EFFECTS OF AUDITORY FEEDBACK ON F_0 TRAJECTORY GENERATION

Hideki Kawahara, Hiroko Kato † and J. C. Williams‡

ATR Human Information Processing Research Laboratories, Kyoto Japan
†Department of Engineering, Hiroshima University, Hiroshima Japan
‡Ohio State Univeristy, OH, USA

ABSTRACT

In this paper, a method is proposed to evaluate contributions of auditory feedback to speech F_0 trajectory generation. This method is based on data obtained in a series of new auditory feedback experiments (TAF: transformed auditory feedback) in which quantitative measurements were taken of interactions between speech perception and production under natural speech conditions. Experimental results revealed that the effects on power spectra of F_0 trajectories vary among subjects and that the maximum magnitude exceeds 10 dB.

1 INTRODUCTION

It is generally believed that humans monitor their own voice while speaking, in order to adapt their speaking style to changing environments. If this is indeed the case, speaking should be affected by auditory stimulation that alters natural feedback conditions.

The basic concept of our experimental method, TAF (Transformed Auditory Feedback), is based on this idea. In a TAF experiment, natural auditory feedback paths are each replaced by an artificial feedback path to introduce a systematic transformation of speech parameters into auditory feedback paths and to analyze effects on produced speech parameters. Using this method it has been demonstrated that there is a compensatory response to pitch perturbations [3, 4].

The main purpose of this paper is to provide a quantitative evaluation of auditory effects on F_0 trajectories. It also intends to underscore the relevance of prediction based on parameters extracted using the TAF procedure.

2 TRANSFORMED AUDITORY FEEDBACK

A brief description of the TAF procedure and a summary of our previous experimental results are given in this section.

Figure 1 shows a schematic diagram of an implementation of the TAF concept for pitch perception experiments. A pseudo random noise (a PN signal) based on the M-sequence was used as the F_0 perturbation signal. The amount of modulation was 3% of the average F_0 (peak--to--peak). Subjects were asked to sustain a Japanese vowel /a/ for about 10 seconds several times with a comfortable pitch. The total effective length of utterances ranged from 30 seconds to 100 seconds. (Sound examples: produced speech[SOUND A288S01.WAV] and fed back speech[SOUND A288S02.WAV].)

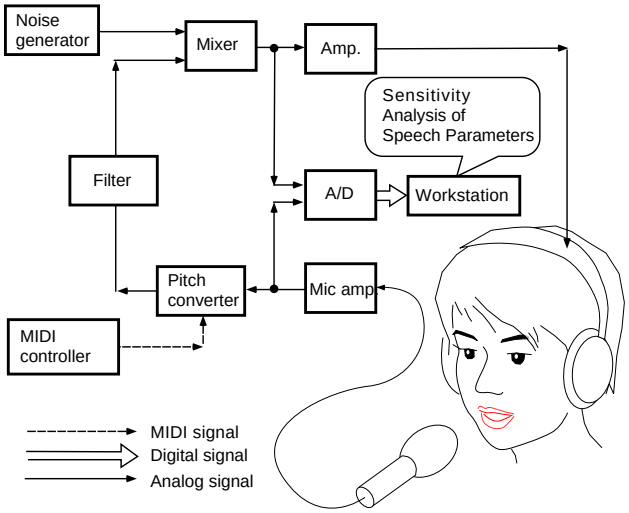

Figure 1: Schematic diagram of an implementation of the TAF concept for F_0 perturbation experiments.

An impulse response representation of the auditory effects on speech production in pitch control was estimated by taking the inverse Fourier transformation of the open-loop transfer function from the feedback F_0 to the produced F_0. A synchronized averaging using the periodic nature of the PN signal was employed to improve the SN ratio in the measurements. The detailed procedures are given in literature[4].

2.1 Parameter Estimation (Example)

Figure 2 shows a typical F_0 impulse response and its decomposition. The broken line in the upper figure represents the estimated open-loop impulse response using a PN signal as the perturbation. The solid line represents the re-synthesized composite response based on two second-order responses which were the result of decomposition. The component responses are illustrated in the lower plot of the same figure.

