Perceptual Calibration to Modest, Predictable Spectral Peaks in Precursor Sounds Influences Vowel Identification

Paul Anderson and Christian Stilp
Department of Psychological and Brain Sciences, University of Louisville

PS-151

INTRODUCTION

Perceptual systems strive to be maximally sensitive to changes in the sensory environment. In audition, when a spectral property is predictable across a preceding acoustic context and subsequent vowel target, perception deemphasizes this cue and increases its reliance on other changing (thus more informative) cues for speech recognition. This process, known as auditory perceptual calibration, has been demonstrated for both spectrally local (second formant frequency, F_2) and global (overall spectral tilt) cues to vowel identity (Kiefte & Kluender, 2008; Alexander & Kluender, 2010). When acoustic energy in the vowel's F_2 region is made predictable throughout the preceding context, listeners attribute less weight to F_2 and more weight to spectral tilt to identify the target vowel, and *vice versa*.

However, conditions that promote perceptual calibration to predictable spectral peaks are poorly understood. In previous experiments, acoustic energy at the vowel's F_2 was made predictable by amplifying this frequency region by 20 dB. Three lines of research suggest these predictable spectral peaks need not be so dramatic in order to induce perceptual calibration:

- Spectral contrast (*i.e.*, amplitude differences between spectral peaks and neighboring valleys) reach 25-30 dB in some vowels but only 5-7 dB in others (Fant, 1973).
- Normal-hearing listeners require only 1-2 dB of spectral contrast to identify vowel sounds (*e.g.*, Leek *et al.*, 1987).
- Experienced listeners can detect intensity increments of only 1 dB for one tone in a multitone complex (Green, 1988).

With exquisite sensitivity to small increments in intensity and extensive experience with modest levels of spectral contrast in speech, research suggests that perceptual calibration will maintain for more modest but still reliable spectral peaks. The present experiment tested this prediction by reducing filter gain for the predictable spectral peak in the preceding acoustic context.

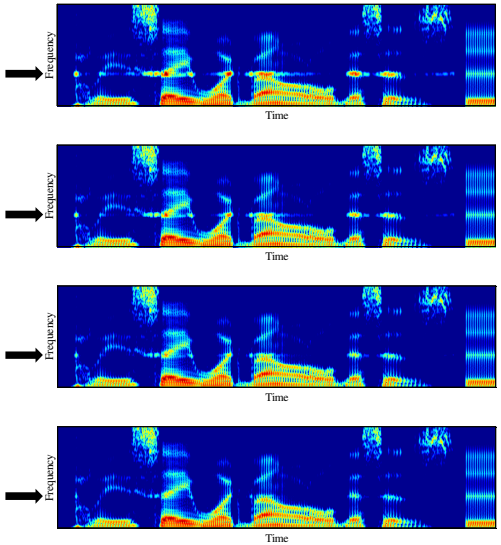


FIGURE 1. Spectrograms of sample trials. Arrows indicate filter center frequency of 1600 Hz, resulting in +20 dB (top row, for reference), +15 dB (second row), +10 dB (third row), or +5 dB amplification (bottom row).

STIMULI

Vowels

5-by-5 vowel matrix perceptually varying from [i] (as in "beet") to [u] (as in "boot")

- spectral tilt varied from -12 to 0 dB/octave Hz in 3 dB/octave steps
- F_2 was orthogonally varied from 1000 to 2200 Hz in 300-Hz steps
- 90 ms duration
- same stimuli as used in Alexander & Kluender (2010)

Precursor

"Please say what vowel this is" produced by AT&T Natural Voices® Text-to-Speech Demo

- http://www2.research.att.com/~ttsweb/tts/demo.php
- Voice and Language: Mike, US English
- 1759 ms duration

Filtering

Precursor was processed by one of five bandpass filters centered at one F_2 frequency used in the vowel matrix (1000, 1300, 1600, 1900, 2200 Hz)

- Filter bandwidth = center frequency ± 50 Hz
- 1200 coefficients using fir2 function in MATLAB
- Amplification set to +5, +10, or +15 dB
  - Acoustic energy in this frequency region is not constant across the entire context, but still waxes and wanes as in unprocessed speech (see figure)

METHODS

Participants

23 undergraduates from University of Louisville

- All native English speakers who reported normal hearing

Procedure

Listeners responded by clicking the mouse to indicate whether the target vowel sounded more like "ee" or "oo"; no feedback was provided

- Stimuli presented diotically over circumaural headphones at 70 dB SPL
- Block 1: vowels presented in isolation
- Blocks 2-4: filtered precursor, 50-ms ISI, then target vowel
  - Amplified frequency region in precursor matched vowel F_2
  - Extent of amplification (+5, +10, +15 dB) blocked and tested in random order

Data

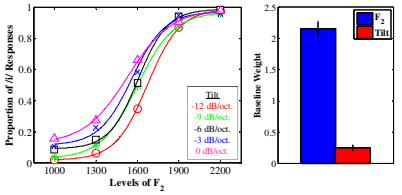
Perceptual weights for F_2 and tilt were estimated using standardized logistic regression coefficients

- Wilcox's (2005) Minimum Generalized Variance method used to remove all data for listeners whose weights for vowels presented in isolation were outliers (n = 2)
- Perceptual calibration was measured as changes in regression coefficients for vowels presented in isolation versus following a filtered precursor

RESULTS

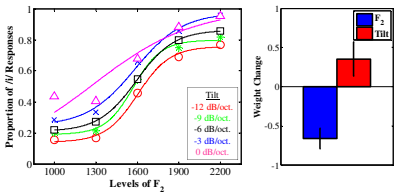
Figures display maximum likelihood psychometric functions fit to listeners' responses (Wichmann & Hill, 2001). Weight changes were analyzed using one-tailed *t*-tests against zero, as F_2 weights were predicted to decrease while tilt weights were predicted to increase across all filter gain conditions.

