Hearing Research 324 (2015) 1–6

Contents lists available at ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

Research paper

Audio–vocal responses of vocal fundamental frequency and formant during sustained vowel vocalizations in different noises

Shao-Hsuan Lee [a], Tzu-Yu Hsiao [b], Guo-She Lee [a, c, *]

[a] Faculty of Medicine, School of Medicine, National Yang-Ming University, Taipei, Taiwan
[b] Department of Otolaryngology, National Taiwan University Hospital and College of Medicine, National Taiwan University, Taipei, Taiwan
[c] Taipei City Hospital, Ren-Ai Branch, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 25 August 2014
Received in revised form
24 February 2015
Accepted 26 February 2015
Available online 5 March 2015

ABSTRACT

Sustained vocalizations of vowels [a], [i], and syllable [mə] were collected in twenty normal-hearing individuals. On vocalizations, five conditions of different audio–vocal feedback were introduced separately to the speakers including no masking, wearing supra-aural headphones only, speech-noise masking, high-pass noise masking, and broad-band-noise masking. Power spectral analysis of vocal fundamental frequency (F0) was used to evaluate the modulations of F0 and linear-predictive-coding was used to acquire first two formants. The results showed that while the formant frequencies were not significantly shifted, low-frequency modulations (<3 Hz) of F0 significantly increased with reduced audio–vocal feedback across speech sounds and were significantly correlated with auditory awareness of speakers' own voices. For sustained speech production, the motor speech controls on F0 may depend on a feedback mechanism while articulation should rely more on a feedforward mechanism. Power spectral analysis of F0 might be applied to evaluate audio–vocal control for various hearing and neurological disorders in the future.

1. Introduction

Speech communication relies on sophisticated sensory-motor integration of both central and peripheral nervous systems. The model of Directions Into Velocities of Articulators (DIVA) is one of the theoretical models that helps to explain the audio–vocal feedback system in terms of neural network and cortical interactions (Guenther, 2006). For keeping a stable speech, DIVA model suggests that the feed-forward control for speech output is performed on the basis of learned motor commands, while the auditory-feedback modification of phonation is mainly induced by the mismatches between the actual auditory feedback signals and the auditory sensory expectations (Tourville et al., 2008). A number of studies have confirmed that auditory feedback is one of the most important sensory information contributing to the learning and stability of phonation and articulation in human speech, and there are interactions between speech production and auditory reception which tend to induce active and reflexive control of vocal-fold vibrations and speech articulation in response to auditory interference. Speakers are likely to show significant changes in vocal fundamental frequency (F0), formant transitions, vocal intensity, speech rate, and/or nasal resonance when auditory feedback of self-generated voice is delayed, pitch-shifted, noise-masked, or greatly attenuated. These observations bolster audio–vocal feedback loop as a key to maintain speech stability (Chen et al., 2007; Ferrand, 2006; Hain et al., 2001; Larson et al., 2007, 2001; Lee et al., 2007).

Even in sustaining an as-steady-as possible vowel, F0s are not constant throughout the entire phonation (Titze, 1991; Titze et al., 1993). Rhythmic fluctuations of F0 do exist and were deduced to originate from the modulations of auditory feedback, aerodynamics of vocal production, or inherent irregularities in the nature of laryngeal muscle contractions (Titze, 1991). Each cycle of vocal fold vibrations is not exactly the same in time. The rhythmic fluctuations of vocal fold vibrations are different in frequencies and are

Abbreviations: ANOVA, analysis of variance; DIVA, directions into velocities of articulators; BBN, broadband-noise masking; F1, first formant; F1SD%, standard deviation of first formant frequency in percentage; F2, second formant; F2SD%, standard deviation of second formant frequency in percentage; NO, no-masking hearing status; EO, wearing headphone only; SN, speech-noise masking; HPN, high-pass noise masking; LFP, low-frequency power; MFP, middle-frequency power; HFP, high-frequency power; LPC, linear predictive coding.

* Corresponding author. Department of Otolaryngology, National Yang-Ming University, No. 155, Sec. 2, Li-Nong Street, Bei-Tou District, Taipei City 112, Taiwan. Tel.: +886 2 28267000x6155; fax: +886 2 28202190.

E-mail addresses: satomilee0701@gmail.com (S.-H. Lee), tyhsiao@ntu.edu.tw (T.-Y. Hsiao), guosheli@gmail.com (G.-S. Lee).