One important observation in this example is that the relatively fast component consists of a delay of 122 ms. In other words, the fast second-order response starts 122 ms after the application of the stimulus. This response time (a fast delay) may reflect the amount of time needed for information processing to compute neural commands from the fed back voice. In this example, the natural frequency for the fast response was 6.4 Hz, and its damping factor was 0.3. The natural frequency for the slow response was

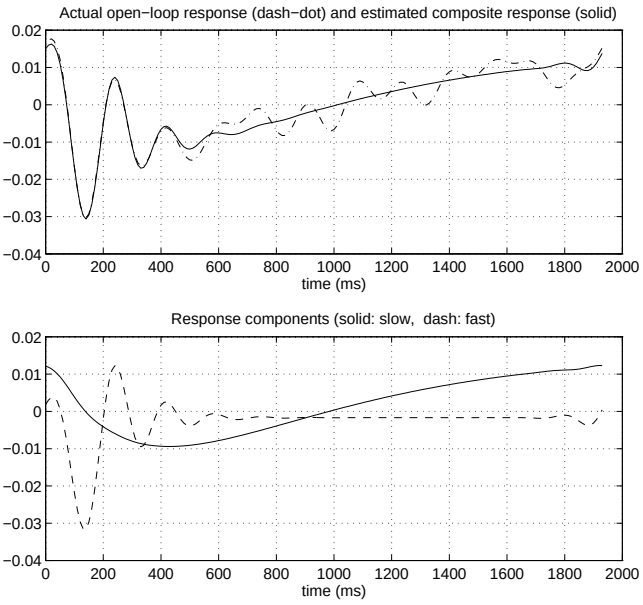

Figure 2: An open-loop impulse response to F_0 perturbation of feed-back speech.

0.48 Hz, and its damping factor was 1.

2.2 Distribution of Parameters

A series of TAF experiments using sustained vowel phonations were conducted. Typical natural frequencies of slower component responses ranged from 0.5 Hz to 1 Hz, while those of faster component responses ranged from 4 Hz to 7 Hz. The open-loop gain in the lower frequency region (under 2 Hz) typically ranged from -10 dB to 0 dB, while that in the higher region (over 2 Hz) ranged from -20 dB to -8 dB. Initial delays were mainly in the 90 ms to 150 ms range, depending on the subject, his/her gender and F_0 [4].

3 F_0 TRAJECTORY GENERATION

Voice F_0 is generally difficult to maintain constant without feedback control, because it depends on various factors that change during voicing. Even though it is unrealistic to keep F_0 constant in natural speech situations, some control mechanisms are still indispensable for producing desired F_0 trajectories to encode prosodic information. A general framework for voluntary movements[5] may also be applicable to this task. Transfer functions found in TAF experiments represent some aspect of this voluntary control mechanism.

Figure 3 shows a functional model of F_0 control which consists of an auditory system as an indispensable component. R and C in the figure represent cerebellum and cerebral functions, respectively. A and P stand for the auditory system and the production system. The broken lines represent a hypothetical correspondence to component responses (F: fast and S: slow responses) found in TAF experiments.

Note that if the transfer function extracted in TAF experiments could also operate under normal speech conditions, the auditory effects on the generated F_0 trajectories could be predicted using this model.

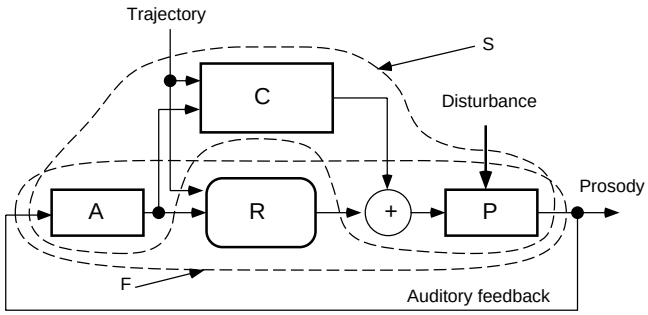

Figure 3: A functional model of F_0 control which consists of auditory system as a component. The figure overlays a hypothetical relationship to component responses found by TAF experiments.

3.1 Constant F_0

TAF experiments conducted so far have mainly used conditions with a sustained vowel /a/ in a constant F_0. That being the case, the parameters extracted in TAF experiments are directly applicable to the situation presented here.

