitch	high	H
	natural	N
	low	L
Repetition	1st	1
	2nd	2
	3rd	3

Table 3: Experiment conditions for the second series of TAF experiments. A data file name is a concatenation of factor IDs: `<Subject><Pitch><Repetition>`.

frequency movement rather than the value itself.

The EMG signal shown here was detected just in front of the CT muscle. The other electrodes did not produce reliable results. Even for the illustrated plots, therefore the reliablilty of the EMG signal is marginal.

4.4 Experiments of hemispheric dominance

The third series of experiments was performed by Mr. Urakami of Ryukoku University during his summer internship under my supervision [45, 20]. This series was done during the end of 1993.8 to the beginning of 1993.9. Prof. Norman D. Cook of Kansai University also participated in this investigation while he was a visiting researcher of ATR HIP from 1993.9 to 1994.3.

The experimental conditions are listed in Table 4. In monaural presentations, pink noise was presented to the opposite ear to reduce the effects of natural side tones.

The level here represents the sound pressure level of the masking noise in dB(A). The noise level was controlled by the attenuator of the amplifier. This manipulation also changed the artificial acoustic feedback gain. No instruction was given to the subjects to maintain the voicing level. Instead, they were instructed to produce voice at a comfortable level under the given feedback conditions.

4.4.1 Review of the results

The major findings of this series of experiments are as follows.

(1) There are no significant effects by the sound pressure level of presentation.

(2) There is a difference in response strength between monotic presentation and diotic presentation.

(3) There seems to be differences between presentation to the left ear and presentation to

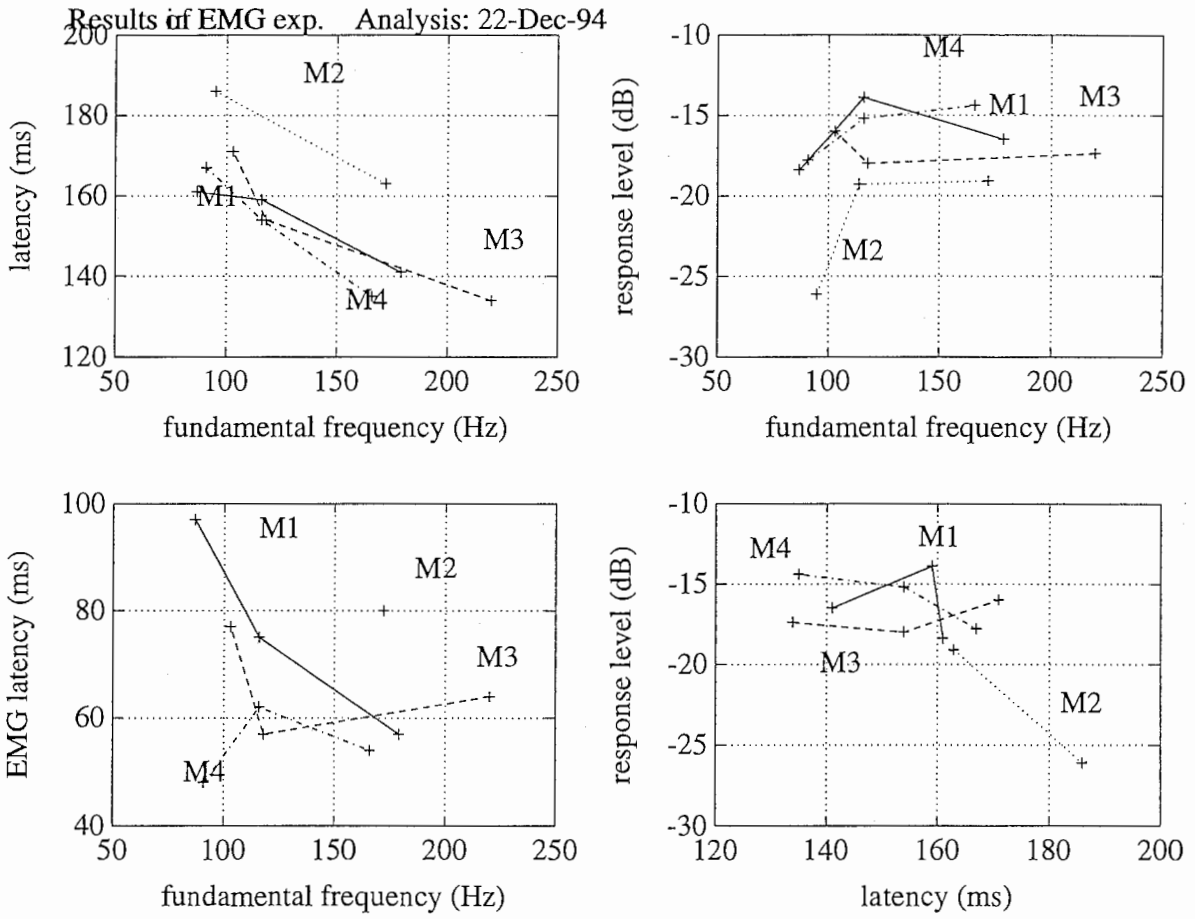

Figure 8: An integrated display of EMG experiments.

the right ear. But they are dependent on the subjects and are not strong enough to be statistically significant.

Figures 9 and 10 show the results of the new analyses. The first finding can be verified by these new results. The sound pressure level effects, however, are not consistent among the subjects. Eight out of ten subjects show no level dependency in response latency. Two subjects show clear increases in response latency with increasing feedback level. This is the opposite of our expectations. Seven subjects show increases in response level with increasing feedback level, but three of them are only slight increases. Three subjects show clear decreases in response level with increasing feedback level. This direction of change is again counter to our expectation.

The second and third points are still not clear. The general trend, however, is the same. The diotic presentation produces slightly faster and stronger responses than the monotic presentation for nine subjects. But one subject shows completely opposite effects. The weakening of responses and increases in latency are more dominant in right ear presentation for five subjects. One subject shows slightly opposite changes. Four subjects

factor	level	factor ID
Subject	Kawahara Hideki	hkawa
	Urakami Hidehiro	hurak
	Andrew Lea	anlea
	Aikawa Kiyoaki	kaik
	Tsuzaki Minoru	mtsuz
	Sato Masaaki	msato
	Tanaka Masako	mtana
	Yoshikawa Noriko	nyosh
	Kawakami Rie	rmori
	Yamada Reiko	ryama
Level	70dB(A)	70
	80dB(A)	80
	90dB(A)	90
Ear	Diotic	bi
	Left monaural	ml
	Right monaural	mr
Repetition	1st	1
	2nd	2
	3rd	3
	4th	4

Table 4: Experiment conditions for the third series of TAF experiments. A data file name is a concatenation of factor IDs: `prn<Ear><Level><Subject><Repetition>`

show no clear difference between right and left ear presentation.

The difference between the diotic presentation and monotic presentation may not be the result of a loudness increase due to binaural presentation, because the level dependency is not strong enough to account for this difference. The left and right ear differences can be interpreted to suggest that the pitch perception center is located in the right hemisphere. But it is too dangerous to make a conclusive statement only with these findings.

It is now obvious therefore that the effective length of the data is not enough to detect the possible diffence reliably. Follow up experiments by controlling the actual listening level directly are necessary.

4.5 Experiments of source characteristics

The fourth series of experiments was performed by Mr. Iwazume of Nara Institute of Science and Technology during his internship under my supervision [21, 22]. This was done during 1993.10 through 1994.3.

The major point in this series of experiments is two fold. One point is to use a weaker perturbation to test interactions. The second point is to measure effects by differentiat-

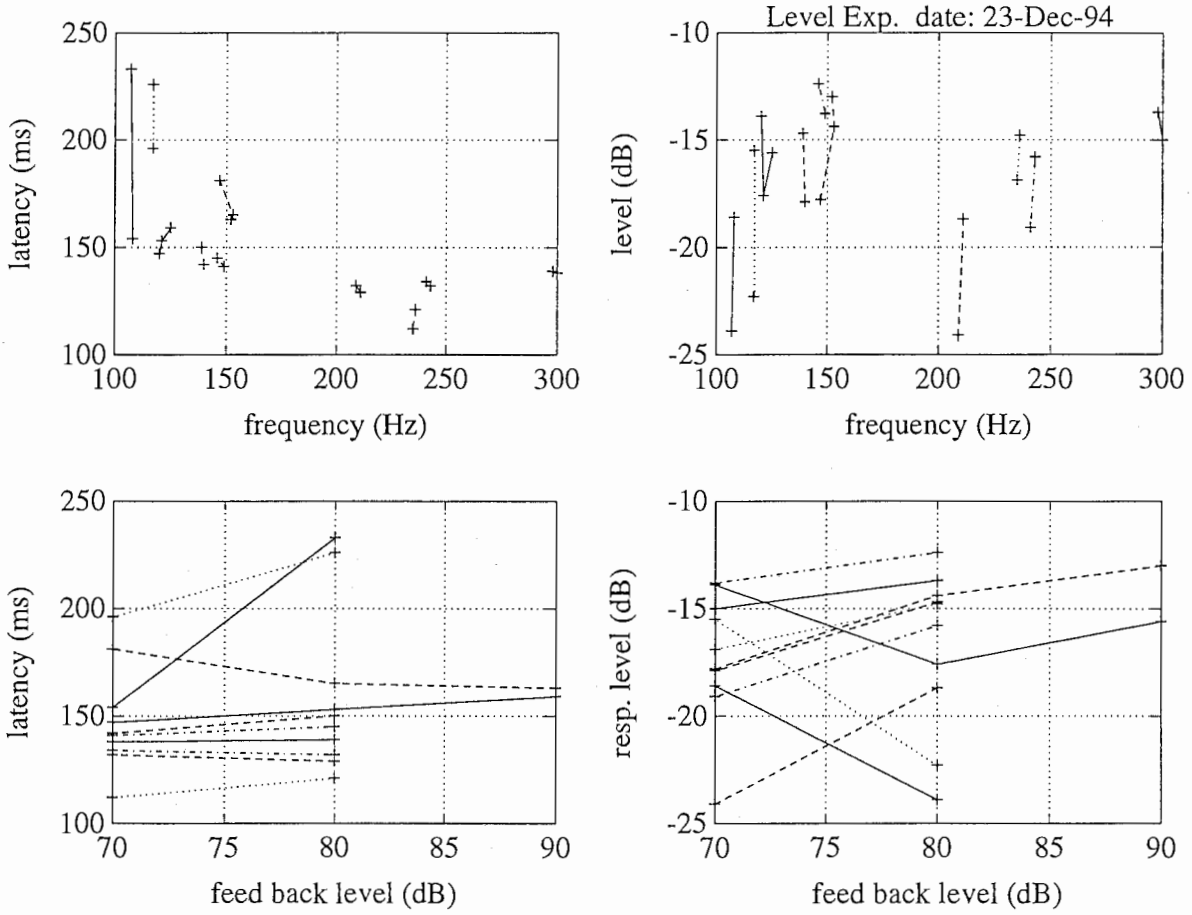

Figure 9: An integrated display of feedback level effects.

ing the timber of the fed-back speech from the original. This differentiation is done by introducing a constant bias in fundamental frequency perturbations. Non-speech stimuli are also used to differentiate the fed-back speech further.

The experimental conditions are listed in Table 5. In monaural presentations, pink noise was presented to the opposite ear to reduce the effects of natural side tones.

4.5.1 Review of the results

The major findings in this series of experiments are listed below.
(1) Perturbations with a half deviation of their previous perturbation still produce measurable effects.
(2) A systematic increase in timber difference causes an increase in response latency.