http://dx.doi.org/10.1016/j.heares.2015.02.005
0378-5955/

generally classified as vocal wow (0–3 Hz), vocal vibrato (3–8 Hz), and vocal flutter (≥8 Hz). A vocal wow is a periodic variation of lower than 3 Hz underlying the vibrations of vocal folds. This essential instability cannot be totally suppressed even though the speaker has the experiences of voice or singing training. The low-frequency fluctuations imbedded in the signals of cycle-to-cycle vocal fold vibrations have been considered related with the audio–vocal interaction in our previous studies and tended to increase significantly while the speaker sustaining the vowel [a] under disturbed auditory input (Lee, 2012; Lee et al., 2004). It should be emphasized that it is the fluctuations of F0 below 3 Hz being analyzed rather than the vocal F0 itself. A faster pulsation of F0, usually between 3 Hz and 8 Hz, is known as vocal vibrato. A vibrato has been considered associated with active modulation of the laryngeal motor neuron pool (Hsiao et al., 1994) and the control of auditory system (Leydon et al., 2003). It can be deliberately produced, suppressed, or modified after training. The rhythms of faster than 8 Hz in F0 are another source of vocal fluctuations known as vocal flutters. The rapid oscillations in F0 might represent a natural oscillating of the glottal adductor–abductor control system during phonation (Aronson et al., 1992).

Our previous findings showed that the low-frequency rhythms in F0 significantly increased in the normal-hearing speakers with noise masking (Lee et al., 2004, 2007) and in the post-lingual and the pre-lingual hearing-impaired speakers (Lee, 2012; Lee et al., 2013). The findings provided evidence that the involuntary modulations of vocal-fold oscillations was associated with the auditory feedback responding to the mismatch between anticipated and actual auditory information from self-generated speech. However, the speech material and the type of noise had been limited to vowel [a] and speech noise, so it remains unclear whether other speech sounds and/or a different type of noise masking will also alter the subsequent audio–vocal feedback modulations of F0 and even speech articulation in the same way. Therefore, we included three speech sounds with different formant frequencies to clarify if there is a dependence of F0 feedback on formant energy. We also used noise masking of different frequency bands to explore the responses of F0, as well as formant frequencies, to the information loss of formant energy. The audibility of vocalization was also evaluated to investigate the relationship between F0 feedback and auditory attention system. All speakers were requested to produce the vowels and syllable in tone 1.

2. Methods

2.1. Participants

Twenty participants (10 males and 10 females), aged between 20 and 40 years, having no medical history of neurological deficits, speech-language disorders, current upper respiratory infection, or the experience of voice singing training were enrolled. All participants passed the hearing screening test which was defined as a pure-tone hearing threshold level of better than or equal to 25 dB HL at the frequencies of 250 Hz, 500 Hz, 1000 Hz, 2000 Hz, 4000 Hz, and 8000 Hz. The participants were all native Mandarin speakers. The research procedures were approved by the Institutional Review Board of National Yang Ming University (IRB-960014), and the informed consent was acquired from each participant.

2.2. Sampling of voice

Voice recordings were conducted in a sound-treated room in which background noise was lower than 40 dBA monitored by a sound-level meter. On the assumption of different audio–vocal feedback for different speech sounds, all participants were instructed to sustain the open vowel [a], the close vowel [i], and the nasalized syllable [mə] as steady as possible for at least 6 s. The nasalized syllable [mə] was included because it elicits an coarticulation of adjacent vowel at the very beginning of the following schwa and serves as a reference for the speaker to purposefully continue the nasalized vowel quality. The vocal intensity was real-time displayed on a laptop computer to help the speakers maintaining their vocal intensity within the range of 70–80 dBA in all auditory conditions.

The microphone-to-mouth distance was maintained at a distance of 15 cm by a stand holder, and the frequency response of the microphone was flat from 31.5 Hz to 8000 Hz (IEC 651 TYPE II, TENMARS Electronics, Taipei, Taiwan). In order to investigate whether and how the different types of nose masking would interfere with the auditory feedback for the speech material, five auditory conditions were introduced to the speakers during vocalizations: no-masking hearing status (NO), wearing headphone only (EO), speech-noise masking (SN, plateau energy from 0.25 kHz to 1 kHz, attenuation by 12 dB per octave from 1 kHz to 11.025 kHz), high-pass noise masking (HPN, plateau energy from 1 kHz to 8 kHz, decay by 12 dB per octave below 1 kHz), and broadband-noise masking (BBN, plateau energy from 0.25 kHz to 11.025 kHz). Two as-steady-possible phonations were recorded for each speech material in each auditory condition, and the analytic results of the two phonations were averaged for later data statistics. The order of the speech sounds and the auditory conditions were both arranged in random for each participant. The introduced noises were generated by a lab-developed program and a built-in sound adapter (ASUS A43S/Realtek high definition audio) and were binaurally introduced to the speakers at the intensity of 85 dBA through the headphones (Telephonics, TDH-50). Calibrations of the noises were accomplished prior to the tests for each participant using a standard sound level meter and a 6-c.c. coupler at the intensity of 80 dBA (Larson Davis system 824, New York, US). To control vocal intensity within the range between 70 and 80 dBA, there was a real-time intensity meter displayed on the screen so as to help the participants control their own vocal intensity. In each listening condition, the phonations were repeated once to acquire averaged data for statistical analysis as our previous works (Lee, 2012; Lee et al., 2004, 2007). No participant reported difficulty of producing the speech materials during voice recordings. The voice signals were obtained with a sampling frequency of 44.1 kHz and stored in a 16-bit format. The software for noise production, hardware controls, signal sampling, and intensity displaying was lab-developed using LabVIEW for Windows (version 6.0i, National Instrument, Austin, Texas, US). For realizing whether or not there is an interaction between auditory awareness and audio–vocal feedback system, right after both vocalizations in each type of auditory conditions, all participants subjectively rated the auditory awareness of their own voices by marking a 12-cm visual analogue scale in which 0 cm denoted “no auditory perception of their own voice” and 12 cm stood for a clear perception of their own voice as in normal listening status.”

