Vocal-tract spectrum estimation method affects the articulatory compensation in formant transformed auditory feedback

Yasufumi Uezu*, Sadao Hiroya† and Takemi Mochida‡

Human Information Science Laboratory, NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, 3–1 Morinosato-Wakamiya, Atsugi, 243–0198 Japan
(Received 17 October 2019, Accepted for publication 17 January 2020)

Abstract: Auditory feedback has a crucial role in stably controlling speaking and singing. Formant-transformed auditory feedback (TAF) is used to investigate the relationship between perturbation to the formant frequency and the compensatory response to clarify the mechanism of auditory-speech motor control. Although previous studies for formant TAF applied linear predictive coding (LPC) to estimate formant frequencies, LPC estimates false formants for high-pitch voice. In this paper, we investigate how different vocal-tract spectrum estimation methods in real-time formant TAFs affect the compensatory response of formant frequencies to perturbations. A phase equalization-based autoregressive exogenous model (PEAR) is applied to the TAF system as a formant estimation method that can estimate the formant frequency more accurately and robustly than LPC can. Fifteen Japanese native speakers were asked to repeat the Japanese syllables /he/ or /hi/ while receiving feedback sounds whose formants F1 and F2 were transformed. From the results for the /he/ condition, the F1 compensatory response for PEAR was significantly larger than that of LPC, and the compensation error in the F1–F2 plane for PEAR was less than that for LPC. Our results suggest that PEAR can increase both the accuracy of formant frequency estimation and the naturalness of the transformed speech sound.

Keywords: Formant frequency, Transformed auditory feedback, PEAR, LPC, Auditory-speech motor control

PACS number: 43.72.Ar, 43.70.–h, 43.70.Mn [doi:10.1250/ast.41.720]

1. INTRODUCTION

Auditory feedback has a crucial role in stably controlling speaking and singing. Speakers monitor their own speech sound in real time while speaking, and if the monitored speech sound has some acoustic errors, they alter their own speech articulators to compensate for them. Elucidating the sensorimotor and cognitive process underlying the articulatory compensation is important for understanding the mechanisms supporting speech communication.

Formant-transformed auditory feedback (TAF) is one of the experimental methods used to investigate the mechanism of auditory-speech motor control [1]. A speaker repeats words or syllables while receiving his/her own speech through headphones. As the speaking task progresses, formant perturbation is gradually given to the speech sound by a real-time formant transformation system. Finally, the speaker receiving the formant-shifted speech sound comes to utter compensatory responses against the perturbations.

In order to perform formant TAF, a real-time formant estimation system is required. Previous studies of formant TAF have used various methods to estimate formant frequencies. Houde and Jordan [1] performed the formant TAF using a technique to estimate the formant frequency from the spectral centroid of the speech signal when the speaker repeated “head” in a whispered voice. This technique was used in the early experiments with formant TAF and in the TAF for fricative sounds such as /s~/[2–4]. However, formants estimated from the spectral centroid are unreliable, and the quality of the transform speech sound is deteriorated.

Another method of performing a pseudo formant transformation without estimating formant frequencies is to use an effector. Rochet-Capellan et al. [5] performed

*a–e-mail: yasufumi.uezu.xh@hco.ntt.co.jp
†e–mail: sadao.hiroya.kv@hco.ntt.co.jp
‡e–mail: takemi.mochida.sw@hco.ntt.co.jp
formant TAF using a VoiceOne (TC Helicon). A transformed speech sound with relatively high naturalness was generated by mixing a low-pass-filtered speech signal, in which the first formant frequency was lowered artificially by VoiceOne, and a high-pass-filtered speech signal. Numerous studies of formant TAF with the VoiceOne system have been reported [6–11]. However, the amount of perturbation given to formant frequencies cannot be controlled appropriately since VoiceOne gives perturbations without formant frequency estimation. In addition, the feasible perturbation pattern is limited because this method gives only a perturbation that shifts the first formant frequency downward. Besides, VoiceOne is no longer in production, making it difficult to replicate the experiments.