The first observation was verified by the new analysis, but there is an important difference between the weak perturbation and the previous perturbation. The response latencies were roughly the same, but a systematic difference in response strength was

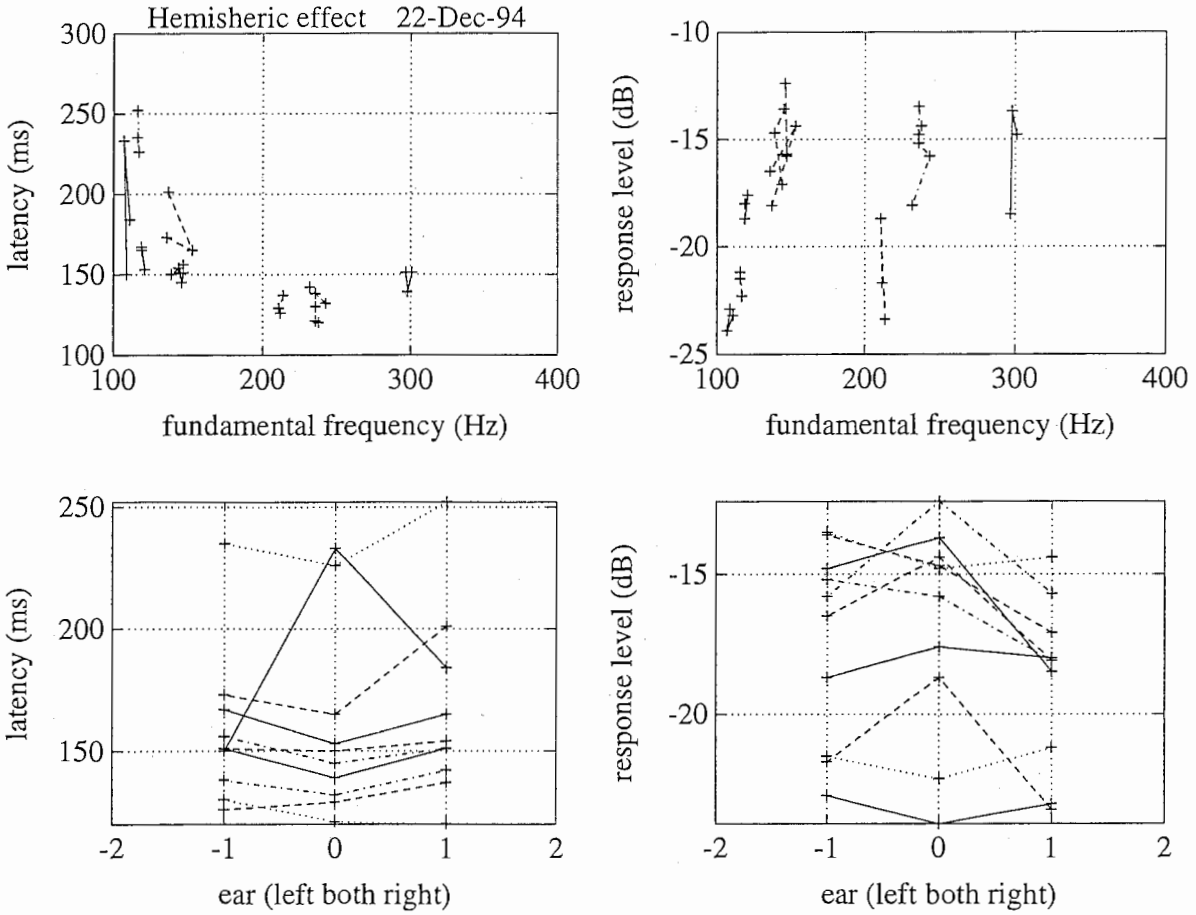

Figure 10: An integrated display of hemispheric effects.

observed in the new results. This will be discussed later.

The new analysis results for a systematic change in timber diffference are shown in Figures 11 and 12. Figure 11 represents the effects of introducing a constant pitch conversion bias. The conversion spans -400cents (76% of the original fundamental frequency) to +400cents (124% of the original fundamental frequency).

The bottom left figure shows the relation between conversion bias and response latency. Except for one subject, the latency increases with increasing difference from the origianl fundamental frequency. The lower right figure shows the effect on the response strength. It seems like the strength decreases when the fundamental frequency increases. The top right figure shows the dependency of the fundamental frequency on pitch shift. An interesting assymmetry is clearly observed for three subjects. The top left figure shows the relation between the fundamental frequency and response latency. The frequency change caused by the introduction of the conversion bias may contribute to the change of the response latency in part. But the vertical trajectory in the top left figure suggests that there still exists a direct effect caused by the difference between intended voice and

factor	level	factor ID
Subject	Kawahara Hideki	kawahara
	Iwazume Michiaki	iwadume
	Yoshikawa Noriko	yoshikawa
	Yamada Reiko	yamada
Source	-400cent shift	-200
	-200cent shift	-100
	0cent shift	NoShift
	200cent shift	+100
	400cent shift	+200
	Pure tone	Pure
	Voice (DX-7)	Control
Repetition	1st	1
	2nd	2
	3rd	3
	4th	4

Table 5: Experiment conditions for the fourth series of TAF experiments. A data file name is a concatenation of factor IDs: `<Subject><Source><Repetition>`.

the fed-back voice.

Figure 12 represents the effects of source difference. The response latency increases with increasing difference between the subject's voice and the fed-back sound. But the response strength does not show a systematic change. It is strange that the response strength seems to have the maximum in case of synthetic voice feedback.

4.6 Experiments on feedback conditions and the model

The fifth series of experiments was performed by Mr. Hirayama of Waseda University during his internship under my supervision [8, 24, 25, 26]. This was done during 1994.8 through 1994.9.

The experimental conditions are listed in Table 6.

This series of experiments tested whether the same response operates under natural (without headphones) conditions.

4.6.1 Review of the results

The major finding in this series of experiments is that the compensatory response found under TAF conditions also operates under natural phonation. This conclusion was derived through a rough modeling of a pitch control process and a sophisticated AR-model based analysis of pitch contour.

The new analysis results make it easyer to understand the power spectral characteristics

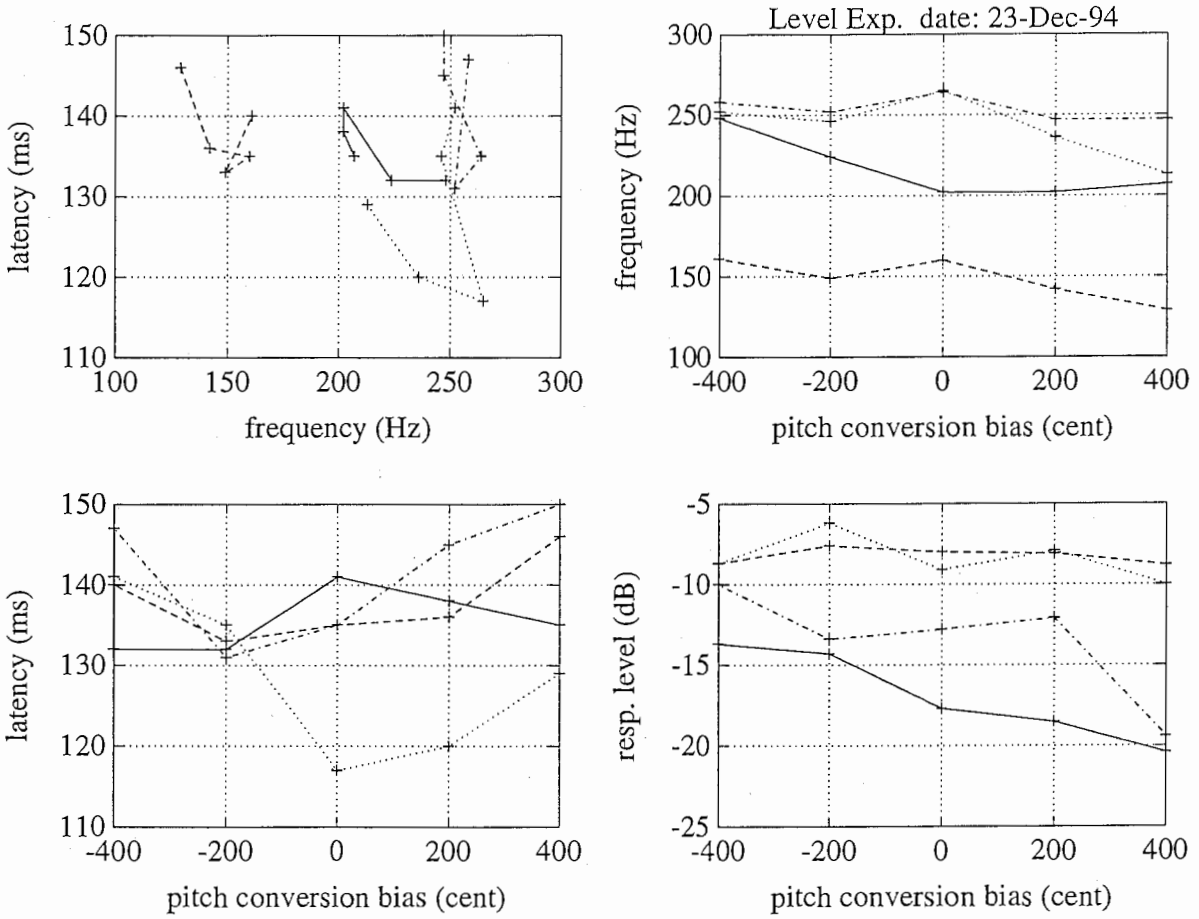

Figure 11: Effects of constant bias in pitch shift on latency and response level.

of pitch trajectories. The phase plot of the estimated transfer function has a zero crossing around 4Hz. This corresponds to the spectral peak in the power spectrum of the naturally phonated pitch trajectory.

TAF experiments were performed to provide a more reliable basis for the fundamental frequency dependency of various response parameters. Some of them are illustrated in Figures 13 and 14.

This series of experiments confirmed the general tendency that the higher the fundamental frequency is the faster the response is. Results by male subjects and female subjects seemed to follow a general curve regardless of gender. This may suggest another way of explaining the dependency of latency on pitch frequency. The major part of the general trend may be a consequence of characteristics of our pitch perception mechanism.

