Modest, reliable spectral peaks in preceding sounds influence vowel perception

Christian E. Stilp[a] **and Paul W. Anderson**
Department of Psychological and Brain Sciences, University of Louisville,
Louisville, Kentucky 40292
christian.stilp@louisville.edu, paul.anderson@louisville.edu

Abstract: When a spectral property is reliable across an acoustic context and subsequent vowel target, perception deemphasizes this cue and shifts toward less predictable, more informative cues. This phenomenon (auditory perceptual calibration) has been demonstrated for reliable spectral peaks +20 dB or larger, but psychoacoustic findings predict sensitivity to more modest spectral peaks. Listeners identified vowel targets following a sentence with a reliable +2 to +15 dB spectral peak centered at F_2 of the vowel. Vowel identifications weighted F_2 significantly less when reliable peaks were at least +5 dB. Results demonstrate high sensitivity to reliable acoustic properties in the sensory environment.

© 2014 Acoustical Society of America

PACS numbers: 43.66.Ba, 43.71.An, 43.71.Es [JH]
Date Received: June 20, 2014 **Date Accepted:** September 17, 2014

1. Introduction

Perceptual systems strive to be maximally sensitive to changes in the sensory environment. When sensory inputs are unchanging or predictable, perception decreases its reliance on this input and increases its reliance on changing and thus informative stimulus properties. Neural adaptation is an elegant example of this principle. When stimuli are constant or predictable, neural firing decreases precipitously or ceases altogether to reflect the unchanging nature of the stimulus. This conserves neural resources (i.e., action potentials) for when stimuli change. Thus, neural adaptation maximizes transmission of information for the organism (Wainwright, 1999; Clifford *et al.*, 2007; Kohn, 2007).

Sensitivity to predictable versus unpredictable stimulus properties scales up to higher-level perception as well. In audition, when a spectral property is reliable (i.e., relatively stable or recurring in the acoustic spectrum, such as a particular spectral shape or local frequency peak) across a preceding acoustic context and subsequent vowel target, perception deemphasizes this cue and shifts toward 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 and Kluender, 2008; Alexander and Kluender, 2010). When acoustic energy in the vowel's F_2 region is made reliable throughout the preceding acoustic context, listeners attribute less weight to F_2 and more weight to spectral tilt in vowel identification, and vice versa. These results extend earlier work by Darwin *et al.* (1989), who reported perceptual compensation for filter properties when a carrier sentence and target word were processed by the same filter.

Demonstrations of perceptual calibration thus far have presented strong evidence for the presence of a reliable spectral property (e.g., large changes in spectral slope, high amplitudes for spectral peaks). In Kiefte and Kluender (2008) and

[a]Author to whom correspondence should be addressed.

Alexander and Kluender (2010), acoustic energy at the vowel's F_2 was made reliable in the precursor sound by amplifying this frequency region by 20 dB or more. Assmann and Summerfield (2004) suggested that perceptual compensation reported by Darwin *et al.* (1989) may occur only under extreme conditions, since it was only present for 30-dB changes in spectral slope. Perceptual sensitivity to more modest but still reliable signal properties is unknown.

Results from three separate lines of research suggest that reliable spectral peaks need not be so extreme in order to induce perceptual calibration. First, listeners have extensive experience perceiving speech with less pronounced spectral peaks. Fant (1973) measured formant frequencies and amplitudes in Swedish vowels, reporting spectral contrast levels (i.e., amplitude differences between spectral peaks and neighboring valleys) in back vowels as little as 5 dB. In favorable listening conditions, such spectral peaks provide more than sufficient spectral contrast for speech recognition (Leek *et al.*, 1987), but additional factors such as simultaneous masking of formants merit consideration (see Kiefte *et al.*, 2010, and references therein).

Second, the amount of spectral contrast required to perceive an otherwise flat-spectrum stimulus as a vowel sound is substantially less than 20 dB. Leek *et al.* (1987) presented listeners composites of logarithmically spaced sine waves. All sine waves were equal-amplitude except for six (three pairs of two adjacent components), located at frequencies corresponding to vowel formants. They measured the minimum intensity increment for these six sine waves that still produced robust vowel percepts. Normal-hearing listeners required only 1–2 dB of spectral contrast to identify vowel sounds (see also Lea and Summerfield, 1994; Alcantara and Moore, 1995). Related studies examining enhancement effects (Summerfield *et al.*, 1987) and detection of spectral troughs (Turner and Van Tasell, 1984) also produced high levels of performance with only 1–2 dB of spectral contrast.

