Weighting of Auditory Feedback Across the English Vowel Space

David Purcell[1], Kevin Munhall[2]
[1]University of Western Ontario, London, ON
[2]Queen's University, Kingston, ON
E-mail: purcelld@nca.uwo.ca

Abstract

Auditory feedback in the headphones of talkers was manipulated in the F1 dimension using a real-time vowel formant filtering system. Minimum formant shifts required to elicit a response and the amount of compensation were measured for vowels across the English vowel space. The largest response in production of F1 was observed for the vowel /ɛ/ and smaller or non-significant changes were found for point vowels. In general, changes in production were of a compensatory nature that reduced the error in the auditory feedback.

1 Introduction

Sensory feedback plays an important role in the acquisition of speech. However, adult speech production is also influenced by feedback both on long and short timescales. The auditory modality is particularly influential in the speech motor control of both segmental and supra-segmental elements. For example, perturbations in the auditory feedback of voice pitch are responded to within a few hundred milliseconds by production changes in a compensatory direction [1]. Similar compensatory production changes also occur in response to the manipulation of vowel formants [2, 3, 4].

Experiments have investigated the role of auditory feedback in the ongoing maintenance of accurate vowel production. These investigations have employed real-time formant tracking and formant manipulation to present altered vowel acoustics through headphones worn by the talker. In response to a shift of the first formant to an adjacent vowel category, the typical talker will change their production to partially compensate for the error. This compensation varies significantly from individual to individual with some talkers fully compensating for the perturbation and others not compensating at all.

The reasons for this variability in response are not known. One possibility is that during the sensorimotor integration that is part of speech planning and control, different individuals might place different weights on the sensory information coming from the auditory, tactile and proprioceptive sensory systems. If a contradiction exists between an induced error in auditory feedback and the expected tactile and proprioceptive feedback, the incongruence would elicit different production responses depending on the sensory feedback weights. Additionally, the weight of auditory feedback may vary across the English vowel space according to the strength of non-auditory cues. A working hypothesis could then be that point vowels like /i/ rely less on auditory feedback because robust tactile feedback is available.

One way to begin to test this hypothesis is to measure the minimum acoustic perturbation required to elicit a compensatory response and the maximum compensatory change for different vowels. Experiments where auditory feedback is manipulated have found that production does not change in response to small perturbations [5, 3]. There appears to be a minimum manipulation that is required to elicit compensatory changes in production. Here we test auditory feedback compensations in F1 for vowels across the vowel space.

2 Methods

The first formant was gradually manipulated in the auditory feedback presented over headphones with a real-time filtering method. Small F1 feedback modifications (4 Hz) were introduced between trials

to a maximum absolute change of either plus or minus 200 Hz. A between-participant design was employed where each vowel and manipulation used a different group of individuals. Data collection continues, but at present there are 10 people per condition, with the exception of vowel /ɛ/ where there are 20. Vowels that have been manipulated in F1 by +200 Hz are /i/, /ɛ/, /ɑ/, and /u/, and by -200 Hz are /i/ and /ɑ/.

A brief pure-tone audiometric hearing assessment was performed for both ears at octave frequencies between 500 and 4000 Hz. The majority of individuals had hearing thresholds below 20 dB HL with the exception of a few with 5 to 10 dB elevation at a single frequency. Participants were seated in a sound attenuated booth in front of a video display where single word prompts were shown. Sennheiser HD 265 headphones were used to present the filtered voice at a level of approximately 80 dBA SPL with background speech shaped noise of 50 dBA SPL. A Shure WH20 headset microphone was worn to measure the speech and deliver it to a formant filtering system based on National Instruments real-time hardware and custom software [3].

Participants first practiced saying /hVd/ words containing seven different vowels from across the English vowel space while the microphone gain was optimized by the operator. Subsequently in random order, six tokens were recorded for each vowel and the best model order for an iterative linear predictive coding (LPC) formant tracker was determined for the test vowel. The participant was then prompted to say the word containing the test vowel a total of 110 times which required approximately six minutes. The first 20 utterances were used to acclimatize the participant to the headphones and pace, and were not analyzed further. They are not shown or included in the data given in the results section. The next 20 trials (referred to as 1 to 20 below) were used to obtain a baseline of speech production prior to the manipulation. For the next 50 trials (labeled 21 to 70) the auditory feedback for the F1 component of the talker's speech was modified in steps of 4 Hz between utterances. The final 20 trials (labeled 71 to 90) and referred to as the hold phase, had the maximum manipulation of auditory feedback (+ or – 200 Hz).

The vocalic portion of each utterance was segmented using a semi-automated procedure and this vowel portion was used in subsequent analyses. From each recorded token, formant values from the middle 80% of the vowel were averaged to obtain a single estimate of F1 and F2 per utterance. Two main types of analyses were performed with this group average data. The difference between the mean of the baseline tokens (trials 1 to 20) and the mean of the "hold" tokens (trials 71 to 90) was represented in terms of percent compensation. Compensation is the difference between the mean hold and baseline trials divided by the maximum manipulation of 200 Hz, and is expressed as a percentage. In addition to percent compensation, values are given for the threshold manipulation size at which a statistically significant change was detected in the production of the first formant using the change point test.