2.3. Contour of F_0 and conversion of cents

The procedure details for digital signal processing had been published in our previous study (Lee, 2012). In short, the 5-s voice signals starting at 0.5 s after the voice onset were extracted for signal processing. A 20-ms window including at least two glottal cycles was used to obtain the fundamental period by counting the time at which the autocorrelation function was maximal. That period is compatible with the interval of a glottal wave that repeats itself. Then, the analytic windows were shifted forward by the fundamental periods, and all fundamental periods were retrieved

using the digital signal processing. The fundamental frequencies were eventually acquired by taking the reciprocals of fundamental period, and the jitter of F0 was also acquired to evaluate the voice quality. A vocalization with a jitter greater than 1% was excluded for data analysis.

Because F_0 can be different from one phonation to another and is also different between participants, the fundamental frequencies were normalized by the conversions of cent to allow comparisons within and between participants. Afterwards, exclusion of extreme values, re-sampling and linear interpolation of F_0 at the period of 20 ms were used to acquire a smooth contour of F_0 in cents. The signal processing was the same as our previously published studies (Lee, 2012; Lee et al., 2004, 2007).

2.4. Spectral analysis of F0 contour

The power spectrum of F0 was acquired using fast Fourier transform of the F0 contour in cent and was then divided into three powers: a low-frequency power (LFP) (0.2–3 Hz), a middle-frequency power (MFP) (3–8 Hz), and a high-frequency power (HFP) (8–25 Hz). The powers were calculated as the summation of the power amplitudes within each frequency range of the F0 spectrum and were expressed in decibels (dB) with the reference power of 1 $cent^2$ as 0 dB. Details of the calculations can be reviewed in our previous works (Lee, 2012; Lee et al., 2004, 2007). In brief, the power of each frequency band (LFP, MFP, and HFP) stands for the extent of modulations of F0 in its corresponding frequency range.

2.5. Formants analysis

For investigating the relationships between the distributions of acoustic energy and audio–vocal feedback, the frequency of the first two formants (F1 and F2) was obtained from each 20-ms analytical window using the digital signal processing of re-sampling (to 8000 Hz), pre-emphasis (+6 dB/octave above 50 Hz), and linear predictive coding (LPC) of Burg's method. The mean formant frequency of each phonation was acquired by averaging the formant frequency of all analytical windows, and the standard deviations of F1 frequency (F1SD%) and F2 frequency (F2SD%) were also obtained and divided by their means to represent the variability of the formant frequency in percentage. The results of the two phonations in each auditory condition were further averaged for statistical analysis.

For the nasalized syllable [mə], the signals exactly being analyzed for exploring the relationship between the nasal resonance and audio–vocal feedback was the data of nasalized vowel [ə] that here carried much higher nasalization than its typical resonance.

2.6. Analysis software and statistics

The mean F0, mean vocal intensity, LFP of F0, MFP of F0, HFP of F0, F1, F2, F1SD%, and F2SD% were analyzed from all signals for each participant, and were then submitted to significance testing using two-way (vowel category × listening condition) repeated measures ANOVA. A significant difference between groups was assumed if $p < 0.05$ using individual pair-wise comparison with Bonferroni correction. The gender difference was conducted by comparing between male participants and female participants the pooled data of all test conditions using Mann–Whitney U test, and a significant difference between groups was assumed if $p < 0.05$. The consistency of measuring F0, F1, F2, LFP, MFP, and HFP was tested using reliability analysis. Spearman rank order correlation analysis was used to assess the relationship between the power of F0 spectrum and the auditory awareness of the phonations. The software used for statistical analysis is SPSS for Windows, Ver. 17.0 (SPSS Inc., Chicago, IL, USA). A statistical significance was assumed if $p < 0.05$. Values are expressed as means ± standard error of the means (SEM).

3. Results

3.1. Vocal intensity

The vocal intensity of all participants was listed in Table 1. The mean vocal intensity was significantly different among the three types of speech sound [$F(2, 16) = 49.6$, $p < 0.05$]. The mean vocal intensity was also significantly different among the five auditory conditions [$F(4, 14) = 11.8$, $p < 0.05$]. The vowel [i] was 2.8 dB and 1.6 dB lower than the vowel [a] and the nasal syllable [mə], respectively ($p < 0.05$). The mean vocal intensity of SN and BBN was also significantly greater (1.1–1.4 dB) than NO condition ($p < 0.05$).