Third, a long line of experiments in profile analysis revealed exquisite sensitivity to increments in intensity of a single component in complex sounds. In these experiments, listeners were presented sine wave composite stimuli where the intensity of one component was manipulated. Listeners compared the incremented stimulus to a standard without an increment. Experienced listeners detected intensity increments of only 1 dB for one tone in a multitone complex (Green, 1988).

Extensive experience hearing modest levels of spectral contrast in speech and great sensitivity to small increments in intensity suggest that perceptual calibration will maintain for more modest (i.e., lower-amplitude) but still reliable (i.e., across time) spectral peaks. The present experiments tested this prediction by reducing gain in the bandpass filter that added a reliable spectral peak to the preceding acoustic context (sentence). Strength of perceptual calibration was measured through the perceptual weights attributed to the reliable (F_2) and unreliable (spectral tilt) cues used for identifying the subsequent target vowel.

2. Methods

2.1 Participants

Sixty-eight undergraduates from University of Louisville participated in exchange for course credit. All were native English speakers who reported normal hearing. Each participated in only one experiment (23 in experiment 1; 23 in experiment 2; 22 in experiment 3).

2.2 Stimuli

2.2.1 Vowels

Target vowels were the same stimuli from Alexander and Kluender (2010), perceptually varying from [i] (as in "beet") to [u] (as in "boot"). Vowels were synthesized using the parallel branch of the Klatt and Klatt (1990) synthesizer at a sampling rate of 22 500 Hz. Vowels had a fundamental frequency of 100 Hz and were 90 ms in duration with 5-ms onset/offset ramps. First, a series of five vowels varying from [u] to [i] was created by varying F_2 from 1000 to 2200 Hz in 300-Hz steps. F_2 was synthesized with 160-Hz bandwidth. F_1 (300 Hz center frequency with 60-Hz bandwidth), F_3 (2700 Hz with 260-Hz bandwidth), and F_4 (3600 Hz with 360-Hz bandwidth) were held constant. Formant amplitudes were manipulated so that each vowel had a reasonably constant spectral tilt of −3 dB/octave as measured by linear regression slope (formant amplitude as a function of log frequency) across all formants as well as between pairs of neighboring formants.

Spectral tilt was systematically manipulated using 90-tap finite impulse response filters in MATLAB (Mathworks Inc., Natick, MA). Between 212 and 4800 Hz, filter gain changed linearly as a function of log frequency. As the series of vowels described above had native spectral tilt of −3 dB/octave, tilt of the filter response varied from −9 to +3 dB/octave in +3 dB/octave steps to achieve final spectral tilts of −12, −9, −6, −3, and 0 dB/octave. Spectral amplitudes below 212 Hz were unmodified by filtering, and all spectral amplitudes above 4800 Hz adopted the amount of gain calculated at 4800 Hz. These five levels of spectral tilt were imposed on each of the five levels of F_2, creating the 5×5 vowel matrix. Vowel targets were upsampled to 48 828 Hz then low-pass filtered with an 86-tap finite impulse response filter with an upper cutoff at 4800 Hz and stopband of −90 dB at 6400 Hz.

2.2.2 Precursor

The precursor was "Please say what vowel this is" produced by AT&T Natural Voices^{TM} Text-to-Speech Synthesizer (Beutnagel *et al.*, 1997). The male talker ("Mike") had an American English accent. Precursor duration was 1759 ms. The precursor was processed by a 100-Hz-wide bandpass filter centered at one of the F_2 frequencies used in the vowel matrix (1000, 1300, 1600, 1900, and 2200 Hz) (Fig. 1). The only specified filter frequencies were the center and corner frequencies (i.e., center frequency ±50 Hz). Filter gain was set to +2, +4, +5, +6, +7, +7.5, +9, +10, or +15 dB at the center frequency and zero elsewhere with no amplification in the stopband. Filters were derived in the fir2 function using MATLAB with 1200 coefficients, a 1201-point Hamming window, and a 600-point grid onto which the desired frequency response was interpolated. Following filtering, target vowels were appended to filtered precursors sharing the same spectral peak (F_2 in the vowel) with a 50-ms silent interstimulus interval.