Linear predictive coding (LPC) [12] is widely used for formant TAF because of its theoretical relationship with the speech production model and its low computational complexity. Purcell and Munhall [13,14] reported formant TAF by applying the iterative Burg algorithm. Villacorta et al. [15] reported formant TAF by applying a formant transformation system implemented with an autocorrelation method. Cai et al. [16] developed “Audapter” by combining LPC and the cepstrum method for formant estimation for high-pitch voices. Since then, formant TAF using LPC has been used to examine the effects of higher functional impairments in Parkinson’s disease and cerebellar degeneration, as well as of stuttering and autism spectrum disorders, on the control of auditory-speech motor control [17–22].

Although LPC is a relatively easy signal processing method for formant estimation, there are problems in the accuracy and robustness of estimation. When the fundamental frequency ($f_0$) of speech sound is high, the formant frequency estimated by LPC is often biased towards the frequency position of $f_0$ or its harmonic component because LPC assumes white noise as a sound source [23]. This degrades the naturalness of speech sound synthesized by the formant frequency estimated by LPC, and the degradation is more remarkable when the formant frequency perturbation is larger in formant TAF. Therefore, instead of LPC, a more accurate and robust formant estimation method should be incorporated into the formant transformation system to understand the mechanism of auditory-speech motor control in detail.

We have developed a phase equalization-based autoregressive exogenous (PEAR) model [24,25] to estimate formant frequencies. PEAR can estimate the formant frequency more accurately and robustly than LPC can even for high-$f_0$ speech sounds. Therefore, synthesized speech sound using formants estimated by PEAR may be more natural.

According to previous studies, the magnitude of the compensations of perturbations in auditory-speech motor control tends to be smaller than that in other motor control. In particular, the ratio of compensation to formant perturbation has been reported to be at most about 40% [26]. This partial compensation is thought to stem from unperturbed somatosensory feedback [18,26] and the existence of vowel categories [27]. However, we suspect that the lack of naturalness in the perturbed auditory feedback generated by the TAF system using LPC is one of the reasons.

In this study, we performed formant TAF experiments and investigated the effect of formant estimation methods (PEAR vs. LPC) implemented in a real-time formant transformation system.

2. TRANSFORMATION METHOD

2.1. Problems with the LPC Method

To shift formant frequencies, decomposition of speech signals into a vocal-tract spectrum and source signals is required based on the source-filter theory [28]. First, LPC analysis calculates LPC coefficients (i.e., the vocal-tract spectrum) from speech signals [12]. Then, LPC residual signals (i.e., source signals) are obtained by an inverse filtering. The formant frequencies are obtained by finding the roots of the LPC polynomial. At the transformation stage, the formant frequencies are shifted, and then the transformed LPC coefficients are generated. The transformed speech signals are obtained by convoluting the transformed LPC coefficients and LPC residual signals.

However, misestimating formant frequencies by LPC would result in formants being included in source signals. Figure 1 shows a vocal-tract spectrum estimated by LPC and the spectrum of LPC residual signals for synthesized speech with $f_0 = 130$ and 190 Hz. If the $f_0$ is low, the spectrum of LPC residual signals shows no peaks and dips.
(i.e., a flat residual spectrum) because of the accurate estimation of formant frequencies. However, if it is high, the first two formant frequencies match the harmonics of the \( f_0 \) due to a misestimation of the formants. As a result, the spectrum of LPC residual signals shows peaks and dips. Thus, the spectrum of transformed speech would include inappropriate peaks and dips, and then less natural sounds of transformed speech signals would be generated. Note that there is no problem if formant frequencies are not transformed such as in the case of speech coding.

### 2.2. PEAR Method

To overcome this problem, we developed a method that can estimate a vocal-tract spectrum using PEAR [25]. The speech signals can be represented by the following speech production model:

\[
s(t) = \sum_{p=1}^{P} a(p)s(t - p) + u(t),
\]

where \( s \) represents the original speech signals, \( a \) denotes the LPC coefficients, \( P \) is the dimension of the LPC coefficients, and \( u \) is the source signals.