In this condition, the target F_0 trajectory in **Figure 3** is constant. The role of the auditory system is to detect F_0 deviations from the constant target value and to compensate the detected deviations.

Effects of auditory feedback suppression Let the transfer function estimated by TAF experiment be $G(z)$. It is convenient to divide noise sources according to their origin. Let n represent a 'neural' noise source and let v represent a 'vocal fold related' noise source. Then, the observed $F0$ time series f_n under the 'constant F_0' condition is represented using an unobservable characteristic P, which represents the transfer function of the production system.

$$f_n = \frac{P(f_0 + n) + v}{1 - G} \tag{1}$$

Elimination of the auditory information yields the time series f_m under the masked condition as follows.

$$f_m = P(f_0 + n) + v \tag{2}$$

Then, the ratio of the normal power spectrum to the masked power spectrum F_n/F_m is reduced to $1/|1-G|^2$, which can be calculated solely using the TAF results.

Effects of delay in auditory feedback Introducing a delay also modifies the power spectrum of the F_0 trajectory. Let τ represent the inserted delay and DAF (delayed auditory feedback) represent this condition. Then, the ratio of the DAF power spectrum to the masked power spectrum F_d/F_m yields the following.

$$\frac{F_d(\omega)}{F_m(\omega)} = \frac{1}{|1 - Ge^{-j\tau\omega}|^2} \tag{3}$$

where ω represents the angular frequency.

3.2 Varying F_0

Strictly speaking, the TAF results are applicable only when F_0 deviates by fluctuations in the speech production system or by some artificial shift operations [6]. The transfer function $G(z)$ is a function of F_0 and many other parameters like registration of the voice. Models given in this section are first-order approximations of this complex behavior.

F_0 variation by an external source Any small F_0 shift introduced in the artificial feedback loop can be represented by an equivalent perturbation to the auditory input. Let p represent this equivalent perturbation. Then, the observed F_0 time series under this condition, f_p, is represented by $f_p = pG/(1-G)$.

TAF results do not provide $G(0)$, since the measurement is based on deviations from the mean F_0. However, based on the findings of Ellman [2] and Larson [6], it is reasonable to assume that $G(0)/(1-G(0))$ is close to -1. (Demonstration: Rondom F_0 shifts make this song out of tune. [SOUND A288S03.WAV] But, the song the singer is hearing is normal. [SOUND A288S04.WAV])

F_0 variation by an internal source It is not possible to predict the auditory effects directly under normal verbal communication and singing, because internal representations of F_0 trajectory information are not known for these varying target situations.

If it is a reasonable hypothesis to assume that the system represented by the results of TAF experiments for the constant F_0 conditions also operates under these varying target situations, the F_0 time series of this situation f_s will be represented as $f_m/(1-G)$. In this case, f_m represents the F_0 time series for the masked condition. A TAF experiment using a repeated sentence with all voiced sounds (/*aioi no oi wa yama no ue no ie ni iru*/ in Japanese) suggested that the system responsible for the constant F_0 TAF results also functions under varying F_0 situations.

4 EXPERIMENTS

A series of experiments were conducted for each F_0 trajectory generation condition. Effects of natural auditory feedback on trajectory generation were simulated using these open-loop characteristics and compared with the experimental results.

4.1 Constant F_0

Power spectral ratios between F0 trajectories under natural feedback conditions and those under conditions without auditory feedback were tested. The case of no auditory feedback was implemented by cutting the microphone input and adding a loud (85 to 90 dB(A)) mixture of pink noise and a sinusoidal tone through headphones. The sinusoidal tone frequency was set equal to the target F_0. Subjects were instructed to sustain a Japanese vowel /a/ at the target F_0. The masking noise was started prior to the first phonation and terminated after the last utterances. The total time for one session was 120 seconds. Two reference conditions were used. One was the natural condition, i.e., without headphones. The other was the control condition where headphones were used and there were no F_0 manipulations.

Figure 4 shows an exemplar comparison between simulation results and the real data. As shown in this figure, typically, natural auditory feedback amplifies F_0 fluctuations around 4 Hz and suppresses them around 6 Hz. The actual frequencies and magnitudes of amplification and suppression differ from subject to subject. The maximum estimated peak--to--peak gain effect exceeds 10 dB. The figure also indicates that by using headphones some bias is introduced in the auditory effects, thereby increasing the peak frequency.