3 Results

Figures 1 and 2 show the average F1 produced for each manipulation direction. The curves have been normalized by subtracting the average baseline F1 and smoothed using a filter 10 trials wide. Due to smoothing end effects, values are not given for trials 1 to 5 and 85 to 90. Average baseline and hold values for F1 of each condition are given in Table 1. Small changes observed in F2 for some vowels are not presented or discussed here due to space constraints. The column labeled "Change" gives the mean difference between the hold and baseline phases across participants. The change was statistically significant at $p<0.05$ using paired t-tests for all conditions except /u/+ and /i/- (the sign denotes the feedback manipulation direction). A univariate analysis of variance (ANOVA) found a significant difference between vowel conditions for the magnitude of compensation [$F(5,64)=13.1$, $p<0.001$]. Post-hoc testing using Tamhane's T2, where equal variances are not assumed, found a significant difference between /ɛ/+ and the conditions /u/+, /i/+, and /i/- ($p<0.001$), where the compensation was larger for /ɛ/. There was no statistical difference between /ɛ/+ and /ɑ/+ or /ɑ/-.

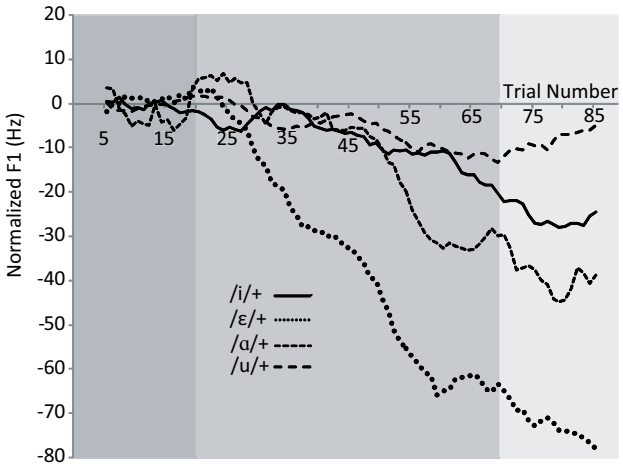

Figure 1. Response to a positive shift of F1. Gray bands indicate the experiment phases.

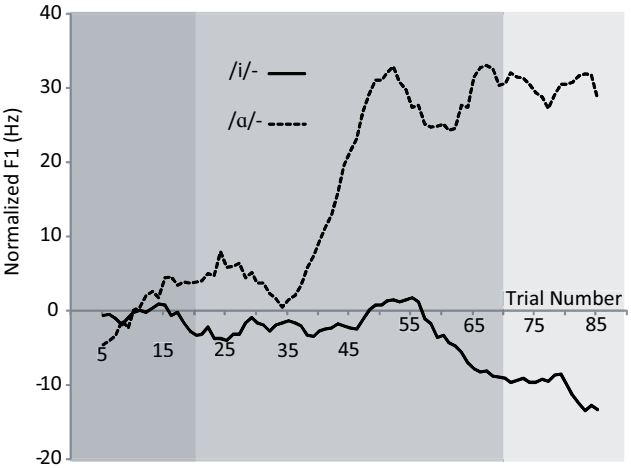

Figure 2. Response to a negative shift of F1.

Thresholds given in Table 1 are the mean of the change points calculated from individual participants' data. For all individuals, a statistically significant change point was found. An ANOVA showed a significant difference between vowel conditions for threshold [$F(5,64)=2.9$, $p<0.05$]. Post-hoc testing using Tamhane's T2 found a significant difference between the thresholds for /ɛ/+ and /i/+, where /ɛ/ had a lower threshold ($p<0.05$).

Table 1: F1 Results for group averages. Standard deviations (s.d.) are given in parentheses. For the baseline and hold phases, two values of s.d. are given. Average trials for each phase were computed across the group, and the s.d. of these average trials is listed first. The s.d. of each individual's tokens in each phase was also computed, and the mean of these s.d. is listed second. * indicates a statistically significant change between hold and baseline phases using a paired t-test ($p<0.05$). Gray shading indicates conditions where F1 was lowered in frequency.

						Threshold	
Vowel	Stimulus Manip. (Hz)	Baseline (Hz)	Hold (Hz)	Change (Hz)	% Comp.	Trial #	Shift Size (Hz)
/i/	200	352 (7, 19)	327 (8, 24)	-26* (15)	13	59 (10)	156
/ɛ/	200	699 (9, 35)	623 (8, 39)	-76* (35)	38	44 (8)	96
/ɑ/	200	818 (17, 43)	780 (15, 38)	-38* (37)	19	54 (16)	136
/u/	200	413 (3, 11)	405 (6, 15)	-7 (16)	4	43 (16)	92
/i/	-200	334 (5, 18)	322 (6, 17)	-12 (32)	-6	50 (15)	-120
/ɑ/	-200	811 (10, 26)	842 (13, 34)	31* (38)	16	49 (8)	-116