LPC assumes Gaussian noise as source signals [12]. However, this assumption deviates from actual signals, especially for a high \( f_0 \). On the other hand, PEAR assumes a periodic impulse source signal for voiced speech and uses phase equalization to convert the phase characteristics of the original speech signals to the minimum phase [29]. The phase-equalized speech signals are computed by

\[
s'(t) = \sum_{\tau=-M}^{M} h(\tau)s(t - \tau),
\]

where

\[
h(t) = e(t_0 - t) \sqrt{\sum_{\tau=-M}^{M} e(t_0 + \tau)^2}.
\]

Taken together, the LPC coefficients \( \hat{a} \) are obtained by solving the following equation:

\[
\begin{pmatrix}
R(0) & \ldots & R(P-1) \\
\vdots & \ddots & \vdots \\
R(P-1) & \ldots & R(0)
\end{pmatrix}
\begin{pmatrix}
a(1) \\
\vdots \\
a(P)
\end{pmatrix}
= \begin{pmatrix}
R(1) - V(1) \\
\vdots \\
R(P) - V(P)
\end{pmatrix},
\]

where

\[
V(p) = \sum_{t=0}^{I} w(t_i - p)w(t_i) \sum_{\tau=-M}^{M} e(t_i - \tau)s(t_i - \tau - \tau),
\]

where \( e \) is a LPC residual signal, \( R \) is an autocorrelation function of the input speech signal, \( w \) is a window function, \( t_i \) is a pitch mark (glottal closing instant or peak time of residual signal), \( I + 1 \) is the number of pitch marks in the frame, and \( 2M + 1 \) is the number of taps of the phase equalization filter.

Assuming Gaussian noise as source signals, \( V(p) = 0 \) is obtained when there is no \( t_i \). In that case, Eq. (4) is equivalent to the autocorrelation method for LPC. Thus, PEAR can be said to be a generalization of LPC. The computational complexity of PEAR is less than twice that of LPC because only the calculation of \( V \) is added. To estimate a time-stable spectrum, TANDEM windows are used in both LPC and PEAR [25].

To accurately determine the pitch mark \( t_i \), especially for a voice having a high fundamental frequency, the glottal closing instants were directly extracted using the differential value of the electroglottography (EGG) signal, and the time at which the value of the residual signal in the vicinity became the maximum was determined as the pitch mark.

### 3. EXPERIMENTS

#### 3.1. Experimental Conditions

Fifteen native Japanese speakers (eight females) participated in the experiments. Their ages ranged from 21 to 42 years, and the mean age was 30.5 years. None of the speakers reported speech impairment, and all had normal hearing thresholds. All gave informed consent to participate in the study, which was approved by the NTT Communication Science Laboratories Research Ethics Committee.

The experiment was conducted in a soundproof room. Figure 2 shows a block diagram of the study. The speakers sat in front of the microphone (SONY ECM-330) with headphones (SENNHEISER HD280 pro) and the two electrodes of the EGG (Glottal Enterprises EG2) attached.
to the speaker’s neck. The speech signals from the microphone and the EGG signals were amplified (M-Audio DMP3), low-pass-filtered (MTT MS 2319-B) at a cut-off frequency ($F_c$) of 4 kHz, A/D converted at 8 kHz, and transmitted to a real-time formant transformation system (Renesas Electronics SH 7785) to generate formant-shifted speech signals. The transformed speech signals were D/A converted, low-pass-filtered at $F_c = 4$ kHz, amplified (Audio-Technica AT-MA 55), mixed with pink noise (Bruel & Kjaer Type 1405) to mask bone-conducted signal, and presented via headphones with a delay of 12 ms. In previous studies, the processing delay was less than 20 ms [15,16], so the delay in the present study is adequate for TAF experiments. The speech and noise were presented at 72 and 45 dBA SPL, respectively. Speakers were encouraged to utter at a natural rate and level with timing controlled by a prompt on a monitor. Each prompt lasted 2 s, and the intertrial interval was approximately 3.5 s.