3.2. Vocal fundamental frequency

Fig. 1 showed the mean F0 of vowels [a], [i], and nasal syllable [mə] of the five auditory conditions. The mean F0 was significantly different among different speech sounds [$F(2, 16) = 17.0$, $p < 0.05$] with the vowels [i] had a significantly higher F0 than the vowels [i] and the nasal syllables [mə] ($p < 0.05$). The mean F0 of nasal syllable [mə] was also significantly higher than the vowel [a] ($p < 0.05$). However, for each of the three speech sounds, the F0 did not show a significant difference among the five auditory conditions [$F(4, 14) = 0.75$, $p > 0.05$], i.e., the F0 did not change significantly by the noise masking. The interaction between the types of speech sound and the auditory conditions was also not significant [$F(8, 10) = 1.5$, $p > 0.05$].

3.3. Power spectral analysis of vocal fundamental frequency

Fig. 2 demonstrates the mean LFP of vowels [a], [i], and syllable [mə] of all participants. The mean LFP was significantly different among different auditory conditions [$F(4, 14) = 15.3$, $p < 0.05$] rather than among different speech sounds [$F(2, 16) = 1.22$, $p > 0.05$]. The interaction between the types of speech sound and

Table 1
The vocal intensity of different speech sounds and auditory conditions.

	Intensity (dBA)[a]
Vowel/Syllable	
[a]	74.7 ± 0.58[†]
[i]	71.9 ± 0.59
[mə]	73.5 ± 0.54[†]
Auditory conditions	
NO	72.7 ± 0.59
EO	72.7 ± 0.54
SN	73.8 ± 0.58[‡]
HPN	73.6 ± 0.63
BBN	74.1 ± 0.52[‡]

NO, normal listening condition; EO, wearing earphone only; SN, speech noise masking; HPN, high-pass noise masking; BBN, broad-band noise masking.
[†] $p < 0.05$, compared with the intensity of vowel [i] using two-way repeated measures ANOVA and post-hoc Bonferroni procedures.
[‡] $p < 0.05$, compared with the intensity of NO using two-way repeated measures ANOVA and post-hoc Bonferroni procedures.
[a] The intensity values are expressed as mean ± standard error of the means.

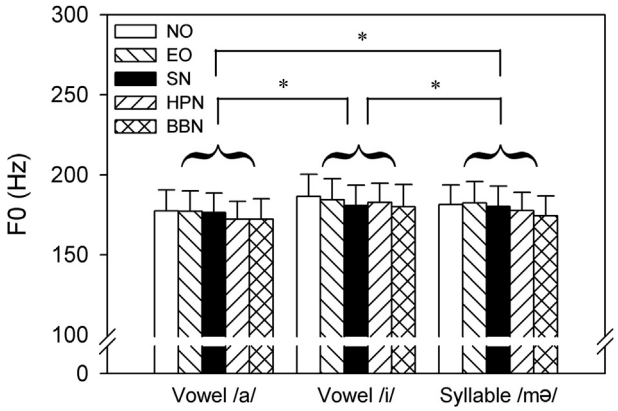

Fig. 1. The vocal fundamental frequency (F0) of vowels [a], [i], and [mə] sustained in the conditions of no noise masking (NO), wearing headphone only (EO), masking with speech noise (SN), masking with high-pass noise (HPN), and masking with broadband noise (BBN). $^{*}p < 0.05$, comparing between vowel group using two-way repeated measures ANOVA and multiple pairwise comparisons with Bonferroni correction.

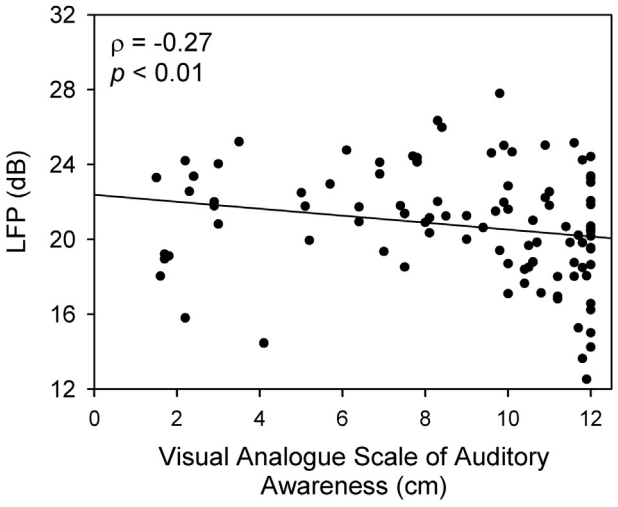

Fig. 3. The correlation between auditory awareness of speaker's own voice and audio–vocal feedback control of F0. LFP, audio–vocal feedback control of F0 evaluated by the low-frequency modulations of F0. The auditory awareness was rated using a 12-cm visual analogue scale with a clear perception of voice at the scale of 12 cm. The correlation was analyzed using Spearman rank order correlation.

the auditory conditions was not significant [$F(8, 10) = 0.79$, $p > 0.05$]. In individual pair-wise comparisons, the LFP of NO condition was significantly lower than the other four auditory conditions ($p < 0.05$) across the speech sounds. These results revealed the low-frequency modulations of F0 were closely associated with the auditory system, and the modulations increased with the attenuation of auditory sensory input by the earphone and or by the noises. Besides, the affected audio–vocal controls of F0 were consistent across the three different types of speech sound. Moreover, there was a significant and negative correlation (Fig. 3) between LFP and rating of auditory awareness of speakers' own voices (Spearman's rank order correlation, rho = −0.27, $p < 0.05$).