2.3 Procedure

All stimuli were resampled at 44 100 Hz sampling rate and presented diotically at 70 dB sound pressure level via circumaural headphones (Beyer-Dynamic DT-150, Beyerdynamic Inc. USA, Farmingdale, NY). Listeners responded by clicking the mouse to indicate whether the target vowel sounded more like "ee" or "oo." No feedback was provided. Listeners participated individually in single-wall sound-isolating booths (Acoustic Systems, Inc., Austin, TX). Following acquisition of informed consent, participants first completed a block of 200 trials (25 vowels × 8 repetitions) where vowels were presented in isolation. Listeners then completed three short sub-conditions where vowels were preceded by the filtered precursor sentence with distinct but overlapping ranges of filter gain. Passband filter gains were +5/+10/+15 dB (experiment 1; $n = 23$), +2/+4/+7 dB (experiment 2; $n = 23$), or +6/+7.5/+9 dB (experiment 3; $n = 22$). Conditions were blocked and tested in random orders, with 200 trials per block (25 vowels each paired with its F_2-matched precursor × 8 repetitions). Participants took short breaks between each experiment. The entire session lasted approximately 1 h.

3. Results

Perceptual weights for F_2 and tilt were estimated using standardized logistic regression coefficients. Weights were calculated separately for vowels presented in isolation and vowels following filtered precursors. Perceptual calibration was operationalized as

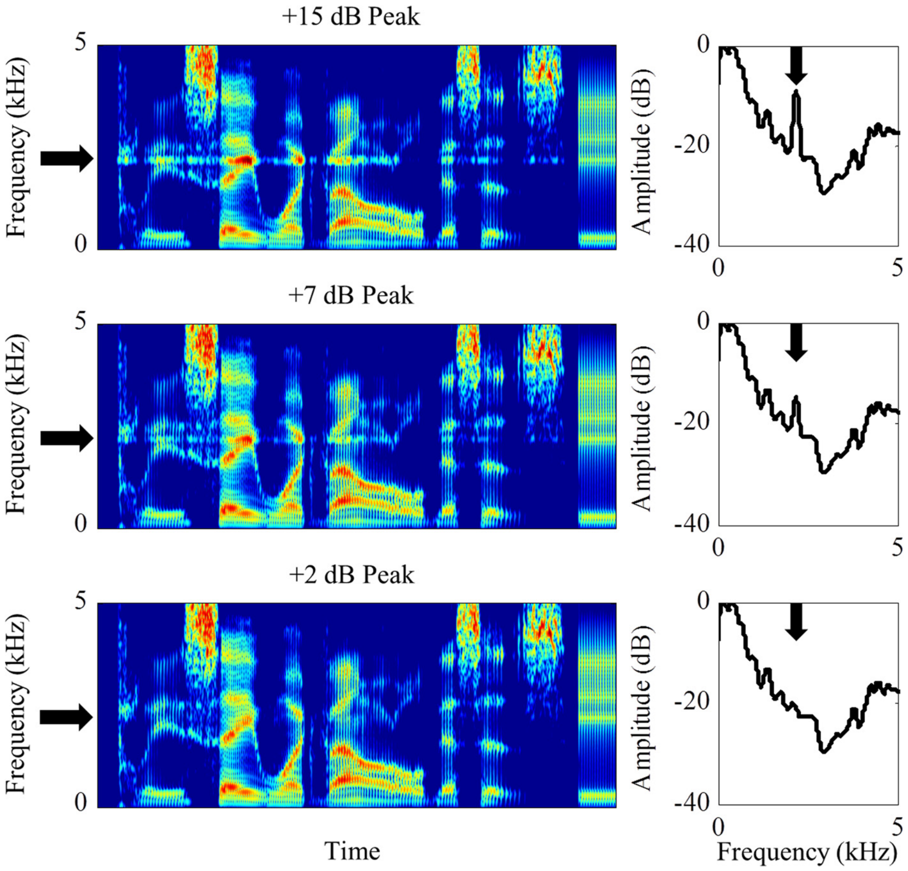

Fig. 1. (Color online) Top row depicts the +15 dB condition, middle row depicts +7 dB condition, bottom row depicts +2 dB condition. The left column depicts spectrograms of trials with the filtered precursor followed by a vowel with matching F_2 (2200 Hz) and spectral tilt of -6 dB/octave. The right column depicts the long-term average spectrum of the precursor. Arrows indicate varying strength of evidence for the presence of a reliable spectral peak at 2200 Hz.