Speakers were asked to produce Japanese vowels /i/ and /e/ in an /hV/ context. The words were shown on the monitor in the hiragana “ち” for /hi/ and “へ” for /he/. Figure 3 shows perturbation patterns of the first (F1) and second (F2) formant frequency in a block including 140 uttering trials. A block consisted of four phases: Baseline (Trial 1–20), Ramp (Trial 21–70), Hold (Trial 71–90), Return (Trial 91–140). In the Baseline phase, speakers produced utterances with no transformed feedback. In the Ramp phase, speakers produced utterances while receiving transformed speech feedback. The absolute value of the magnitude of perturbation in the formant frequency was increased by a certain amount per trial. In the Hold phase, speakers uttered while receiving feedback speech with formant frequencies transformed at the maximum level of perturbation. In the Return phase, speakers produced utterances with normal feedback, which was the same as in the Baseline phase, to wash out aftereffects. The maximum levels of perturbation for formant frequencies were $(F1, F2) = (-150, +200)$ Hz for the /hi/ and $(F1, F2) = (+200, -200)$ Hz for the /he/ conditions. These values were determined based on preliminary experiments and previous studies [30].

Speech signals were pre-emphasized by a first-order high-pass filter. Then, a 16-ms Blackman window was applied, and LPC coefficients were obtained every 4 ms. Either PEAR or LPC were selected as the formant estimation method implemented in the real-time formant frequency transform system. Four test conditions (PEAR or LPC) × (/i/ or /e/) were examined separately in a series of four blocks. In order to increase the reliability of the results, speakers performed the four test conditions twice (totally eight blocks), in the order of blocks randomized within each series of four blocks.

The numbers of LPC coefficients (8 or 10) and taps of the phase equalization filter (9 or 17) for each of the speakers were determined by calibration. Note that the EGG electrodes were attached for both LPC and PEAR. The vowel interval was determined based on the amplitude of the speech signals and the autocorrelation of EGG signals.

3.2. Analysis

Formant analysis of the speech uttered by speakers was performed offline, and the median value from 40 to 80% of the vowel interval was used as a representative value of each trial. Here, PEAR method was applied for both PEAR and LPC conditions to estimate formant frequencies offline precisely.

Amounts of compensatory response in formant frequencies were determined by subtracting the value in the Baseline phase from that in the Hold phase. Baseline and Hold values of formant frequencies in each block were set...
by computing the median value from Baseline (6–20 trials) and Hold (71–90 trials) formants, respectively [17].

We obtained the amount of compensatory responses to formant frequencies F1 and F2 for every combination of the vowels (/i/ or /e/) and formant estimation method used in TAF (PEAR or LPC) for each subject. The mean values of compensatory responses in F1 and F2 of all subjects in each condition were calculated as the average compensatory responses.

### 4. RESULTS

Figures 4 and 5 show the amount of change in F1 and F2 in each trial when speakers repeated /hi/ and /he/ syllables in formant TAF. Shaded regions denote the standard error.

The formant change in the /he/ for PEAR is larger than for LPC. In particular, in the F1 Hold phase, the formant change appeared to saturate in LPC, whereas in PEAR it continued to increase until the middle of the Hold phase. On the other hand, the formant changes in /hi/ did not differ between PEAR and LPC.

Figure 6 shows the average compensatory responses of the formant frequencies when PEAR or LPC is used as a real-time formant transform system for formant TAF with repeated syllables /hi/ and /he/. An exact Wilcoxon signed rank test revealed that compensatory response at Hold was significantly larger than zero ($p < 0.05$), except for F1 of /hi/ using LPC.