For MFP and HFP, there was no significant between-group difference of the five auditory conditions ($p > 0.05$). The results are consistent with the findings of our previous study, where the MFP and HFP showed no significant change in response to noise masking.

3.4. First formant and second formant

Fig. 4 showed the mean frequency of first formant (F1) and second formant (F2) of [a], [i], and [mə] of all participants in five different auditory conditions. The F1 frequency was significantly different among the speech sounds [$F(2, 16) = 135.21$, $p < 0.05$], where the vowel [a] had the highest F1 frequency of all, and the F1 of nasal syllable [mə] was also significantly higher than vowel [i] ($p < 0.05$). There was no significant between-group difference of the five auditory conditions [$F(4, 14) = 2.45$, $p > 0.05$]. For F2, the mean frequency was significantly different among the three vowels [$F(2, 16) = 98.15$, $p < 0.05$], and the vowel [i] was significantly higher than the other two speech sounds ($p < 0.05$). However, there was also no significant between-group difference among the five auditory conditions for all speech sounds [$F(4, 14) = 0.52$, $p > 0.05$].

For the variability of formant frequency, the only significant difference was present only among F1SD% of different vowels [$F(2, 16) = 3.43$; $p < 0.05$]. The between-group difference of the five auditory conditions, however, was not significant [$F(4, 14) = 2.67$, $p > 0.05$]. For the variability of F2 frequency that was indicated by F2SD%, there was neither between-vowel [$F(2, 16) = 0.17$, $p > 0.05$] difference nor between-condition difference of auditory feedback [$F(4, 14) = 0.95$, $p > 0.05$].

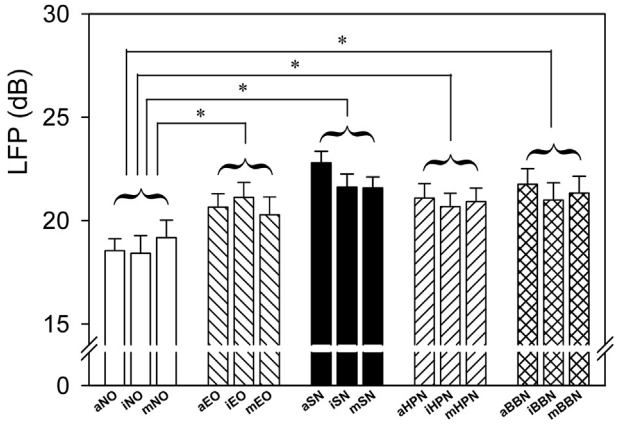

Fig. 2. The mean LFP of [a], [i], and [mə] in NO, EO, SN, HPN, and BBN conditions for all participants. $^{*}p < 0.05$, comparing between groups of auditory condition using two-way repeated measures ANOVA and multiple pairwise comparisons with Bonferroni correction.

3.5. Gender differences of LFP and formants

The mean F0, F1, F2, and LFP of male group and female group in 5 auditory conditions were listed in Table 2. The mean F0, mean F1 frequency, and mean F2 frequency of all speech sounds in females were significantly higher than those of males ($p < 0.05$, Mann–Whitney U). Both group showed an increase of LFP in the disturbed auditory conditions, although the female group revealed a more susceptibility to the auditory interferences. Moreover, the LFP of the female group was also significantly lower than the male group ($p < 0.05$, Mann–Whitney U).

3.6. Test–retest reliability

The consistency of measuring F0, F1, F2, LFP, MFP, and HFP of the two vocalizations in each auditory condition was tested using the reliability analysis. The values of Cronbach's alpha for the two measurements of F0, F1, F2, LFP, MFP, and HFP were 1.00, 0.96, 0.98, 0.78, 0.85, and 0.72, respectively.