changes in weights across the two sessions (i.e., perceptual adjustment in response to the precursor possessing a reliable spectral peak). Wilcox's (2005) Minimum Generalized Variance method was used to remove all data for listeners whose weights for vowels presented in isolation were outliers, which complicates interpreting measures of perceptual calibration ($n = 2$ in experiment 1; no outliers in experiment 2; $n = 3$ in experiment 3). On the basis of previous experiments (Kiefte and Kluender, 2008; Alexander and Kluender, 2010), and if perceptual calibration effects extend to more modest amplitude manipulations, F_2 weights were predicted to decrease while tilt weights were predicted to increase for all conditions. Significant weight changes from baseline (i.e., weights for vowels presented in isolation) were of primary interest. Given the consistent directionality of these predictions (negative weight changes for F_2, positive weight changes for tilt), weight changes were analyzed using one-tailed t-tests against zero. As each participant group contributed six measures (F_2 weight change and tilt weight change in each of three conditions of filter gain), t-tests were Bonferroni-corrected using $\alpha = 0.05/6 = 0.0083$.

Weight changes for each listener are presented in Fig. 2(a). In Experiment 1 (top row), F_2 weights significantly decreased following +15 dB (mean weight change $= -0.66$, one-tailed t-test against zero: $t_{20} = 5.09$, $p < 0.0001$), +10 dB (mean $= -0.56$, $t_{20} = 3.89$, $p < 0.001$), and +5 dB spectral peaks in the precursor (mean $= -0.67$,

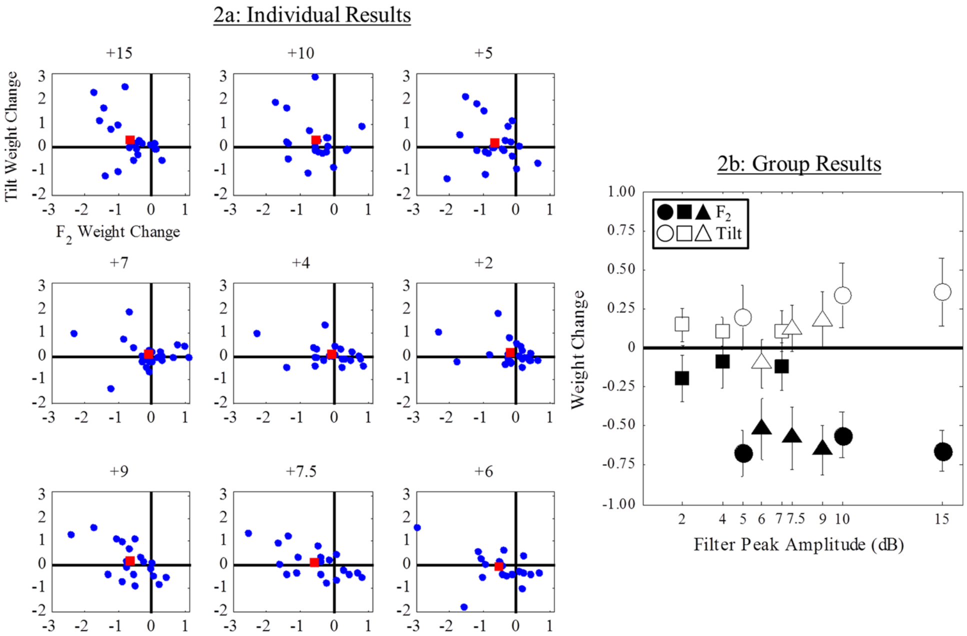

Fig. 2. (Color online) (a) Each circle represents one participant's weight changes for spectral tilt (ordinate) and F_2 (abscissa) from vowels presented in isolation to vowels following filtered precursors (one participant's weights at [1.54 0.00] in +7 dB condition not shown). Squares indicate group means. Negative weight changes on the abscissa (decreased F_2 weight) and positive weight changes on the ordinate (increased tilt weight) both reflect perceptual calibration to a reliable spectral peak. Results are arranged by participant group (experiment 1 = top row, experiment 2 = middle row, experiment 3 = bottom row). (b) Mean weight changes for each listener group. Filled shapes represent weight changes for F_2 and unfilled shapes represent weight changes for tilt. Circles indicate results from experiment 1, squares indicate results from experiment 2, and triangles indicate results from experiment 3. Error bars depict ±1 standard error.