In one experiment, natural voicing and phonating under DAF conditions were acquired to check spectral characteristics of fundamental frequency trajectories. These trajectories were analyzed by MAICE based on the AR model of pitch control, and latencies were estimated from pole frequencies. Figure 15 shows the relation between fundamental fre-

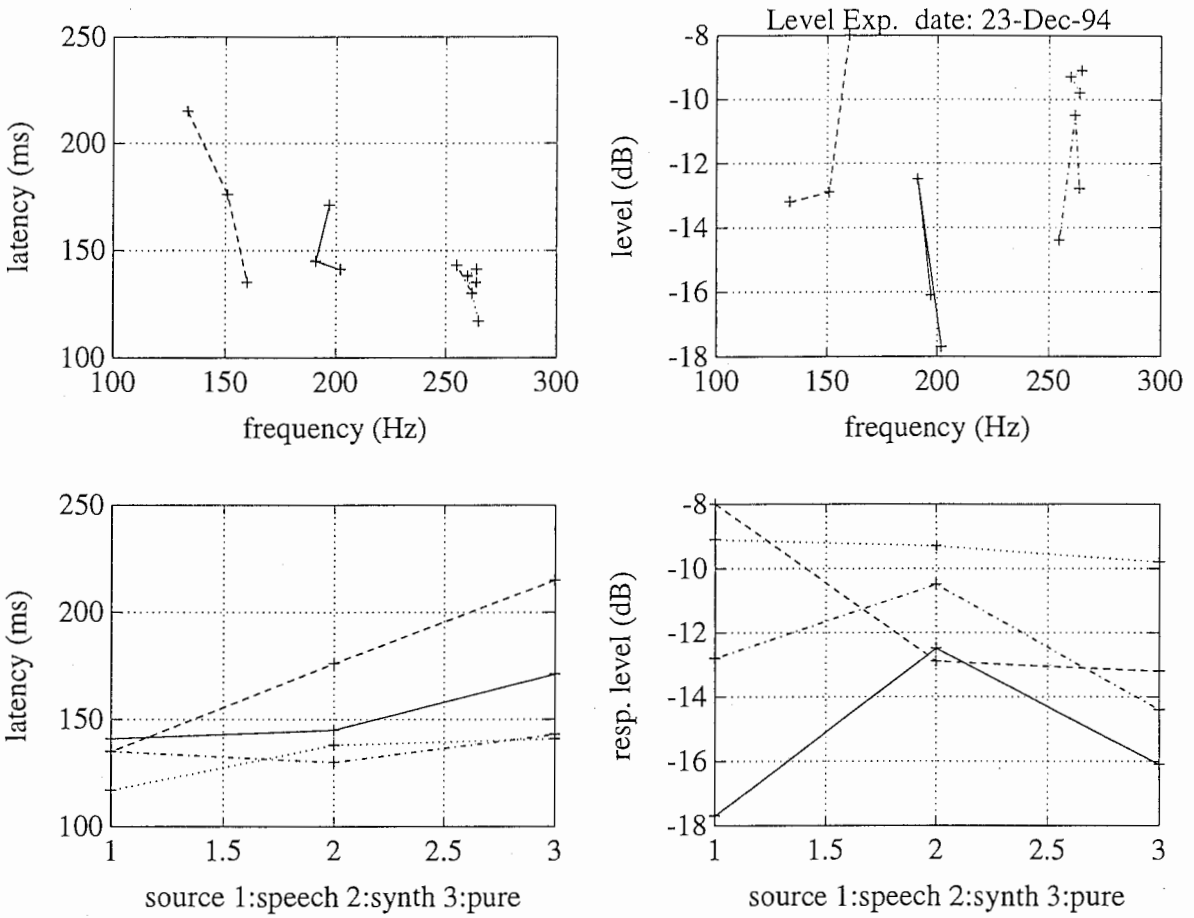

Figure 12: Effects of source difference on latency and response level.

quency and estimated latency. A trend between fundamental frequencies and estimated latencies similar to that observed under TAF conditions was also found in this analysis. This correspondence is important, because it shows that completely different estimation methods and measurement procedures can still produce similar results.

The other important agreement comes from the DAF experiment. Figure 16 shows the relation between inserted delay and estimated latency. This new analysis result replicated the previous finding that the estimated latency increases approximately the same amount as the inserted delay, except for one subject out of six subjects. The re-analysis suggested that it seems like there is a systematic increase in estimated latencies under DAF with a 100ms delay.

This new observation again gives support to our hypothesis that the auditory system actually plays as the role of regulator of the voice fundamental frequency, because close inspection of the phase characteristics obtained from the TAF experiments shows systematic downward deviations from linear relations at lower frequency regions, namely around 2Hz or less. Such deviations correspond to systematic increases in estimated latencies,

factor	level	factor ID
Subject	Kawahara Hideki	TAF-kawahara
	Hirayama Kazuhiko	TAF-kazuhiko
	Nishida	TAF-nishi
	Tanaka Masako	TAF-tanaka
	Morita Rie	TAF-morita
	Yamada Reiko	TAF-yamada
Pitch	G2 (male) or G3(female)	g2
	C3 (male) or C4(female)	c3
	E3 (male) or E4(female)	e3
	G3 (male) or G4(female)	g3
Delay	0ms	0
	25ms	25
	50ms	50
	75ms	75
	100ms	100
Modulation	PN signal	taf
	None (without headphones)	nat
Repetition	1st	1
	2nd	2
	3rd	3
	4th	4
	5th	5

Table 6: Experiment conditions for the fifth series of TAF experiments. There are no systematic relations to the data file names and conditions, because the conditions were randomized to check the ordering effect. Results by TAF experiments with the PN perturbation are analyzed in this report.

because the pole frequency caused by the auditory contribution is located at the frequency where the phase of the loop transfer function is zero, if the amplitude component around there is approximately constant.

In other words, the pole frequency $f_p = 2\pi\omega$ caused by the auditory feedback is the solution of the following equation.

$$\phi(\omega) - \tau\omega = 0 \tag{53}$$

here $\phi(\omega)$ is the phase component of the loop transfer function obtained from the TAF experiments.

Figure 17 shows a simulation of DAF effects on latency estimation. The estimation replicates the essential feature found in Figure 16. The agreement is within the confidence interval estimated from the TAF measurements.

Figure 18 shows a more interesting result. This figure suggests that under certain

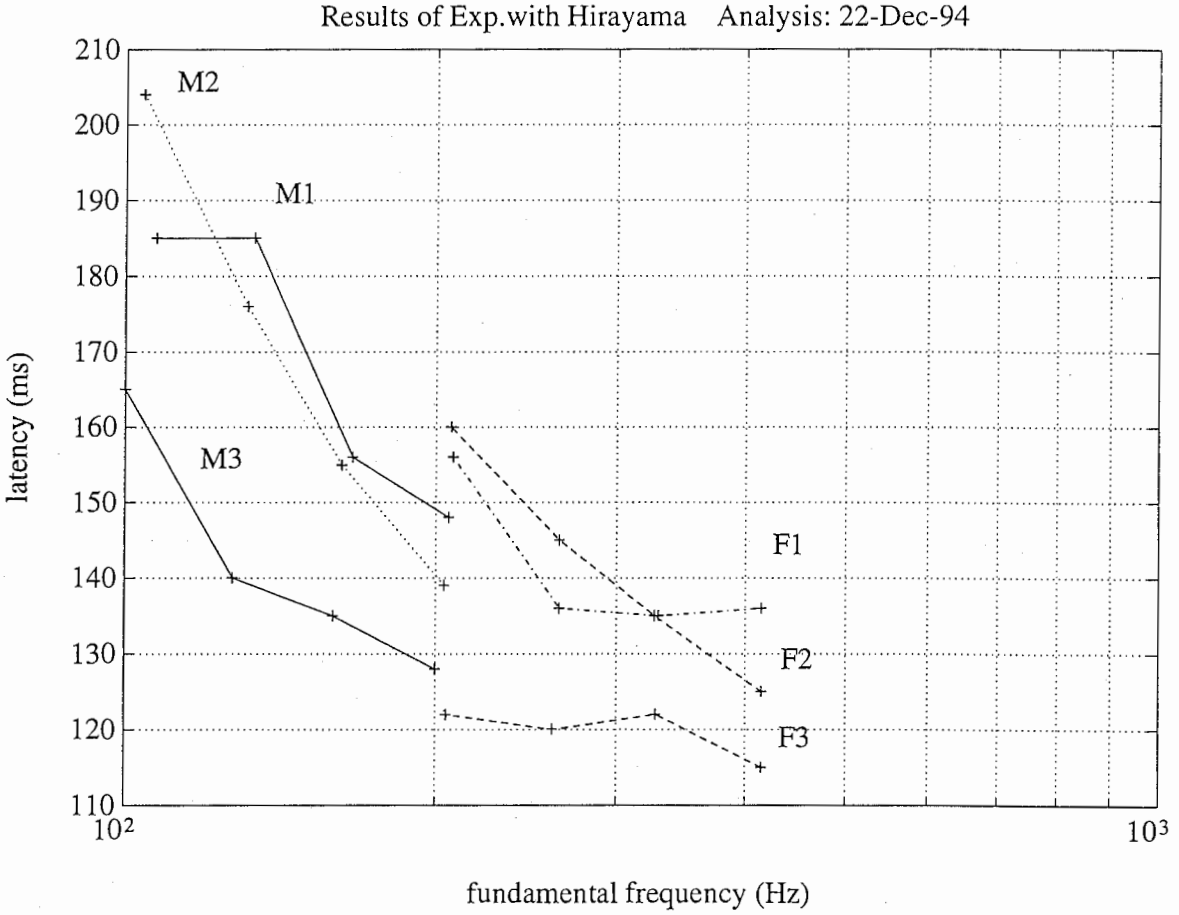

Figure 13: Dependency of response latency on voice fundamental frequency.

conditions, spectral characteristics will drastically change from a single auditory related pole to three poles and vice versa.

This delay-to-delay representation is not very suitable for investigating the validity of the model, because this representation requires the selection of a pole from several poles. Such selection is not free from subjective biases. The more desirable display is a delay-to-pole frequency representation. This display does not require any selection process after the decision of order based on AIC is completed. If our simulation method to predict the auditory induced pole/poles is valid, then some of the estimated poles may be located around the predicted line.

5 Preliminary experiments using read speech

The last series of experiments was performed to test for the existence of the same response in read speech. This was done in 1994.12. This experiment is still continuing and will be

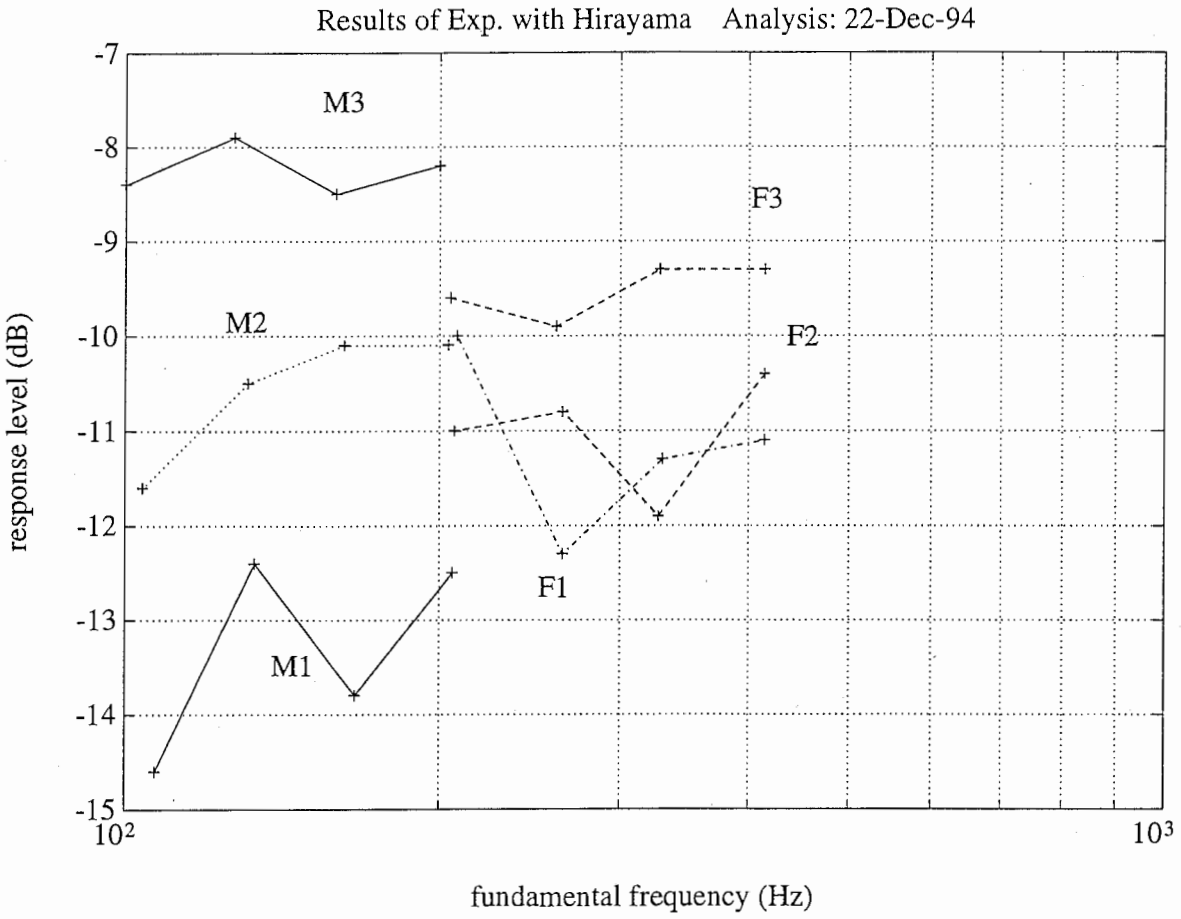

Figure 14: Dependency of response strength on voice fundamental frequency.

reported in detail elsewhere. Several new analysis method is introduced and described in this section.