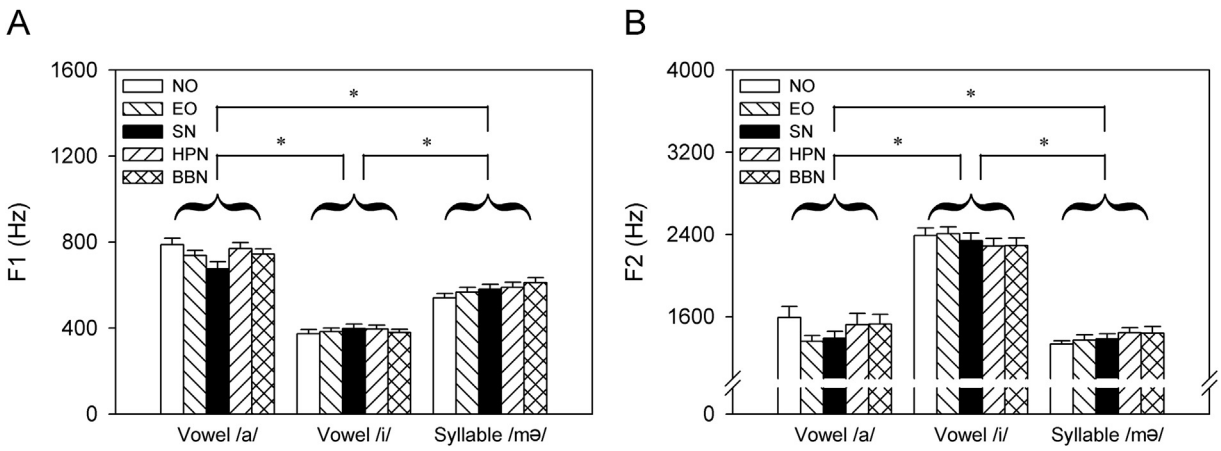

Fig. 4. The F1 (A) and F2 (B) of speech sounds [a], [i], and [mə] in the conditions of NO, EO, SN, HPN, and BBN for all participants. $^{*}p < 0.05$, comparing between vowel group using two-way repeated measures ANOVA and multiple pairwise comparisons with Bonferroni correction.

4. Discussion

The results revealed that the speakers tended to show a significant increase of LFP, i.e. the low-frequency modulation of F0, when their auditory feedback was interfered by noises or even only by the masking of a headphone no matter what speech sound was produced. However, there was no significant effect of disturbed auditory conditions on vocal F0 and formant frequencies for different speech materials used in this study. These findings provided an additional important clue that audio–vocal feedback acts predominately on the modulation of F0, especially the modulations of frequency below 3 Hz, rather than on the articulation of the first two formants in vowel production.

Speech acquisition and production require complex sensory-motor integration, and it has been well documented that auditory feedback shows a close relationship with the refinement and stabilization of speech motor control. Many researches focusing on audio–vocal feedback control were conducted to simulate the neural network model of speech production as well as examine the influences of auditory confliction on speech behaviors. Current behavioral and electrophysiological evidences have concluded that auditory inputs could not be recognized immediately after being transduced into electrical signals. These acoustic signals are initially stored in echoic memory for 250 ms–300 ms after stimulus offset (Massaro, 1975), and attention does not have a significant influence on the neural processing until the next stage in which the auditory stimuli are transformed into a percept for later pattern recognition (Naatanen, 1992; Tiitinen et al., 1994). The cortical event-related potential of mismatched negativity (MMN) is elicited by auditory deviations in a repetitive auditory features without auditory attention, and the latency is about 170 ms after stimulus offset (Näätänen et al., 1978). Coincidently, the neural transmission delay of corrective speech motor commands triggered by auditory conflictions in DIVA model is approximately 75 ms–150 ms (Guenther et al., 2006), and the responsive time of auditory error detection for a speech sound without involving auditory attention is therefore deduced to be within the time frame of an echoic memory (75–250 ms). These neural activations resulted from auditory errors are automatic and need no attention allocation.

Table 2
The mean F0, LFP, F1, and F2 values in the different auditory conditions.

	F0 (Hz)	LFP (dB)	F1 (Hz)	F2 (Hz)
Auditory conditions				
NO	179 ± 2.1	18.86 ± 0.31	565 ± 9.0	1759 ± 32.7
EO	175 ± 2.3	20.95 ± 0.33*	558 ± 9.6	1730 ± 34.9
SN	175 ± 2.1	22.08 ± 0.31*	545 ± 9.0	1699 ± 32.7
HPN	174 ± 2.1	21.02 ± 0.31*	579 ± 9.0	1738 ± 32.7
BBN	177 ± 2.1	21.37 ± 0.31*	578 ± 9.0	1754 ± 32.7
Gender				
Male	129 ± 1.5	21.99 ± 0.21	512 ± 13.4	1627 ± 36.7
Female	232 ± 1.9†	19.46 ± 0.29†	622 ± 14.8†	1868 ± 49.3†

The values of F0, LFP, F1, and F2 are expressed as mean ± standard error of the means.
*$p < 0.001$, compared with the LFP in normal listening condition using two-way (vowel category × listening condition) repeated measures ANOVA and multiple pairwise comparisons with Bonferroni correction.
†$p < 0.001$, compared with the male group using Mann–Whitney U test.