It seems that PEAR has a larger compensatory response than LPC for F1 in /he/ utterance condition. An exact Wilcoxon signed rank test showed that there was a significant greater compensation in F1 with PEAR than with LPC ($p = 0.04$) while /he/ was uttered. This shows that PEAR provides a large formant compensatory response compared to LPC as a formant estimation method for the formant transform system used in formant TAF.
Figure 7 shows the average compensatory responses shown in the Fig. 6 in the F1–F2 space during /he/ production. The two arrows in the figure represent the compensation of PEAR and LPC as vectors. Origin (0, 0) Hz indicates the baseline. The dotted line is a straight line through the origin and the maximum perturbation (+200, –200) Hz. Thus, the ideal compensation would be on the dotted line. The LPC’s compensation vector diverges considerably from the dotted line, while PEAR’s compensation vector is near. This suggests that using PEAR instead of LPC may not only improve the amount of compensatory response but also reduce the error in the compensatory response.

5. DISCUSSION

We investigated how PEAR and LPC, which are formant frequency estimation methods incorporated in a real-time formant transform systems, affect the compensatory response in speech under formant TAF. As a result, it was confirmed that PEAR produced a larger F1 compensatory response than LPC in the /he/ condition. Moreover, the compensation error in the F1–F2 plane for PEAR was less than that for LPC during /he/ production. The results from our study can be attributed to some differences between PEAR and LPC.

5.1. Formant Estimation Error

The LPC method often misestimates formant frequencies for high-pitch voice. Cai et al. [16] noted that this formant estimation problem against $f_0$ should be solved in formant TAF. Oohashi et al. examined the accuracy of formant estimation by PEAR and LPC methods using synthesized speech sounds with various $f_0$s [25]. They showed that the F1 estimation error is smaller in PEAR method than in LPC method when the $f_0$ of synthesized speech sounds is high. This suggested that PEAR method has a smaller F1 estimation error for actual speech with a high $f_0$ compared to LPC method.

Using this experimental data, we examined whether the F1 estimation error of LPC method is large for a high $f_0$. Here, the correct value of the formant frequency of the uttered speech sounds is not known. Therefore, the absolute values of the difference between F1 estimated by PEAR and LPC methods were regarded as the F1 estimation errors. Baseline speech data (6–20 trials) under PEAR and LPC conditions for /he/ were used because the speech data was considered to be stable. The same speech data were analyzed in both PEAR and LPC methods. Figure 8 shows that there is a marginally significant correlation ($p = 0.05$) between the $f_0$ height and the F1 estimation error in the /he/ condition. Assuming that F1 estimated by PEAR method was more accurate than that estimated by LPC method, this indicates that PEAR method can perform F1 estimation relatively robustly even in actual speech with a high $f_0$. 

Fig. 6 Average compensatory responses of formant frequencies F1 and F2 when PEAR or LPC is used in a real-time formant transformation system for formant TAF with repeated syllables /hi/ (left) and /he/ (right). The error bar indicates the standard error.

Fig. 7 Vectors of compensatory responses in F1 and F2 for /he/ using PEAR or LPC.
5.2. Difference in Compensatory Response

Now, where did the difference in compensatory response between PEAR and LPC come from? One possible cause is that F1 estimation error simply biases the F1 compensation response. We examined the relationship between the difference in compensatory responses of F1 between PEAR and LPC and the F1 estimation error under the /he/ utterance condition. There was no correlation between the difference in the amount of compensatory response of F1 and the F1 estimation error ($r^2 = 0.18749$). Moreover, not only individual correlation but also average value were examined. As shown in Fig. 8, the average F1 estimation error was 13 Hz under the /he/ utterance condition. However, as shown in Fig. 6, the average difference between PEAR and LPC in F1 compensatory response under the /he/ condition was more than 20 Hz, which was larger than the average F1 estimation errors. These results indicate that the difference in the amount of compensatory responses between PEAR and LPC cannot be explained only by differences in formant estimates.