Practice



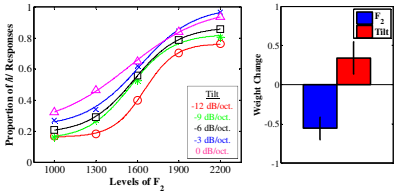
Practice weights reveal heavy reliance on F_2 for vowel identification, consistent with previous research. This reliance is evident in steep slopes of psychometric functions and small leftward shifts as spectral tilt increases.

+15 dB Filters



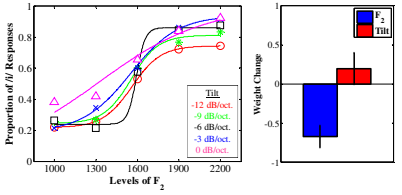
F_2 weights decreased (mean = –0.66, $t_{20} = 5.09$, $p < .0001$)
Tilt weights modestly increased (mean = +0.35, $t_{20} = 1.63$, $p = .059$)

+10 dB Filters



F_2 weights decreased (mean = –0.56, $t_{20} = 3.89$, $p < .001$)
Tilt weights modestly increased (mean = +0.34, $t_{20} = 1.65$, $p = .057$)

+5 dB Filters



F_2 weights decreased (mean = –0.67, $t_{20} = 4.75$, $p < .0001$)
Tilt weights did not significantly increase (mean = +0.19, $t_{20} = 0.96$, $p = .175$)

DISCUSSION

Auditory perception calibrates to a reliable spectral peak in an acoustic context, but demonstrations thus far have presented overwhelmingly strong evidence for this peak by using +20 dB filters. The present results reveal that reliable spectral properties need not be particularly robust to elicit perceptual calibration, as amplifying frequency regions in precursors by as little as +5 dB influenced vowel identification. Results shed light on underlying principles supporting auditory perceptual calibration and how perception attunes to predictable properties in the environment most broadly.

Results raise questions regarding the minimum spectral contrast (filter amplitude) necessary in order to elicit perceptual calibration. A follow-up study was conducted with new listeners and the same experimental design but using +7, +4, and +2 dB filters. All weight changes were in the predicted directions (range of mean F_2 weight changes: -0.09 to -0.19; range of mean tilt weight changes: +0.10 to +0.15), but none significantly differed from zero. One possible reason why +5 dB filters induced perceptual calibration but +7 dB filters did not was enhanced sensitivity to modest spectral peaks due to experience with much larger peaks (+15, +10 dB). Further research is needed to test this prediction.

Low thresholds in profile analysis and spectral contrast detection may predict perceptual calibration will maintain down to very modest spectral peaks, but several significant differences between methodologies bear mention. In studies of profile analysis and spectral contrast, detection is based on short-time sampling of an intensity increment known to be present, often judged against a flat-spectrum background of equal-amplitude components. Perceptual calibration involves accumulation of intermittent evidence for a spectral peak in complex and rapidly changing spectral shapes. No explicit instruction is given to participants to detect this increment; listeners are only asked to identify the following vowel sound. All three lines of research converge on considerable sensitivity to modest spectral peaks or increments, but under very different circumstances.

Greater amounts of spectral contrast are needed for speech perception by hearing-impaired listeners (Leek *et al.*, 1987; Summers & Leek, 1994; Dreisbach *et al.*, 2005) and cochlear implant users (Loizou & Poroy, 2001). To date, perceptual calibration has not been explored with hearing-impaired listeners. The present results are encouraging in that reliable spectral peaks may not need be excessively large (> 20 dB) to induce perceptual calibration in these listeners.

REFERENCES

Alexander, J.M., & Kluender, K.R. (2010). *JASA*, *128(6)*, 3597-3613.
Dreisbach, L.E., Leek, M.R., & Lentz, J.J. (2005). *JSLHR*, *48*, 910-921.
Fant, G. (1973). *Speech Sounds and Features* (MIT Press: Cambridge, MA).
Green, D.M. (1988). *Profile Analysis* (Oxford University Press: New York).
Kiefte, M., & Kluender, K.R. (2008). *JASA*, *123(1)*, 366-376.
Leek, M.R., Dorman, M.F., & Summerfield, Q. (1987). *JASA*, *81(1)*, 148-154.
Loizou, P.C., & Poroy, O. (2001). *JASA*, *110(3)*, 1619-1627.
Summers, V., & Leek, M.R. (1994). *JASA*, *95(6)*, 3518-3528.
Wichmann, F.A., & Hill, N.J. (2001). *Percept. Psychophys*, *63(8)*, 1293-1313.
Wilcox, R.R. (2005). *Introduction to Robust Estimation and Hypothesis Testing*, 2nd ed. (Elsevier Academic Press, London), pp. 228–231.