In this research, the reductions of auditory feedback lead to a significant change in F0 modulations but not in formant frequencies for the participants. The difference supports the idea that the F0 stability of a sustained phonation essentially relied on a close-loop control of speech production and was mainly maintained by a continuous audio–vocal feedback. Disturbances to the auditory input would result in reflex-like adjustments of F0 that made the F0s resemble a sway in a low frequency of <3 Hz. On the contrary, speech articulation appeared to be primarily guided by the stored feedforward motor commands that were established by previous language learning. The speakers were able to maintain the articulation even though they did not have enough auditory feedback of their own phonations, and thus the formant frequencies did not shift significantly in response to different types of noise masking.

Based on a parallel-distributed processing model of higher-order cognitive processing of acoustic stimulus (Rumelhart and McClelland, 1986), auditory information is supposed to be processed simultaneously in several regions of sensory memory system rather than to be moved from memory structure to another in a sequence. Thereby, this parallel processing model may allow for various neural mechanisms relative to audio–vocal feedback carrying on at the same time, such as deriving the phonetic features of an auditory stimulus, mismatching the acoustic characteristics to the target lexicon, analyzing the semantic structure, and perhaps modulating the vocal-fold rhythmic movements during sustained phonation. In the pitch shift studies, the speakers showed a voluntary and long-latency (310 ms–680 ms) feedback vocal response (VR2) to the pitch shift stimulus that was considered to be generated by a pathways having longer processing times or neural distances reasonably include the cerebral cortex and may reflect cognitive processing of perceived changes in auditory feedback (Burnett et al., 1998; Hain et al., 2000). The neural pathways might

reasonably include the cerebral cortex and reflect cognitive processing of perceived changes in auditory feedback. This long-latency response may be implemented as part of normal strategies for control of voice F0 that needs auditory attention system. In this study, the modulation of F0 showed a consistent and significant increase in the low-frequency range of $\leq$3 Hz across all three types of noise masking and three vowels. The responsive F0 modulations that remained in such a low-frequency range to noise masking might suggest the audio–vocal feedback control of F0 needs a long neural pathway to complete the action. Besides, the significant correlation of LFP with rating of auditory awareness of speakers' own voices suggests the involvement of auditory attention system in this type of feedback control. Most importantly, our results showed that the audio–vocal feedback of F0 could be evaluated using power spectral analysis of F0 with the speech material of sustained vowels, and the responses showed a consistency across vowels.

In our previous studies, the healthy participants presented with a significant increase of LFP under speech noise masking for the sustained vowel/a/ (Lee et al., 2004, 2007), and the low-frequency modulations of F0 also significantly increased in the hearing-impaired subjects (Lee, 2012; Lee et al., 2013). Our results were compatible with the above findings, however, they further revealed that the increase of LFP was significant in different sustained vowels that having different spectral energy when the auditory inputs were reduced by wearing earphone and by masking with noises that having different energy distributions. The audio–vocal control of F0 did not seem to be affected by the spectral energy distributions of auditory input information but appeared to be significantly affected by the amount of auditory input from speakers' own voices. From this point of view, the mechanism was deduced to be very different between control of F0 and articulation.

In a tonal language like mandarin Chinese, pitch is used to distinguish lexical or grammatical meaning by applying tonemes in words, and the control of tone is highly dependent on the control of F0. According to the pitch shift study, online control of voice F0 during vocalization is sensitive to language experience of the tonal language (Liu et al., 2010). However, the tonal features in the running speech of such language system, as well as the pitch shift stimuli, usually last for hundreds of ms. The F0 control might be different from a sustained vocalization of over 6 s of this research. The involvement of auditory attention system was evident for such sustained vocalizations. Nonetheless, the difference of F0 responses for sustained vowels with different tones remains unclear and needs further research to clarify.

In this research, the results showed that the auditory feedback is more important for voicing than articulation. The conclusion was supported by the significant increase of F0 modulations in the low frequency of <3 Hz and by no significant shift of F1 and F2 frequencies with the interference to auditory feedback. Moreover, for the acoustic analyses on F0 and formant frequencies to be parallel, the modulations of formant frequency revealed by F1SD% and F2SD % also did not alter significantly to the disturbance of auditory feedback. However, the effect of auditory feedback on articulation may be further explored by a more sensitive measurement. The analysis of vowel inherent spectral change and temporal variability of formant trajectory may be used to clarify the influence of auditory feedback on articulation in the future.

Although a full understanding of audio–vocal controls on the F0 rhythms and formants requires further investigations in running speech, the current findings provide evidence that there are different mechanisms of motor speech control for phonation and articulation movements in a sustained speech production. Speech articulation appears to be more associated with feedforward motor commands than with the real-time auditory inputs, and the rhythmic modulations of F0 are more sensitive to the decrease of auditory feedback. The measurement of low-frequency modulations of F0 is a good indicator for auditory feedback status The audio–vocal control of F0 and articulation can be explored using the model of this research for sustained speech production, and the methodology might be applied clinically to evaluate the audio–vocal motor control for hearing, neurological, voice disorders.

Conflict of interest

There was no conflict of interest to disclosure in this work.