$t_{20} = 4.75$, $p < 0.0001$). Tilt weights increased nominally but weight changes did not significantly differ from zero (+15 dB: mean = +0.35, $t_{20} = 1.63$, $p = 0.06$; +10 dB: mean = +0.34, $t_{20} = 1.65$, $p = 0.06$; +5 dB: mean = +0.19, $t_{20} = 0.96$, $p = 0.18$). In Experiment 2 (middle row), F_2 weight decreases did not significantly differ from zero (+7 dB: mean = −0.12, $t_{22} = 0.80$, $p = 0.22$; +4 dB: mean = −0.09, $t_{22} = 0.55$, $p = 0.30$; +2 dB: mean = −0.19, $t_{22} = 1.32$, $p = 0.10$). Tilt weights increased by small, nonsignificant amounts (+7 dB: mean = +0.11, $t_{22} = 0.86$, $p = 0.20$; +4 dB: mean = +0.10, $t_{22} = 1.14$, $p = 0.13$; +2 dB: mean = +0.15, $t_{22} = 1.30$, $p = 0.10$). In Experiment 3 (bottom row), F_2 weights significantly decreased following reliable +9 dB (mean = −0.65, $t_{18} = 4.16$, $p < 0.001$), +7.5 dB (mean = −0.58, $t_{18} = 2.82$, $p < 0.006$), and +6 dB peaks in the precursor spectrum (mean = −0.52, $t_{18} = 2.69$, $p < 0.008$). Tilt weights again did not significantly differ from 0 (+9 dB: mean = +0.18, $t_{18} = 1.00$, $p = 0.17$; +7.5 dB: mean = +0.12, $t_{18} = 0.83$, $p = 0.21$; +6 dB: mean = −0.10, $t_{18} = 0.63$, $p = 0.73$). Mean weight changes are plotted as a function of filter gain in Fig. 2(b).

4. Discussion

Auditory perception calibrates to a reliable spectral peak in an acoustic context, but demonstrations thus far only examined prominent peaks introduced by high-gain filters. The present results reveal that reliable spectral properties need not be particularly pronounced to elicit perceptual calibration. Introducing a spectral peak as modest as +5 dB in a precursor sentence influenced identification of the following vowel. Results reveal remarkable perceptual sensitivity to reliable acoustic properties in the sensory environment.

Low thresholds in profile analysis and spectral contrast detection tasks predicted perceptual calibration will maintain for very modest spectral peaks. However, 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. While participants are instructed to listen for the intensity increment in profile analysis and spectral contrast detection, no such explicit instruction was given to participants in the present experiments; they were only asked to identify the following vowel sound. Finally, profile analysis and spectral contrast investigations may include extensive training to establish sensory thresholds, but no such training component was included here (aside from the listener's prodigious experience perceiving speech). All three lines of research converge on considerable sensitivity to modest spectral peaks or increments, but under very different circumstances.

Reliable spectral peaks consistently produced significant increases in tilt weights in Alexander and Kluender (2010), but the same was not observed here despite using the same target vowels. Two differences in experimental design might explain this discrepancy. First, Alexander and Kluender (2010) tested a filter gain of +24.5 dB, reporting larger weight changes (F_2 mean change = −0.77, tilt mean change = +0.44 for 2000-ms precursors, best approximating duration of the precursor presented here) than those following +15 dB filter gain (F_2 mean change = −0.66, tilt mean = +0.35). Second, studies differed widely in selection of precursor. Alexander and Kluender (2010) presented a nonspeech precursor where four narrowband (formant-like) filters sampled a harmonic source. Filters were sinusoidally frequency-modulated across time, resulting in continuously varying spectral peaks. Reliable spectral peaks occurred semi-regularly throughout the duration of the precursor (see their Figs. 2 and 10). Contrast this with the speech precursor shown in Fig. 1, where evidence for the spectral peak is highly variable and particularly weak toward the end of the sentence. In addition to relative amplitude, sampling across time might also define the reliability of a given spectral property. Additional research is needed to understand the relationship between calibration and specific spectrotemporal characteristics of precursor sounds.