As described in Sect. 2.1, transformed speech signals are synthesized using the transformed formants and residual signals. Thus, the flatness of the residual spectrum and the direction and magnitude of the perturbation affect the naturalness of transformed speech sounds. Note that naturalness of a transformed speech means that a speaker can speak it, i.e., the natural transformed speech should be synthesized from a source signal and vocal-tract spectrum suitable for human articulatory constraints. LPC method often generates a non-flat residual spectrum and an incorrect formant, resulting in a buzzer sound in transformed speech. Thus, it is considered that such unnatural transformed speech makes the compensation difficult in TAF using LPC method. On the other hand, PEAR method is more likely to produce more a natural transformed speech than LPC method does, because it provides residual signals and a vocal-tract spectrum suitable for human articulatory constraints. Therefore, we speculate that the difference in the compensatory response between PEAR and LPC is related to the naturalness of transformed speech overall by a combination of formant estimation error, the flatness of the residual spectrum, etc.

In the introduction, we stated that the ratio of compensation for formant perturbation tends to be smaller than that of other types of motor control. Mitsuya et al. [17] reported that Japanese native speakers produced a smaller formant compensation than English native speakers, about 10% for perturbation. Our results showed that the ratio of compensation for formant perturbation in LPC is equivalent to results of previous Japanese studies [17], but the compensation in PEAR exceeded it. This suggests that one of the reasons for the small amount of compensatory response may be the vocal-tract spectrum estimation method, i.e., the naturalness of transformed speech.

As described in Sect. 2.1, transformed speech signals are synthesized using the transformed formants and residual signals. Thus, the flatness of the residual spectrum and the direction and magnitude of the perturbation affect the naturalness of transformed speech sounds. Note that naturalness of a transformed speech means that a speaker can speak it, i.e., the natural transformed speech should be synthesized from a source signal and vocal-tract spectrum suitable for human articulatory constraints. LPC method often generates a non-flat residual spectrum and an incorrect formant, resulting in a buzzer sound in transformed speech. Thus, it is considered that such unnatural transformed speech makes the compensation difficult in TAF using LPC method. On the other hand, PEAR method is more likely to produce more a natural transformed speech than LPC method does, because it provides residual signals and a vocal-tract spectrum suitable for human articulatory constraints. Therefore, we speculate that the difference in the compensatory response between PEAR and LPC is related to the naturalness of transformed speech overall by a combination of formant estimation error, the flatness of the residual spectrum, etc.

In the introduction, we stated that the ratio of compensation for formant perturbation tends to be smaller than that of other types of motor control. Mitsuya et al. [17] reported that Japanese native speakers produced a smaller formant compensation than English native speakers, about 10% for perturbation. Our results showed that the ratio of compensation for formant perturbation in LPC is equivalent to results of previous Japanese studies [17], but the compensation in PEAR exceeded it. This suggests that one of the reasons for the small amount of compensatory response may be the vocal-tract spectrum estimation method, i.e., the naturalness of transformed speech.

However, this study did not directly examine the relationship between the naturalness of transformed speech sound and the compensatory response, so it cannot be concluded that our results can be explained simply by naturalness. Therefore, it is necessary to establish a method of controlling the naturalness of transformed speech appropriately to examine the relationship between the naturalness and compensatory response directly, and to investigate the naturalness of transformed speech during speech production perceptually.

5.3. Compensation for F2

Under formant TAF in the /he/ utterance condition, there was a significant difference in the amount of F1 compensatory responses between PEAR and LPC; however, there was no significant difference in the amount of F2 compensatory response. One of the reasons is that the estimation error of F2 is small even when LPC method is used [25]. However, as is clear from Fig. 7, the vector of PEAR is closer to the diagonal dashed line that indicates an ideal articulatory compensatory response. Thus, it is considered that since the compensation by PEAR was reasonable in relation to F1 and F2, there was no significant difference in the amount of F2 compensatory responses. This suggests that improving the naturalness of transformed speech can produce a more ideal articulatory compensatory response under the formant TAF.