Acknowledgments

This study was partially supported by the grant from National Science Council, Taiwan NSC 102-2314-B-010-029.

References

Aronson, A.E., Ramig, L.O., Winholtz, W.S., Silber, S.R., 1992. Rapid voice tremor, or "flutter," in amyotrophic lateral sclerosis. Ann. Otol. Rhinol. Laryngol. 101, 511.
Burnett, T.A., Freedland, M.B., Larson, C.R., Hain, T.C., 1998. Voice F0 responses to manipulations in pitch feedback. J. Acoust. Soc. Am. 103, 3153–3161.
Chen, S.H., Liu, H., Xu, Y., Larson, C.R., 2007. Voice F responses to pitch-shifted voice feedback during English speech. J. Acoust. Soc. Am. 121, 1157.
Ferrand, C.T., 2006. Relationship between masking levels and phonatory stability in normal-speaking women. J. Voice 20, 223–228.
Guenther, F.H., 2006. Cortical interactions underlying the production of speech sounds. J. Commun. Disord. 39, 16.
Guenther, F.H., Ghosh, S.S., Tourville, J.A., 2006. Neural modeling and imaging of the cortical interactions underlying syllable production. Brain Lang. 96, 22.
Hain, T.C., Burnett, T.A., Larson, C.R., Kiran, S., 2001. Effects of delayed auditory feedback (DAF) on the pitch-shift reflex. J. Acoust. Soc. Am. 109, 2146.
Hain, T.C., Burnett, T.A., Kiran, S., Larson, C.R., Singh, S., Kenney, M.K., 2000. Instructing subjects to make a voluntary response reveals the presence of two components to the audio-vocal reflex. Exp. Brain Res. 130, 133–141.
Hsiao, T.-Y., Solomon, N.P., Luschei, E.S., Titze, I.R., 1994. Modulation of fundamental frequency by laryngeal muscles during vibrato. J. Voice 8, 224–229.
Larson, C.R., Sun, J., Hain, T.C., 2007. Effects of simultaneous perturbations of voice pitch and loudness feedback on voice F0 and amplitude control. J. Acoust. Soc. Am. 121, 2862–2872.
Larson, C.R., Burnett, T.A., Bauer, J.J., Kiran, S., Hain, T.C., 2001. Comparison of voice F0 responses to pitch-shift onset and offset conditions. J. Acoust. Soc. Am. 110, 2845.
Lee, G.-S., 2012. Variability in voice fundamental frequency of sustained vowels in speakers with sensorineural hearing loss. J. Voice 26, 24–29.
Lee, G.-S., Hsiao, T.-Y., Kuo, T.B.J., 2004. Relationship between fundamental frequency system and audio-vocal reflex: illustration by power spectral analysis of vocal fundamental frequency. J. Taiwan Otolaryngol. Head Neck Surg. 39, 145–152.
Lee, G.-S., Liu, C., Lee, S.-H., 2013. Effects of hearing aid amplification on voice F0 variability in speakers with prelingual hearing loss. Hear. Res. 302, 1–8.
Lee, G.-S., Hsiao, T.-Y., Yang, C.C.H., Kuo, T.B.J., 2007. Effects of speech noise on vocal fundamental frequency using power spectral analysis. Ear Hear. 28, 343–350.
Leydon, C., Bauer, J.J., Larson, C.R., 2003. The role of auditory feedback in sustaining vocal vibrato. J. Acoust. Soc. Am. 114, 1575.
Liu, H., Wang, E.Q., Chen, Z., Liu, P., Larson, C.R., Huang, D., 2010. Effect of tonal native language on voice fundamental frequency responses to pitch feedback perturbations during sustained vocalizations. J. Acoust. Soc. Am. 128, 3739–3746.
Massaro, D.W., 1975. Perceptual processing in dichotic listening. J. Exp. Psychol. Hum. Learn. 2, 331–339.
Näätänen, R., Gaillard, A.W.K., Mäntysalo, S., 1978. Early selective-attention effect on evoked potential reinterpreted. Acta Psychol. 42, 313–329.
Naatanen, R., 1992. Attention and Brain Function. Lawrence Erlbaum Associates, Hillsdale, New Jersey.
Rumelhart, D., McClelland, J., 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Cambridge, MA.
Tiitinen, H., Sinkkonen, J., May, P., Näätänen, R., 1994. The auditory transient 40-Hz response is insensitive to changes in stimulus features. Neuroreport 6, 190–192.
Titze, I.R., 1991. A model for neurologic sources of aperiodicity in vocal fold vibration. J. Speech Lang. Hear. Res. 34, 460.
Titze, I.R., Baken, R., Herzel, H., 1993. Evidence of chaos in vocal fold vibration. In: Vocal Fold Physiology: New Frontier in Basic Science, pp. 143–188.
Tourville, J.A., Reilly, K.J., Guenther, F.H., 2008. Neural mechanisms underlying auditory feedback control of speech. Neuroimage 39, 1429–1443.