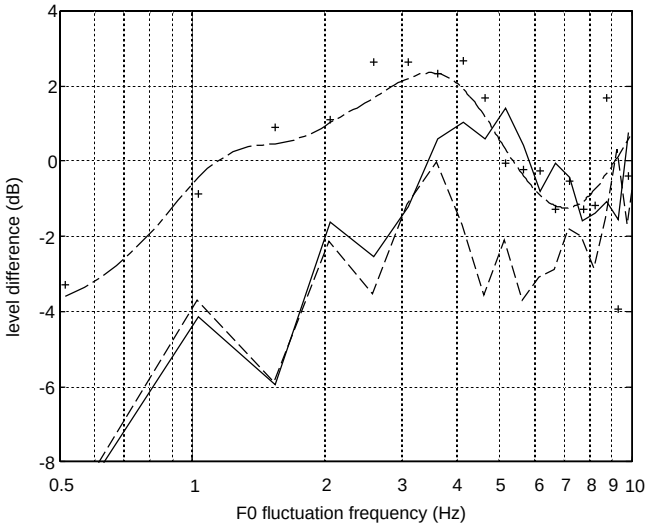

Figure 4: Simulated effects of auditory feedback and the experimental results. The plot shows F_n/F_m. The smooth dash-dot line represents the simulation results. The solid line shows results for the natural condition without headphones. The broken line shows results for the control condition with headphones.

Simulation results with delays predicted interesting effects of DAF on F_0 trajectories in a lower fluctuation frequency region. For example, it has been suggested that DAF with a longer delay (200 ms or more) induces strong amplification or instability of F_0 fluctuations around 0.5 Hz, which was an observation mentioned in our previous report [3]. This behavior was replicated by the simulation with delays.[IMAGE A288G01.GIF]

4.2 Varying F_0

The response to step shifts in fed back F_0 was simulated and compared with the real data. A 50-cent (a half semitone) shift (peak--to--peak pitch deviation) was introduced into the artificial feedback paths with random inter step intervals longer than 1 second.

Figure 5 shows the results. The solid line in the upper plot represents a simulated step response based on the estimated TAF response. The introduced step was a positive F_0 shift. The broken line in the same plot represents a result based on composite response approximation. The actual response is an averaged response, normalized by the introduced steps. This behavior replicated the normal response for small step shifts [6]. The simulation results also replicated the actual response.

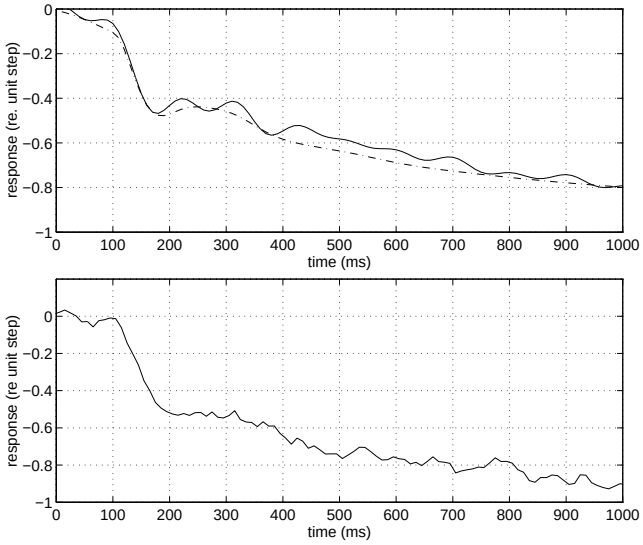

Figure 5: A simulated response (upper plot) and the actual response (lower plot) to an F_0 shift. The actual response is an average of 40 repetitions.