5.4. Compensation for /hi/

In contrast to the results for the /he/ speech condition, there was no significant difference in formant compensatory response between PEAR and LPC under the /hi/
speech condition, although the amount of compensatory response for the baseline of F1 of /hi/ in LPC was not significant. Mitsuya et al. pointed out the effect of somatosensory feedback in formant TAF [18]. They performed formant TAF with various English vowels and showed an asymmetric compensatory response of F1 to its perturbation direction for /i/-like vowels but not for /e/. This suggested that /i/-like vowels are influenced by somatosensory feedback caused by narrowing of the tongue and palate rather than auditory feedback. Therefore, since there was no difference in unperturbed somatosensory feedback between PEAR and LPC, there would be no significant difference in formant compensatory response for /hi/ between them.

6. CONCLUSIONS
In this study, we investigated the effect of formant estimation methods implemented in a real-time formant transformation system on the F1 and F2 compensatory response. PEAR produced a greater compensatory response than LPC. In addition, it was suggested that PEAR reduces the error in compensatory responses. These results indicate that PEAR, which can achieve high estimation accuracy of formant frequency and the resulting transformed speech sound with high naturalness, is a useful technique for formant TAF to better understand the interaction between hearing and speech. Furthermore, these findings suggest that the naturalness of speech sound fed back to the speaker may play a very important role in auditory-speech motor control.