4 Discussion

Talkers' sensitivity to perturbations of auditory feedback is not constant across the vowel space. Both the compensatory thresholds and magnitudes of maximum compensation for perturbations varied with vowel quality. For an increase in F1, the vowels /i/, /ɛ/, and /u/ have neighbouring vowel category boundaries that would likely be crossed in the F1 dimension by a manipulation of +200 Hz. If the perceptual categorization of vowels was driving compensatory behaviour, the auditory-vocal feedback system would produce strong corrective changes in speech production in response to such an error. However, this cannot be the sole basis of the explanation for the varying compensatory behaviour. Compensatory behaviour is initiated following small perturbations that are perceptually still the target vowel. Also, differences between vowels exist even for vowels that don't have neighbouring vowels in one direction. The vowels /i/+ and /ɛ/+ had statistically significant compensatory changes in production of the first formant, where the change in /ɛ/ was the largest (38% or -76 Hz). Of the vowels with statistically different hold and baseline phases, production of /ɛ/ also changed for the smallest manipulation of 96 Hz. However, the change points were only statistically significantly different between /ɛ/+ and /i/+. For vowel /i/, the compensation was larger than /u/, but still small at 13% (-26 Hz) with a high threshold of 156 Hz. A large compensatory change in production of /i/ or /u/ would require hyper-articulation, whereas /ɛ/ is closer to the middle of the F1 dimension.

For /ɑ/+, the manipulation produces a first formant outside the normal vowel space. The compensation of 19% (-38 Hz) is second in magnitude only to /ɛ/, although they did not differ statistically in conservative post-hoc testing. The compensatory response to perturbations in a frequency direction where there is no neighbouring vowel category suggests that the response is initiated by deviation from a target region defined independently for the vowel.

For /i/, the manipulation of F1 in a negative direction also produces a first formant that is outside the normal English vowel space. However, the formant filtering system may not work optimally with this vowel and manipulation direction due to the proximity of the first formant and fundamental. The average fundamental for this group of participants was 206 Hz and the average first formant was 334 Hz for the baseline trials. Since the larger manipulations would place a pair of spectral poles below the fundamental where there is no harmonic energy to emphasize, it is unlikely that the auditory-vocal feedback system would interpret the filtered voice as having an error as large as intended. The skirts of the filter would raise the level of the first harmonic F0, but the result may not be as effective as for other conditions. The response to this manipulation approached zero, and the baseline and hold trials were not statistically different. This may be in part due to the limitations of the formant filtering method.

The vowel /ɑ/ has neighbouring vowel category boundaries that would likely be crossed in the F1 dimension by a manipulation of -200 Hz. For this condition, the compensation in production was 16% (31 Hz) with a threshold of -116 Hz. As for /i/ or /u/ shifted in the positive direction, a large compensatory change in production of /ɑ/ would require hyper-articulation.

With the exception of /ɛ/, all the tested vowels are at extremes of the English F1 x F2 vowel space. The response for /ɛ/ was the largest and had the lowest numerical threshold for vowels with significantly different baseline and hold phases. Being away from the extremes of articulation, /ɛ/ does not benefit from a saturation effect [6, 7] and may therefore produce weaker kinesthetic and tactile feedback. Auditory feedback may therefore be weighted more heavily, which would explain the robust response to altered auditory feedback observed here. For point vowels, the availability of stronger non-auditory feedback may lead to a lower relative weight being applied to auditory feedback. The contradiction amongst different modalities may then elicit a smaller response.

Data collection is ongoing to complete the negative manipulation for /u/ and /ɛ/, as well as to add the point vowel /æ/ and the high front vowel /ɪ/.

References

[1] Burnett, T. A., Freedland, M. B., Larson, C. R., and Hain, T. C. (**1998**). "Voice F0 responses to manipulations in pitch feedback," J. Acoust. Soc. Am. **103**, 3153–3161.

[2] Houde, J. F. and Jordan, M. I. (**2002**). "Sensorimotor adaptation of speech I:Compensation and adaptation," J. Speech Lang. Hear. Res. **45**, 295–310.

[3] Purcell, D. W. and Munhall, K. (**2006**). "Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation," J. Acoust. Soc. Am. **120**, 966-977.

[4] Villacorta, V. M., Perkell, J. S., and Guenther, F. H. (**2007**). "Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception," J. Acoust. Soc. Am. **122**, 2306-2319.

[5] Jones, J. A. and Munhall, K. G. (**2000**). "Perceptual calibration of F0 production: evidence from feedback perturbation," J. Acoust. Soc. Am. **108**, 1246–1251.

[6] Svirsky, M. A. and Tobey, E. A. (**1991**). "Effect of different types of auditory stimulation on vowel formant frequencies in multichannel cochlear implant users," J. Acoust. Soc. Am. **89**, 2895–2904.

[7] Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Perrier, P., Vick, J., Wilhelms-Tricarico, R., and Zandipour, M. (**2000**). "A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss," J. Phonetics **28**, 233–272.