The experimental conditions are listed in Table 7. In this series of experiments, the unit recording length was set 2 minutes, because a procedure to analyze the intermittent data was available at that time.

5.1 Results

The results of these experiments are all new and not reported anywhere else. The important points are listed below.

(1) One shot analysis of data from long recordings is now possible.
(2) A systematic method is introduced to integrate separate measurements.
(3) Intermittent read speech of over 36 minutes was analyzed.
(4) An instability condition exists in pitch control.
(5) A method is available to separate the information processing duration and mechanical

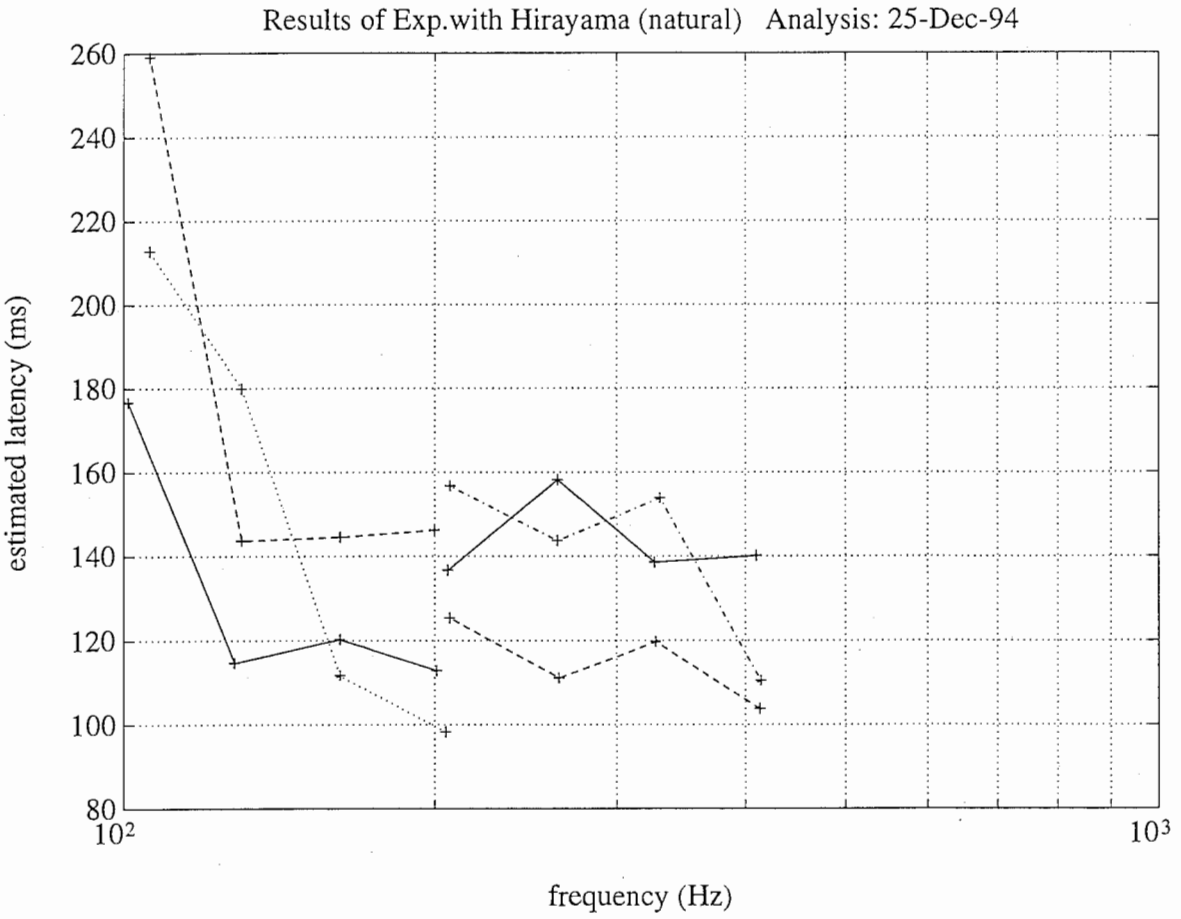

Figure 15: Dependency of estimated latency from MAICE of AR model on voice fundamental frequency.

response time, based on a 2nd order model fitting to impulse responses.

The first and second points make it possible to increase accuracy when necessary. The one shot analysis also reduces possible artifacts by apparent phase alignment between the perturbation signal with the observed signal.

The third points takes advantage of (1) and (2). The result is still close to the confidence limit, but it may be safe to say that the same fast compensatory response to perturbations also operates when speech is read aloud. This may suggest that the same process also operates in spontaneous speech.

The fourth point is a validation test of the new finding, i.e., a slow and band limited response. The fact that all subjects under a specific condition, designed based on the phase plot, make the fundamental frequency control unstable indicates that the finding is not an artifact.

Figure 19 shows one example of such unstability. Under this DAF condition, a 500ms delay shifts phase $-\pi$ at 0.5Hz, which corresponds to the amount of phase shift compen-

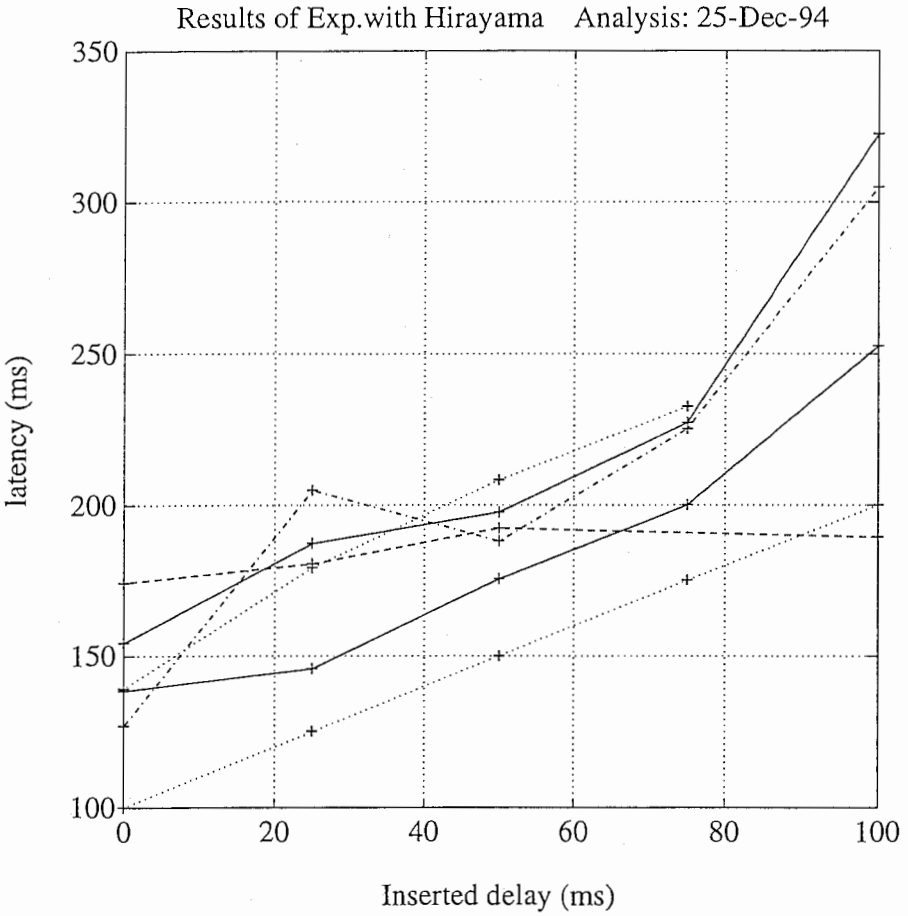

Figure 16: Dependency of estimated latency on inserted delay under DAF conditions.

sating the phase shift of the loop transfer function at 0.5Hz.

5.2 Step response, rise time and processing time

It is possible to convert the estimated transfer function into the step response of the system. This situation is the situation reported by Larson at the last VFPS in Kurume [30]. An example plot of an estimated step response from TAF experiments is shown in Figure 20.

The bottom left figure shows the initial part of the estimated step response. The latencies used in our previous report corresponded to the response delay of the initial part. The delay was 135ms this time. The rising time (time to change from 0.1 to 0.9 of the target value) is about 55ms in this example. This corresponds to a cut off frequency of 5.2Hz for a 2nd order system with a damping ratio of 0.5. If we assume that the response wave form represents the transient of that second order system, the expected delay to a step input is about 45ms. The difference between the measured delay and this