REFERENCES
[1] J. F. Houde and M. I. Jordan, “Sensorimotor adaptation in speech production,” Science, 279(5354), 1213–1216 (1998).
[2] J. F. Houde and M. I. Jordan, “Sensorimotor adaptation of speech I: Compensation and adaptation,” J. Speech Lang. Hear. Res., 45, 295–310 (2002).
[3] D. M. Shiller, M. Sato, V. L. Gracco and S. R. Baum, “Perceptual recalibration of speech sounds following speech motor learning,” J. Acoust. Soc. Am., 125, 1103–1113 (2009).
[4] D. M. Shiller, V. L. Gracco and S. Rvachew, “Auditory-motor learning during speech production in 9–11-year-old children,” PLOS ONE, 5(9), e12975 (2010).
[5] A. Rochet-Capellan and D. J. Ostry, “Simultaneous acquisition of multiple auditory–motor transformations in speech,” J. Neurosci., 31, 2657–2662 (2011).
[6] M. Shum, D. M. Shiller, S. R. Baum and V. L. Gracco, “Sensorimotor integration for speech motor learning involves the inferior parietal cortex,” Eur. J. Neurosci., 34, 1817–1822 (2011).
[7] D. R. Lametti, A. Rochet-Capellan, E. Neufeld, D. M. Shiller and D. J. Ostry, “Plasticity in the human speech motor system drives changes in speech perception,” J. Neurosci., 34, 10339–10346 (2014).
[8] L. Max and D. G. Maffett, “Feedback delays eliminate auditory-motor learning in speech production,” Neurosci. Lett., 591, 25–29 (2015).
[9] N. J. Bourguignon, S. R. Baum and D. M. Shiller, “Please say what this word is-Vowel-extrinsic normalization in the sensorimotor control of speech,” J. Exp. Psychol. Hum. Percept. Perform., 42, 1039–1047 (2016).
[10] F. Mollaei, D. M. Shiller, S. R. Baum and V. L. Gracco, “Sensorimotor control of vocal pitch and formant frequencies in Parkinson’s disease,” Brain Res., 1646, 269–277 (2016).
[11] M. R. van den Bunt, M. A. Groen, T. Ito, A. A. Francisco, V. L. Gracco, K. R. Pugh and L. Verhoeven, “Increased response to altered auditory feedback in dyslexia: A weaker sensorimotor magnet implied in the phonological deficit,” J. Speech Lang. Hear. Res., 60, 654–667 (2017).
[12] F. Itakura and S. Saito, “A statistical method for estimation of speech spectral density and formant frequencies,” Electron. Commun. Jpn., A, 53(1), 36–43 (1970).
[13] D. W. Purcell and K. G. Munhall, “Compensation following real-time manipulation of formants in isolated vowels,” J. Acoust. Soc. Am., 119, 2288–2297 (2006).
[14] D. W. Purcell and K. G. Munhall, “Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation,” J. Acoust. Soc. Am., 120, 966–977 (2006).
[15] V. M. Villacorta, J. S. Perkell and F. H. Guenther, “Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception,” J. Acoust. Soc. Am., 122, 2306–2319 (2007).
[16] S. Cai, S. S. Ghosh, F. H. Guenther and J. S. Perkell, “Adaptive auditory feedback control of the production of formant trajectories in the Mandarin triphthong /iu/ and its pattern of generalization,” J. Acoust. Soc. Am., 128, 2033–2048 (2010).
[17] T. Mitsuya, E. N. MacDonald, D. W. Purcell and K. G. Munhall, “A cross-language study of compensation in response to real-time formant perturbation,” J. Acoust. Soc. Am., 130, 2978–2986 (2011).
[18] T. Mitsuya, E. N. MacDonald, K. G. Munhall and D. W. Purcell, “Formant compensation for auditory feedback with English vowels,” J. Acoust. Soc. Am., 138, 413–424 (2015).
[19] M. K. Franken, D. J. Acheson, J. M. McQueen, F. Eisner and P. Hagoort, “Individual variability as a window on production-perception interactions in speech motor control,” J. Acoust. Soc. Am., 142, 2007–2018 (2017).
[20] P. Trudeau-Fisette, M. Tiede and L. Ménard, “Compensations to auditory feedback perturbations in congenitally blind and sighted speakers: Acoustic and articulatory data,” PLOS ONE, 12(7), e0180300 (2017).
[21] C. Demopoulos, H. Kothare, D. Mizuiri, J. Henderson-Sabes, B. Fregeau, J. Tjernagel, J. F. Houde, E. H. Sherr and S. S. Nagarajan, “Abnormal speech motor control in individuals with 16p11.2 deletions,” Sci. Rep., 8, 1274 (2018).
[22] C. D. Martin, C. A. Niziolek, J. A. Duñabetia, A. Perez, D. Hernandez, M. Carreiras and J. F. Houde, “Online adaptation to altered auditory feedback is predicted by auditory acuity and not by domain-general executive control resources,” Front. Hum. Neurosci., 12 (2018).
[23] S. Hiroya, “Formant analysis of vowels: Process and hypotheses,” J. Acoust. Soc. Jpn. (J), 70, 538–544 (2014) (in Japanese).
[24] S. Hiroya and T. Mochida, “Phase equalization-based autoregressive model of speech signals,” in Proc. Interspeech, pp. 42–45 (2010).
[25] H. Oohashi, S. Hiroya and T. Mochida, “Real-time robust formant estimation system using a phase equalization-based autoregressive exogenous model,” Acoust. Sci. & Tech., 36, 478–488 (2015).
[26] S. Katseff, J. Houde and K. Johnson, “Partial compensation for altered auditory feedback: A tradeoff with somatosensory
feedback?" *Lang. Speech*, **55**, 295–308 (2012).

[27] C. A. Niziolek and F. H. Guenther, “Vowel category boundaries enhance cortical and behavioral responses to speech feedback alterations,” *J. Neurosci.*, **33**(29), 12090–12098 (2013).

[28] G. Fant, *Acoustic Theory of Speech Production* (Walter de Gruyter, The Hague, 1970).

[29] M. Honda, “Speech coding using waveform matching based on LPC residual phase equalization,” *Proc. ICASSP 90*, pp. 213–216 (1990).

[30] T. Caudrelier, J.-L. Schwartz, P. Perrier, S. Gerber and A. Rochet-Capellan, “Transfer of learning: What does it tell us about speech production units?” *J. Speech Lang. Hear. Res.*, **61**, 1613–1625 (2018).