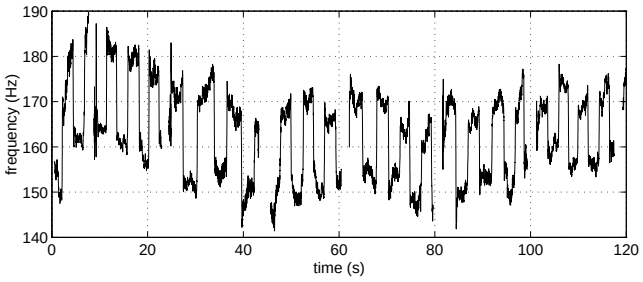

Figure 6: Voluntary F_0 alternation without auditory feedback.

Voluntary F_0 shifts Finally, F_0 trajectory differences under the masked condition and the natural condition were tested for voluntary F_0 alternation. The size of the alternation was set to two semitones. **Figure 6** shows the observed behavior under the masked condition: the normal trajectory is modified by the amount predicted by $1 - G$. The slow drift in F_0 corresponds to a high gain in a lower frequency region (namely 1 Hz or less). **Figure 7** shows the averaged power spectral difference between two conditions around F_0 transitions. The smooth peak around 4 Hz corresponds to the predicted auditory feedback effects. The jaggy shape in the 7 Hz to 10 Hz region in the plot is due to a low signal to noise ratio caused by the observation noise in the F_0 estimation.

5 DISCUSSION

These results support a hypothesis that the auditory system plays an indispensable role in F_0 control in normal speech production processes, even though it is a highly skilled behavior. This suggests that higher control processes co-exist with lower automatic

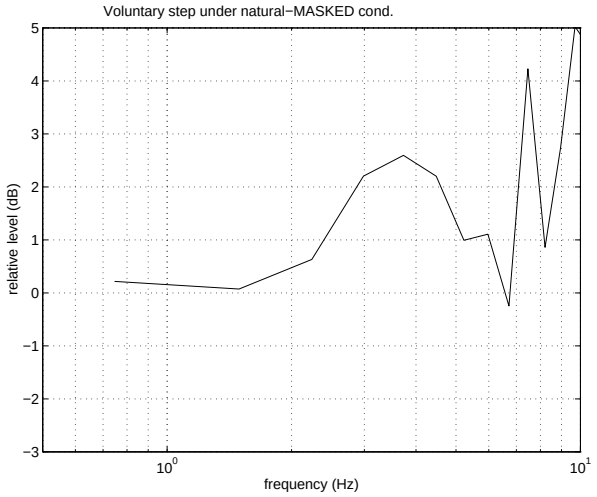

Figure 7: Power spectral difference between the normal condition and the masked condition around F_0 transitions.

processes and do not subsume lower functions completely. It may also be safe to conclude that the parameters extracted by the TAF procedure represent the dynamics of the F_0 control quantitatively.

6 CONCLUSION

It has been demonstrated that auditory effects on speech production are not negligible and can be estimated using impulse responses measured by the TAF procedure. It is interesting to investigate the relationship between results obtained by the TAF procedure and dynamic models of prosodic control.

REFERENCES

1. K. Aikawa, M. Tsuzaki, H. Kawahara and Y. Tohkura. ‘‘Pitch Ringing Induced by Frequency-Modulated Tones’’, *J. Acoust. Soc. Am.*, 98(5), Pt.2, p.2926, 1995.
2. J. L. Elman. ‘‘Effects of frequency shifted feedback on the pitch of vocal productions’’, *Journal of the Acoustical Society of America*, 70:45--50, 1981.
3. H. Kawahara. ‘‘Effects of Natural Auditory Feedback on Fundamental Frequency Control’’, *ICSLP’94* , 24.2, 1994.
4. H. Kawahara and J. C. Williams. ‘‘Effects of Auditory Feedback on Voice Pitch’’, *The 9th Vocal Fold Physiology Symposium*, Sydney, 1995, (in Print).
5. M. Kawato and H. Gomi. ‘‘A Computational Model of Four Regions of the Cerebellum based on Feedback-error Learning’’, *Biological Cybernetics* , 68:95-103, 1992.
6. C. R. Larson, T. D. Carrell, J. E. Senner, T. A. Burnett, and L. L. Nichols. ‘‘A proposal for the study of voice f0 control using the pitch shifting technique’’, *in* O. Fujimura and M. Hirano, eds, *Vocal Fold Physiology -- Voice Quality Control* , Singular Publishing Group Inc., 321--331, 1995.