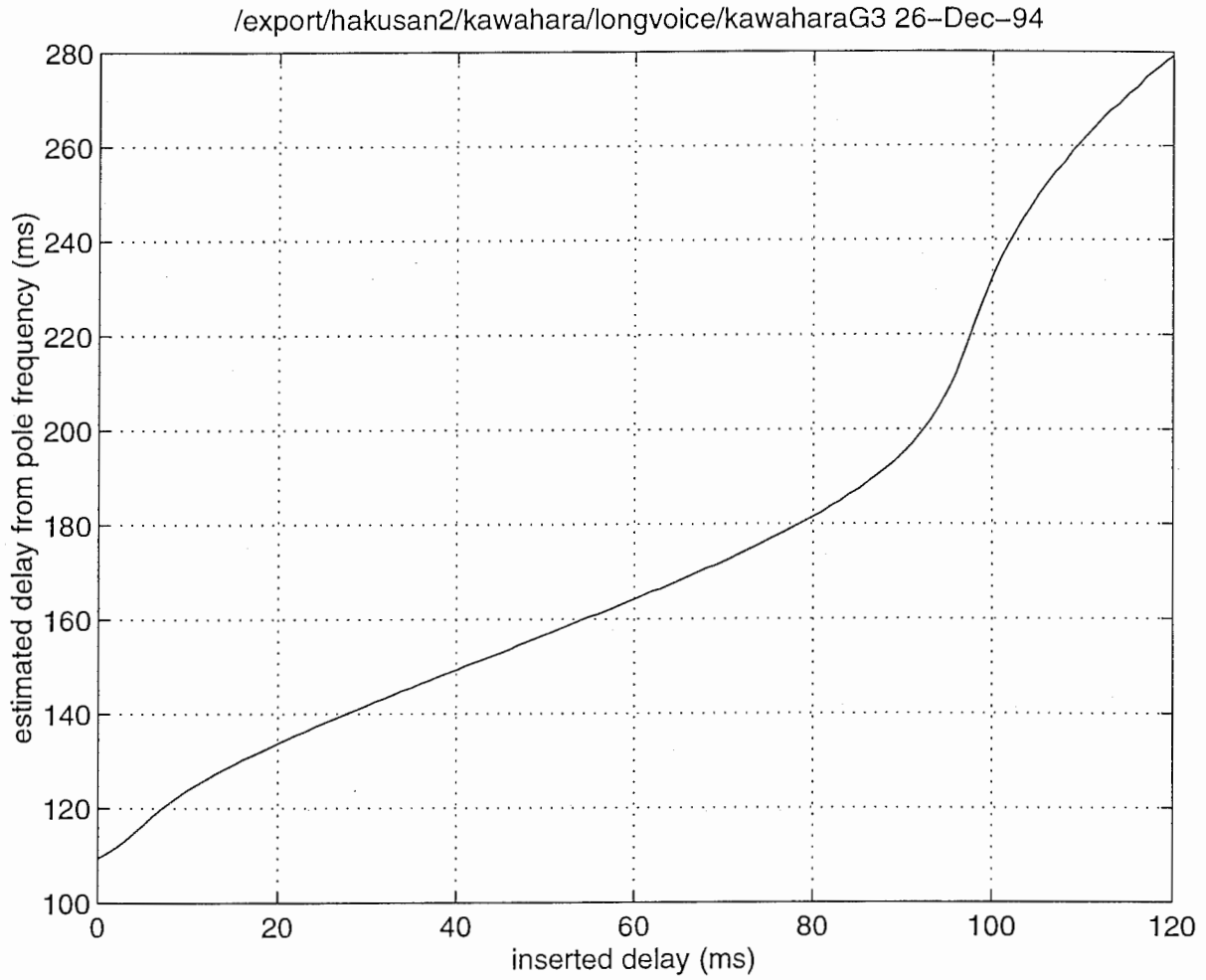

Figure 17: Simulated dependency of estimated latency on inserted delay under DAF conditions. The simulation is based on TAF results.

expected delay may represent the time for information processing in our neural system. The difference in this case is 90ms. This value is somewhat larger than the value indicated by the EMG measurements. One possible explanation is that our auditory-to-production system makes use of frequency change information as well as the frequency itself.

The previous paragraph described the basic idea. But there are several issues to be considered. The estimated impulse response for a closed loop system was used to estimate the step response. This is reasonable to predict an actual step response. But an open loop response should be used to investigate component processes. In addition, A non-linear optimization method should be employed to separate the time for information processing and the mechanical response. It is also ncessary to use pre-processing to eliminate the strong low frequency component generally found in estimated transfer functions.

The optimization is performed to minimize the squared error between the estimated impulse response and the impulse response of the 2nd order system. The parameters to be

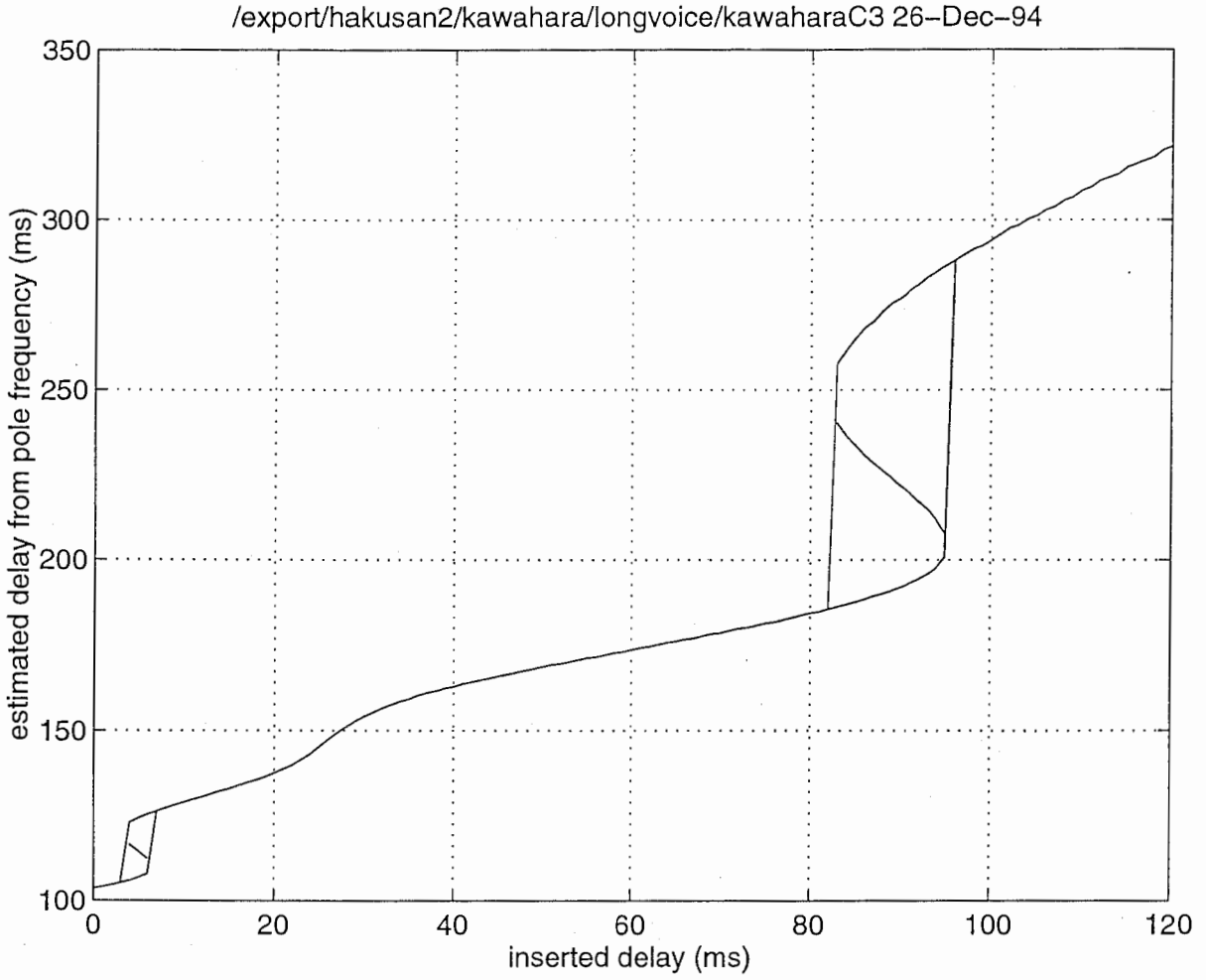

Figure 18: Simulated dependency of estimated latency on inserted delay under DAF conditions. The simulation is based on TAF results. A strange split of an auditory induced pole into three poles can be observed around 5ms and 90ms.

optimized are dumping factor ζ, characteristic frequency ω_n and information processing delay τ. The optimization uses the following template function of a 2nd order system.

$$\text{(i) } \zeta < 1$$

$$g(t) = \frac{\omega_n}{\sqrt{1-\zeta^2}} e^{-\zeta\omega_n t} \sin\sqrt{1-\zeta^2}\omega_n t \tag{54}$$

$$\text{(ii) } \zeta = 1$$

$$g(t) = \omega_n^2 t e^{-\omega t} \tag{55}$$

$$\text{(iii) } \zeta > 1$$

$$g(t) = \frac{\omega_n}{\sqrt{1-\zeta^2}} e^{-\zeta\omega_n t} \sinh\sqrt{1-\zeta^2}\omega_n t \tag{56}$$

factor	level	factor ID
Subject	Kawahara Hideki	kawahara
	Aikawa Kiyoaki	aikawa
	Kato Hiroaki	kato
sentence	/aoiaoinoewa yamanouenoieniaru/	aoia
	/aioinooiwa yamanouenoieniiru/	aioi
	long sustained vowel	[name][note]
	Text from Piaget	Piaget

Table 7: Experiment conditions for the sixth series of TAF experiments.

$$\text{where } g(t) = 0 \qquad \text{for } t < 0$$

The cost function is defined as a squared sum of the observation and the corresponding estimation.

$$L = \sum_{t=0}^{T} |g(t-\tau) - h(t)|^2 \tag{57}$$

The lower right figure of Figure 20 shows the fitted results and the optimum parameters. The information processing delay of 82.7ms gives the best fit. The previous rough estimate was not too bad. This new value is reasonably close to the EMG data.

5.3 Composite model of auditory feedback and decomposition

The impulse response derived from the estimated loop transfer function seems to have a component other than this relatively fast response. Some of estimated transfer functions have a dip around 2Hz in the gain component. This frequency region sometimes corresponds to the low coherency region. Moreover, a downward deviation from a linear phasic relation occurs around the same region. All these suggest that there are two different auditory response systems and they work in parallel.

It could be unrealistic to hypothesize that the response of the slow system behaves like that of a second order system, but it may be a reasonable first step, because a "correcting pitch error proportional to the magnitude of error with some delay in detection, decision and action chain" type strategy is approximated by the second order system.

This line of thought leads to a procedure to decompose the open loop impulse response into two components. The steps are outlined as follows.

(1) Estimate impulse responses corresponding to the loop transfer function. One response uses the whole reliable frequency range and the other uses only the middle frequency range, namely 2Hz to 8Hz.

(2) Estimate optimum fast response from middle range response.

(3) Calculate the first estimate of residual response by subtracting estimated fast response from the whole signal.

(4) Estimate slow response from band limited residual response.

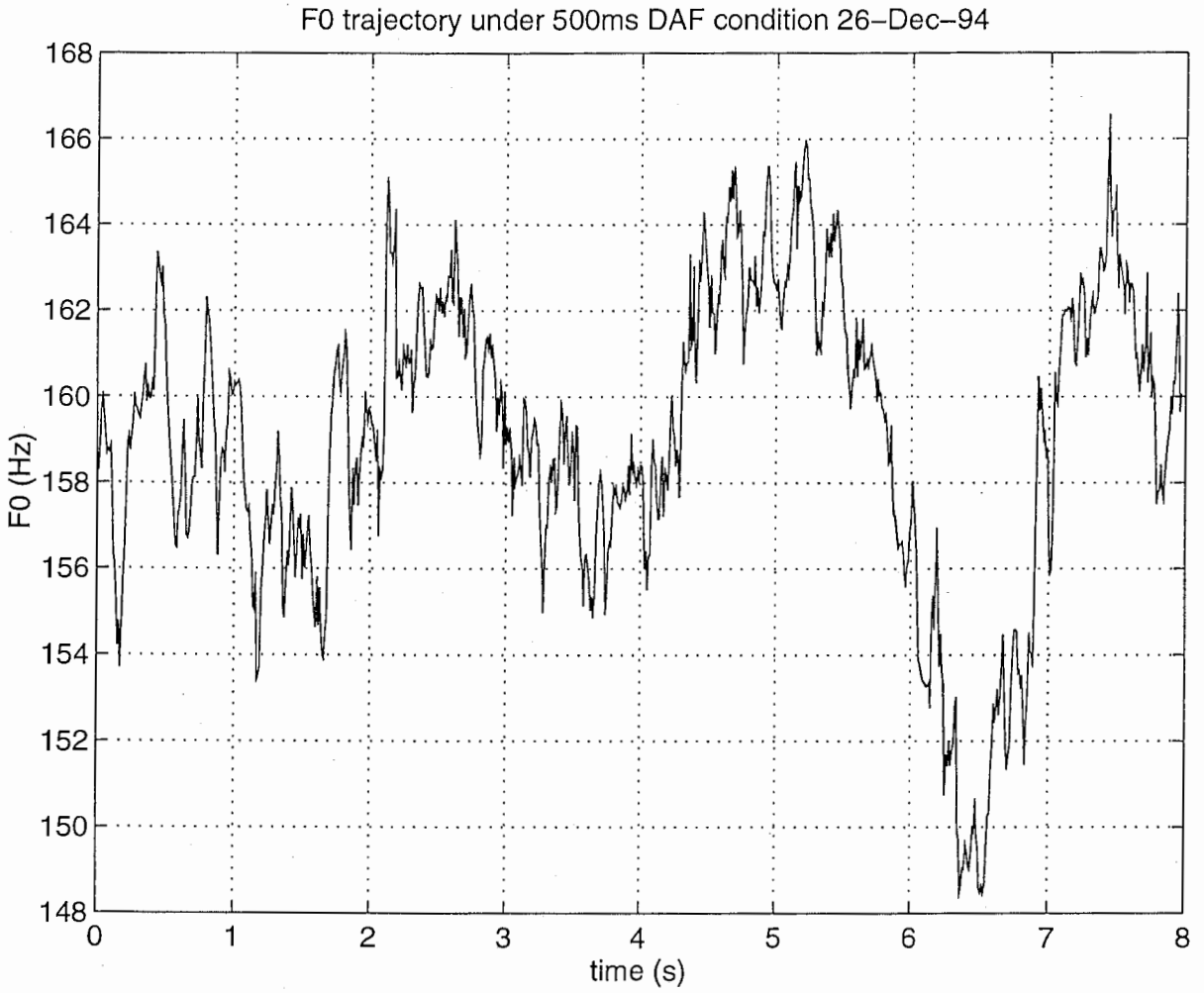

Figure 19: An example of an unstabilized pitch trajectory under DAF with a 500ms delay.

(5) Calculate updated middle range response by subtracting slow response from the whole signal.
(6)Iterate from 2 through 5 to revise estimates.

Some technically difficult problems exist in how to set the intial values for the optimization process. They will be discussed later.

5.4 Initial value for slow response optimization

The first estimate of the residual may consist of a considerable amount of noise especially in a high frequency region. The main characteristics of a slow response may exist in a lower frequency region. The first step is to filter out the higher frequency region using FFT.

If the system is a second order system, AR model estimation can be applied to this

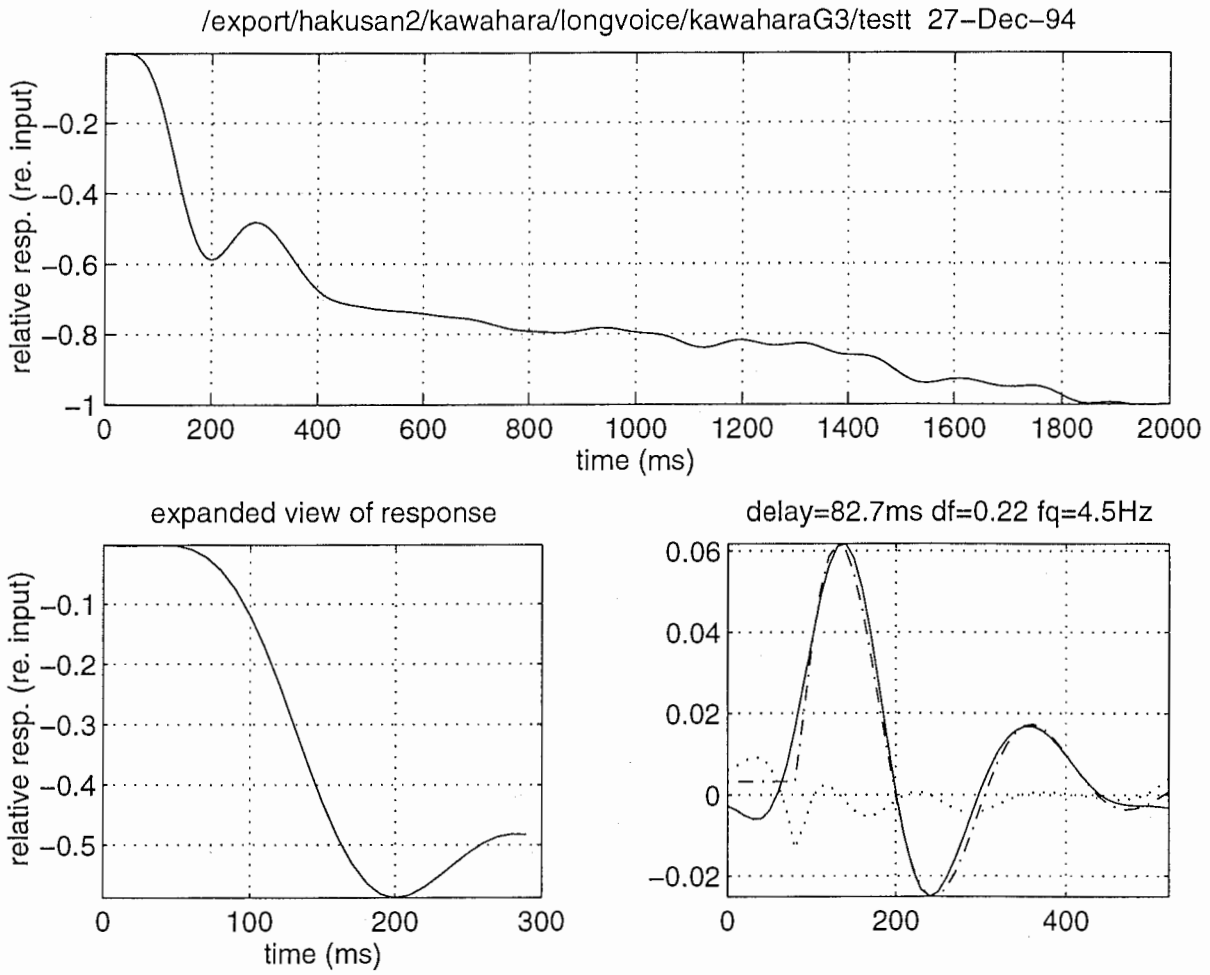

Figure 20: An example of an estimated step response from TAF experiments. The bottom right figure shows the best fit impulse response of the 2nd order system. The best parameteres are shown in the figure.

analysis. However, the direct application of an AR model to the estimated slow residual signal will fail to give a reasonable initial estimate. The reason is in the pole location and the periodic nature of the original signal.

The expected pole location caused by a slow auditory process is around 0.5Hz to 1Hz. That is only 1% of the sampling frequency. This strong bias makes the estimation process of AR parameters very sensitive to error.

The second difficulty is periodicity. This filtered signal is a periodic signal, because it is the result of periodic averaging and periodic convolution. The computation of auto correlation using DFT produces cyclic auto correlation. It is strongly biased when the pole frequency is very low and the damping ratio is small; this is the current case. Therefore, the typical method of computing the auto correlation has difficulty meeting the required accuracy.

The covariance method of auto correlation is used here, because it will give the best fit even with a small amount of data[12].

A brief description of the covariance method of AR parameter estimation is given below. Let x_t be an N element vector of observation values starting from time t, and v_t be the observation noise vector. Then, the AR model with p regression coefficients is represented as follows.

$$x_t = -H\alpha + v_t \tag{58}$$

where

$$H = [x_{t-1}, x_{t-2}, ..., x_{t-p}] \tag{59}$$

Then, the least squares estimate of the coefficients is given by the following calculation. The expected total variance of the parameter $V(\hat{\alpha})$ is also given.

$$\hat{\alpha} = -(H^T H)^{-1} H^T x_t \tag{60}$$

$$L_{min} = \|x_t + H\hat{\alpha}\|^2 \tag{61}$$

$$V(\hat{\alpha}) = \frac{L_{min}}{N-p} \mathrm{tr}\left[(H^T H)^{-1}\right] \tag{62}$$

Response parameters are calculated using the following formula from the estimated AR coefficients ($\{\alpha_k\}_{k=0}^2$, $\alpha_0 = 1$).

$$\omega_n = \sqrt{a^2 + b^2} \tag{63}$$

$$\zeta = \frac{a}{\omega_n}$$

where

$$a = -\log\left(\frac{\alpha_2}{2}\right)$$

$$b = \begin{cases} \cos^{-1}\left(\frac{-\alpha_1}{2e^{-a}}\right) & \frac{-\alpha_1}{2e^{-a}} < 1 \\ \cosh^{-1}\left(\frac{-\alpha_1}{2e^{-a}}\right) & \frac{-\alpha_1}{2e^{-a}} > 1 \end{cases}$$

Figure 21 shows an example of this decomposition. The figure illustrates that the decomposition is successful. It may indicate that the response to perturbations consists of two components, which work in parallel. The fact that the DAF condition with a 500ms delay causes unstable fundamental frequencies indicates that the slow response is not an artifact or error in the estimation process. Considering the relation between pole frequencies and the estimated transfer function as well as the above evidence of slow response, it may now be safe to say that we have identified two auditory feedback paths to control the fundamental frequency in voicing.

There still remain many points to be clarified. One is the confidence interval for these estimation processes. It is also necessary to derive AIC for these procedures to check the validity of our model. There is no proof at present that the estimation process can converge. In fact, there were a few exceptional cases in which the process produced strange results. This is the reason why we need a measure for the reliability of estimates.

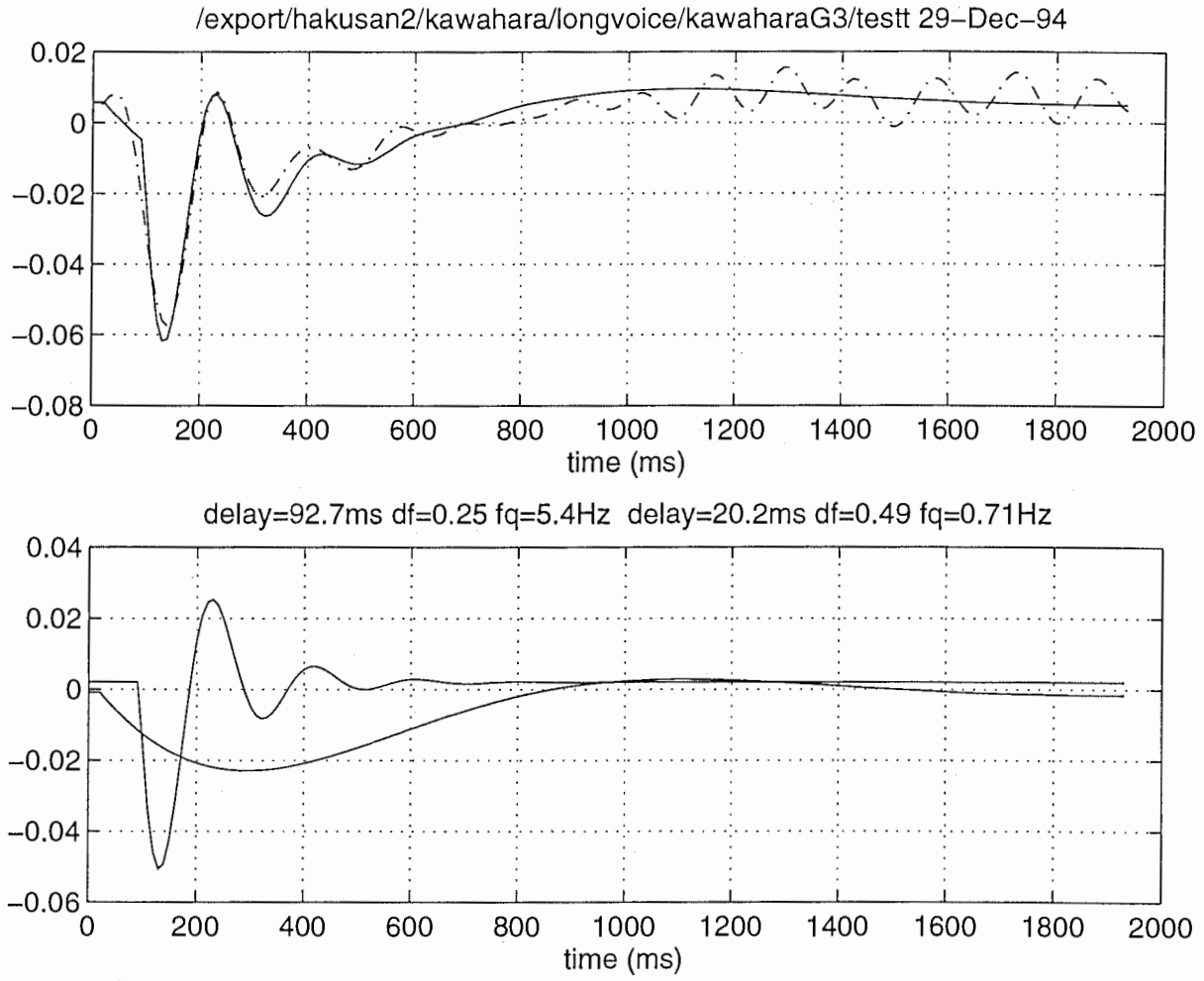

Figure 21: An example of response decomposition into fast and slow components.

5.5 Decomposition of the response for read sentences

The data obtained under TAF for read speech was analyzed by the proposed method. This analysis could be characterized as tentative, because the natural variation caused by the prosodic component was still comparable to the level of the expected response. Further data acquisition is necessary to get more reliable results. Figure 22 shows the decomposition result. An integrated display for the data can be found in the last part of Appendix B.

At this time, the slow component extracted by the decomposition procedure seems to represent the prosodic effect, because the characteristic frequency of the slow component corresponds to the moraic timing in the read speech and the component does not show decay with time.

The fast component extracted seems very similar to the fast response found in sustained vowel phonation. This may suggest that similar responses, or at least a fast response

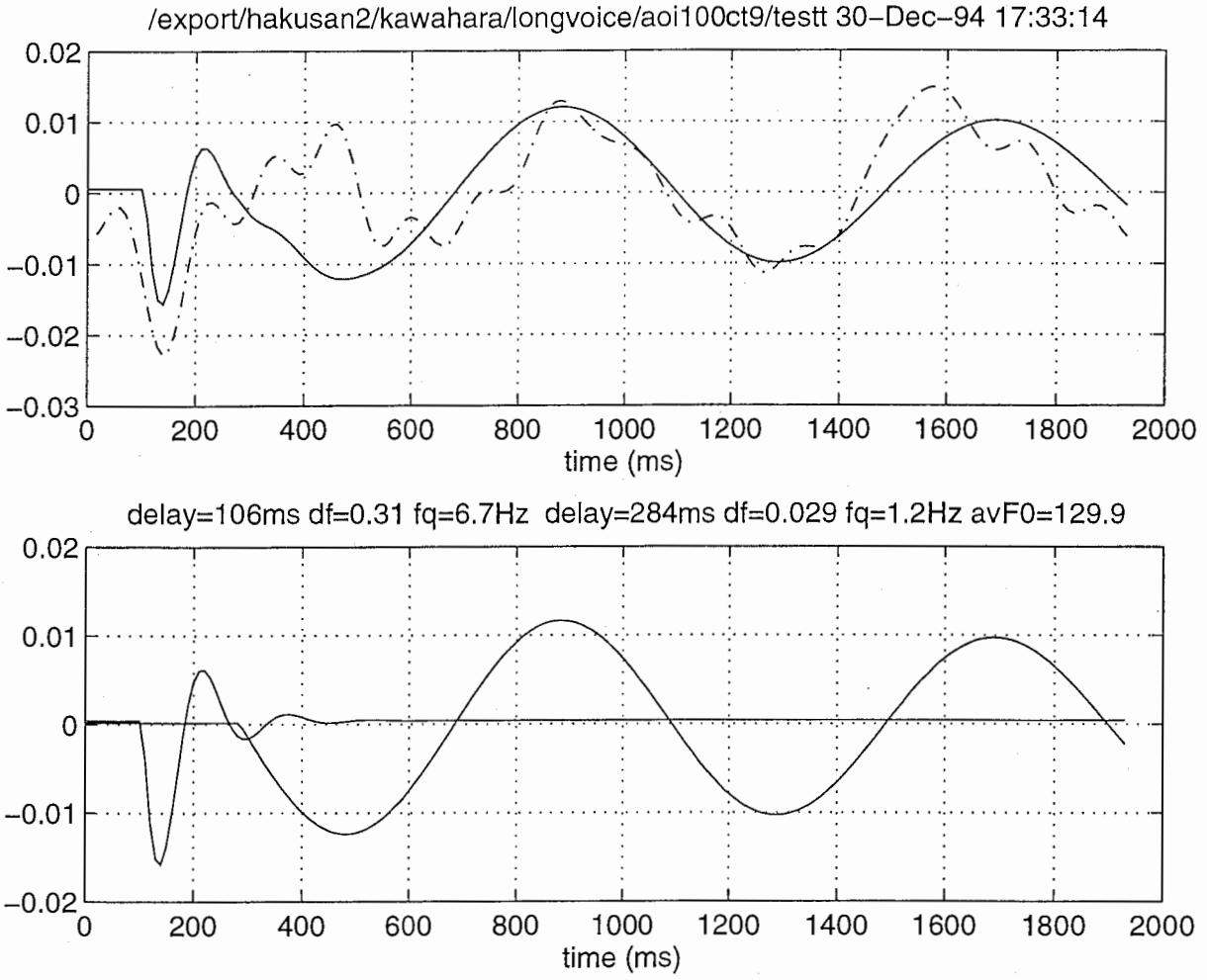

Figure 22: An example of response decomposition into fast and slow components.

operate *in parallel* with higher control like prosody.

6 Discussion

First of all, let us summarize the procedures and findings obtained in this series of re-analyses and the just introduced new analysis.

The data was analyzed in a uniform manner, i.e.,

(1) Fundamental frequency trajectories

(2) Periodic average representations of fundamental frequencies of fed-back and produced speech (phonation).

(3) The coherency of each variation frequency component.

(4) The loop transfer function of fed-back-to-produced interactions with confidence intervals.

Representations from (2), (3) and (4) were averaged over separate measurements of the same conditions and illustrated. Mathematical descriptions and calculation procedures of all statistical values were given in detail. This made it possible to estimate impulse and step responses to auditory stimulations.

The new findings using this set of procedures include (1) a strong but slow response around a 0.5Hz region and (2) the possible existence of the same response for natural speech. The first finding was demonstrated by the fact that introducing about a 500ms delay in the artificial auditory feedback path makes the pitch contour very unstable. A new analysis method introduced in analyzing the final part of the result suggested that the observed response can be modeled by two major auditory components. The first is (1) a fast and wide-band compensatory response which may be mediated by the cerebellum, and (2) a slow and narrow-band compensatory response which may be mediated by the cognitive cortex. The dynamic characteristics of pitch perception and laryngeal control are overlaid to these responses. One important thing about this is that they are not tandem. In fact, they seem to be working in parallel. This may be evidence that skilled tasks like speech production actually implement the so-called subsumption architecture [4, 5].

This type of dual level control system resembles motor control models and a computational model of vision proposed by Kawato et.al. [28, 29]. It is very likely that the same control principles are also applied to other tasks like prosodic control. These findings may provide a clue in helping to establish a computational theory of prosody control in Marr's terminology [32]. Or they may be more complicated. In experiments with Hirayama, there was one subject who showed a stable pole even with different DAF delay conditions. This may suggest the existence of a non-auditory feedback mechanism. If this were a neurally mediated response, it may suggest a triple level control model. This possibility is not very high. There may be a chance that more careful analyses of pole allocation and simulation of pole allocation based on the estimated transfer functions will resolve this puzzle in Hirayama's data. But this line of investigation has not yet been done.

Further experiments and advanced analyses of data based on the ARMA (Auto Regressive Moving Average) model or more sophisticated dual or triple level control models are important. One example is given in the most recent section. That model can be classified as an ARMA model with a special architecture. Also the MAICE principle may be applicable to the testing of various competing models, because even with ARMA and the other models it is possible to write the log likelihood functions when the Gaussian noise source can be assumed. This is also an important point. But again, this has not yet been done.

The frequency characteristics of loop transfer functions illustrate that the cut off frequency of dynamic system may be 8Hz or more. This contradicts with the result of wave form decomposition, where the mechanical responses seem to have resonance around 5Hz. It is therefore necessary to re-design the PN signal for TAF experiments to make measurements possible around the 7Hz to 14Hz region for more reliability, in other words, for higher coherency.

There are other interesting findings. For instance, some frequency regions exist in which the power spectrum levels are high but the coherency is low. This indicates that there are special noise sources or chaotic behaviors in fundamental frequency deviations.

The conclusions in our previous ICSLP paper should be revised based on these new findings. It is also necessary to introduce a better model to explain the behaviors of vocal fold vibrations. The first step is to fit the two-component model to the observed data.

Other interesting experiments to be conducted include so-called selective feedback paradigm combined with TAF, measurement of the response function of Lombard effects and measurement of the response function to formant perturbations. For the first experiment, there should be some difference between the feedback of higher harmonics and that of lower harmonics. For the last one, it is interesting to compare the results with Perkell et.al. [35, 42].

This series of experiments can shed light on pitch perception theories by providing an objective measure for evaluating competing theories. The estimated latencies provide bounds for the contribution of temporal information and frequency information.

7 Conclusion

All data gathered during the period spanning 1993.1 to 1994.12 under TAF conditions were analyzed using the same procedures. These analyses revealed that there are consistent features in responses to pseudo random sequence perturbations. There are two types of dominant responses to stimulations. One response is fast ($\approx$ 150ms) and the other is slow ($\approx$ 400ms). These are compensatory responses. It was found that they work together (in parallel) to speed up responses to perturbations. Identical responses were found to exist under natural sustained phonation conditions and also while reading a text. It may be reasonable to believe therefore that responses also operate in spontaneous speech in everyday life.

Acknowledgement

We appreciate the participation of our intern students in this project. They are Mr. Iwatani, Mr. Urakami, Mr. Iwazume and Mr. Hirayama. Mr. Iwatani replicated classical DAF experiments which led to the discovery of some TAF effects. His skillful programming and experimental preparations were indispensable for starting this project. Mr. Urakami acquired data for testing hemispheric dominance. Mr. Iwazume prepared MIDI routines to make the experimental design flexible. He also gathered data on sound source dependencies. Mr. Hirayama improved the emplyed data acquisition paradigm. This enabled efficient and reliable data acquisition, which eventually led to the proof of our pitch control model.

The authors also appreciate Dr. Honda, Dr. Kusakawa and Mr. Hirai for acquiring laryngeal data as well as their discussions. Ms. Williams of Ohio State University was very helpful with her discussions and criticism of our methods. Professors Sapir and Larson of Northwestern University provided many important papers and personal communications. Professor Cook of Kansai University offered new directions in our thinking concerning the new experiments on hemispheric intetactions. Dr. Ko'ichi Mori suggested a method of improving the measurement accuracy by the synchronous accumulation of data which

enabled us to detect TAF effects under reading situations. Professor Kasuya made an important criticism on our method at the 1994 fall meeting of ASJ. It led to the new decomposition method of TAF responses. It is important to mention that coutinuous push and support by our president Dr. Tohkura were indispensable to make this project possible. Finally, we appreciate members of our department and friends for doing research on auditory perception.

References

[1] H. Akaike: "Information Theory and an Extention of the Maximum Likelihood Principle," 2nd Inter. Symp. on Information Theory (Petrov, B. N. and Csaki, F. eds.), Akademiai Kiado, Budapest pp.267-281, (1973).

[2] H. Akaike : "What is 'AIC', an infotmation criteria," Mathematical Science, 153, pp.5-11, (1976). [in Japanese].

[3] J. Bendat and A. Piersol: "Random data – Analysis and measurement procedures (2nd Ed.)," John Wiley & Sons, (1986).

[4] R. A. Brooks: "A Robust Layered Control System for a Mobile Robot," IEEE J. of Robotics and Automation, RA-2, pp.14-23, (1986).

[5] R. A. Brooks: "Intelligence Without Reason," Proceedings of the 1991 International Joint Conference on Artificial Intelligence, pp.569-595, (1991).

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

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

[8] K. Hirayama and H. Kawahara: "Effects of Auditory Feedback Conditions on Fundamental Frequency Fluctuations," Technical Report of IEICE, SP94–48, (October 1994). [In Japanese]

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

[10] F. Itakura and S. Saito: "Digital filtering technique for speech analysis and synthesis, " 7th Int. Congr. Acoust., Budapest, 25 C1, (1971).

[11] S. Iwatani and H. Kawahara: "Preliminary Investigations on Transformed Auditory Feedback – Pitch Variations Induced by Time Variant Pitch Conversion – ," ATR Technical Report, TR-H-009, HIP-ATR, (March 1993). [in Japanese]

[12] H. Kawahara, K. Tochinai and K. Nagata : "On the Linear Predictive Analysis using a Small Analysis Segment and its Error Evaluation," J. Acoust. Soc. Jpn., 33, 9, pp.470-479, (1977). [in Japanese]

[13] H. Kawahara: "On Interactions Between Speech Production and Perception using Transformed Auditory Feedback," Technical Report of Acoust. Soc. Jpn., H–93–24, (May 1993). [In Japanese]

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

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

[16] H. Kawahara: "Impact of Transformed Auditory Feedback in Hearing Research," ITE Technical Report, 17, 53, VAI93–29, pp.1-6, (september 1993). [In Japanese]

[17] H. Kawahara: "Transformed Auditory Feedback: Effects of Fundamental Frequency Perturbation," Proc. 1993 Fall Meeting of Acoust Soc. Am., 5aSP28, pp.1883-1884, (October 1993).

[18] H. Kawahara: "Transformed Auditory Feedback: Effects of Fundamental Frequency Perturbation," ATR Technical Report, TR-H-040, HIP-ATR, (December 1993).

[19] H. Kawahara: "Implementation of Auditory Models," J. of IEIECJ, 76, 11, pp.1197-1202, (1993). [in Japanese]

[20] H. Kawahara: "Interactions between Speech Production and Perception under Transformed Auditory Feedback," Technical Report of IEICE, SP93–70, (January 1994). [In Japanese]

[21] H. Kawahara: "On Interactions between Speech Production and Perception using Transformed Auditory Feedback in Fundamental Frequency Control," Proc. of Spring Annual Meeting of the Acoust. Soc. Jpn, 1-8-20, (1994). [in Japanese]

[22] H. Kawahara and M. Iwazume: "Source characteristics effects on fundamental frequency responses under transformed auditory feedback," Technical Report of Acoust. Soc. Jpn., H–94–28, (May 1994). [In Japanese]

[23] H. Kawahara: "Interactions between speech Production and perception under auditory feedback perturbations on fundamental frequencies," J. Acoust. Soc. Jpn. (E), 15, 3, pp.201-202, (1994).

[24] H. Kawahara: "A Fundamental Frequency Control Model which Consists of Auditory Mediated Regulation," Technical Report of IEICE, SP94–34, (July 1994). [In Japanese]

[25] H. Kawahara: "Effects of Natural Auditory Feedback on Fundamental Frequency Control," Proc. of ICSLP'94, Yokohama, Japan, S24-2.4, pp.1399-1402, (1994).

[26] H. Kawahara: "Contributions of Auditory Feedback on Fundamental Frequency Fluctuations in Sustained Vowel Production," Proc. of Fall Annual Meeting of the Acoust. Soc. Jpn, 3-7-1, (1994). [in Japanese]

[27] H. Kawahara: "Toward the Computational Theory of Audition," Technical Report of Acoust. Soc. Jpn., H–94–63, (November 1994). [In Japanese]

[28] M. Kawato, K. Furukawa and R. Suzuki: "A Hierarchical Neural-network Model for Control and Learning of Voluntary Movement," Biological Cybernetics, 57, pp.169-185, (1987).

[29] M. Kawato, H. Hayakawa and T. Inui: "A Forward-inverse Optics Model of Reciprocal Connections between Visual Cortical Areas," Network, 4, pp.415-422, (1993).

[30] C. R. Larson et.al. :"A Proposal for the Study of Voice Fo Control using the Pitch Shifting Technique", unpublished manuscript, 1994.

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

[32] D. Marr: "Vision – A Computational Investigation into the Human Representation and Processing of Visual Information," Freeman, New York, (1982).

[33] H. Miyakawa: "Probabilistic Systems and Estimation of Dynamic Characteristics," Colona Publishing Co., Tokyo, (1978). [in Japanese]

[34] J. No and V. Kumar: "A New Family of Binary Pseudorandom Sequences Having Optimal Periodic Correlation Properties and Large Linear Span," IEEE Trans. Information Theory, 35, 2, pp.371-379, (1989).

[35] J. Perkell, H. Lane, M. Svirsky and J. Webster: "Speech of cochlear implant patients: A longitudinal study of vowel production," J. Acoust. Soc. Am., 91, pp.2961-2978, (1992).

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

[37] Y. Sakamoto, M. Ishiguro and G. Kitagawa: "Information Theory and Statistics," Kyoritsu publishing Co., Tokyo, (1983). [in Japanese].

[38] 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).

[39] 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).

[40] 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).

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

[42] M. Svirsky, H. Lane, J. Perkell and J. Wozniak: "Effects of short-term auditory deprivation on speech production in adult cochlear implant users," J. Acoust. Soc. Am., 92, pp.1284-1300, (1992).

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

[44] 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).

[45] H. Urakami and H. Kawahara: "Hemispheric Effects under Transformed Auditory Feedback," ATR Technical Report, TR-H-026, HIP-ATR, (September 1993). [in Japanese]

[46] N. Zierler and J. Brillhart: "On primitive trinomials (mod 2) II," Information and Control, 14, pp.556-569, (1969).

A M-sequence, PN signal and perturbation

The M sequence is a binary sequence generated by the following recursive equation.

$$\begin{aligned} X_n \quad &= \quad X_{n-p(1)} \oplus ... X_{n-p(m)} \qquad (64) \\ &\text{Where} \\ & \qquad p(1) < ... < p(i) < ... < p(m) \\ & \qquad p(m) = K; \end{aligned}$$

Here $\oplus$ denotes on 'exclusive OR' operation. The tap information $p(i)$ is given based on Garoi's theory, and listed in Table 8. This recursion generates a binary sequence which has the period of $2^K - 1$ elements.

The PN signal is derived from the M-sequence simply by introducing a bias term.

$$\begin{aligned} s(n) \quad &= \quad X_n \left(1 \pm 2^{-\frac{K}{2}}\right) - 1 \qquad (65) \\ &= \quad \begin{cases} 1 \pm 2^{1-\frac{K}{2}} & (\text{ if } \ X_n = 1) \\ -1 & (\text{ if } \ X_n = 0) \end{cases} \end{aligned}$$

Then, the PN signal is normalized to satisfy the following normalization condition. *The normalization constant should be given in a closed form. This part will be revised.*

$$\frac{1}{T_p} \sum_{k=1}^{T_p} \tilde{s}(k)\tilde{s}(k-l) = \begin{cases} 1 & (\text{if } l = 0) \\ 0 & \text{otherwise} \end{cases} \qquad (66)$$

order (K)	tap position $(K - p(1) + 1)$
2	2
3	3
4	4
5	4
6	6
7	7
8	does not exist
9	6
10	8
11	10
12	does not exist
13	does not exist
14	does not exist
15	15
16	does not exist
17	15
18	12
19	does not exist
20	18
21	20
22	22
23	19

Table 8: The list of tap positions to produce an M-sequence with 2-taps.

Over sampling is simply a process of inserting zeros between adjacent samples. Let M be the rate of over sampling. Then, the over sampled PN signal $s_M(n)$ is represented as follows.

$$s_M(n) = \begin{cases} s\left(\lfloor \frac{n}{M} \rfloor\right) & (\text{ if } \frac{n}{M} = \lfloor \frac{n}{M} \rfloor) \\ 0 & \text{otherwise} \end{cases} \tag{67}$$

In general, this over sampling introduces spurious repetition in higher frequency regions. This causes artifacts for general signals. However, the PN signal produces a flat spectrum without any problems, because of its orthonormality. The final step is to smooth out the signal for it to be used to modulate the fundamental frequency of speech sounds.

There are several conditions for designing such a signal.

(1) The signal has to produce a better signal to noise ratio when the peak to peak value is limited.

(2) The signal shall not exceed the limit of the slope or velocity.

(3) The signal shall not have dominant side lobes that exceed the predetermined limit.

(4) The signal has to be band limited under the specific frequency.

(5) The length of the impulse response associated with the smoothing has to be as small

as possible.
(6) The impulse response associated with the smoothing has to be symmetric in time, in other words, a linear phase.
(7) The step response associated with the smoothing shall not have overshoot or undershoot.

A non-linear optimization procedure may be applicable to the design of such a kernal for interpolation. The up sampling function provided in MATLAB is not optimum in this sense. Tentatively, we use the Blackmann window function as a close approximation of the optimal kernel.

B Integrated display of all data

Integrated displays of all the data gathered during this period are shown in a separate volume as the supplement to this technical report. They are listed in the following order.

1. Experiments from 1993.1 through 1993.3
   Figures B-1 through B-25.

2. Experiments of EMG
   Figures B-26 through B-41 for voice data.
   Figures B-42 through B-53 for EMG data.

3. Experiments of hemispheric dominance
   Figures B-54 through B-95

4. Experiments of source characteristics
   Figures B-96 through B-123

5. Experiments on feedback conditions and the model
   Figures B-124 through B-147

6. Decomposition of estimated impulse responses
   Figures B-148 through B-180

7. Power spectrum plots
   Figures B-181 through B-190

Another type of analysis like MAICE also took place. Phonation without the artificial feedback conditions were analyzed using an AR model based on AIC. These results will be presented in the other report.