Hearing-impaired (HI) listeners display elevated spectral contrast thresholds compared to normal-hearing (NH) listeners (Leek *et al.*, 1987; Summers and Leek, 1994; Leek and Summers, 1996; Dreisbach *et al.*, 2005). For a spectral peak at 2000 Hz (which is appropriate for F_2 in [i] as tested here), HI listeners exhibited contrast thresholds of approximately 11 dB, much larger than that for NH listeners (2–5 dB) (Dreisbach *et al.*, 2005). For vowel identification where amplitudes of the first three formants were manipulated, HI listeners required 6–7 dB spectral contrast to achieve 75% correct while NH listeners required only 1–2 dB (Leek *et al.*, 1987). Loizou and Poroy (2001) extended this approach to cochlear implant (CI) users, reporting spectral contrast thresholds of 4–6 dB. In all cases, HI and CI listeners' contrast thresholds are elevated compared to NH listeners due to poorer frequency resolution. However, these thresholds are considerably smaller than 20-dB-plus spectral peaks tested in earlier studies of perceptual calibration. The present results suggest HI and CI listeners might also demonstrate perceptual calibration to modest spectral peaks in acoustic contexts, likely larger than +5 dB for NH listeners but not as extreme as +20 dB peaks tested previously.

Acknowledgments

We thank Joshua Alexander for providing stimulus materials as well as helpful comments and suggestions, which were also provided by James Hillenbrand. We thank Andrew McPheron, Elizabeth Niehaus, Ijeoma Okorie, McKenzie Sexton, and Lillian Slaughter for assistance with data collection.

References and links

Alcantara, J. I., and Moore, B. C. J. (**1995**). "The identification of vowel-like harmonic complexes: Effects of component phase, level, and fundamental frequency," J. Acoust. Soc. Am. **97**(6), 3813–3824.

Alexander, J. M., and Kluender, K. R. (**2010**). "Temporal properties of perceptual calibration to local and broad spectral characteristics of a listening context," J. Acoust. Soc. Am. **128**(6), 3597–3613.

Assmann, P. F., and Summerfield, Q. (**2004**). "The perception of speech under adverse conditions," in *Speech Processing in the Auditory System*, edited by S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay (Springer, New York).

Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., and Syrdal, A. (**1997**). "AT&T natural voices text-to-speech computer software," http://www2.research.att.com/~ttsweb/tts/demo.php (Last viewed September 4, 2014).

Clifford, C. W. G., Webster, M. A., Stanley, G. B., Stocker, A. A., Kohn, A., Sharpee, T. O., and Schwartz, O. (**2007**). "Visual adaptation: Neural, psychological and computational aspects," Vis. Res. **47**(25), 3125–3131.

Darwin, C. J., McKeown, J. D., and Kirby, D. (**1989**). "Perceptual compensation for transmission channel and speaker effects on vowel quality," Speech Comm. **8**(3), 221–234.

Dreisbach, L. E., Leek, M. R., and Lentz, J. J. (**2005**). "Perception of spectral contrast by hearing-impaired listeners," J. Speech Lang. Hear. Res. **48**, 910–921.

Fant, G. (**1973**). *Speech Sounds and Features* (MIT Press, Cambridge, MA).

Green, D. M. (**1988**). *Profile Analysis: Auditory Intensity Discrimination* (Oxford University Press, New York).

Kiefte, M., Enright, T., and Marshall, L. (**2010**). "The role of formant amplitude in the perception of /i/ and /u/," J. Acoust. Soc. Am. **127**(4), 2611–2621.

Kiefte, M., and Kluender, K. R. (**2008**). "Absorption of reliable spectral characteristics in auditory perception," J. Acoust. Soc. Am. **123**(1), 366–376.

Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**(2), 820–857.

Kohn, A. (**2007**). "Visual adaptation: Physiology, mechanisms, and functional benefits," J. Neurophys. **97**(5), 3155–3164.

Lea, A. P., and Summerfield, Q. (**1994**). "Minimal spectral contrast of formant peaks for vowel recognition as a function of spectral slope," Percept. Psychophys. **56**(4), 379–391.

Leek, M. R., Dorman, M. F., and Summerfield, Q. (**1987**). "Minimum spectral contrast for vowel identification by normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **81**(1), 148–154.

Leek, M. R., and Summers, V. (**1996**). "Reduced frequency selectivity and the preservation of spectral contrast in noise," J. Acoust. Soc. Am. **100**(3), 1796–1806.

Loizou, P. C., and Poroy, O. (**2001**). "Minimum spectral contrast needed for vowel identification by normal hearing and cochlear implant listeners," J. Acoust. Soc. Am. **110**(3), 1619–1627.

Summerfield, Q., Sidwell, A., and Nelson, T. (**1987**). "Auditory enhancement of changes in spectral amplitude," J. Acoust. Soc. Am. **81**(3), 700–708.

Summers, V., and Leek, M. R. (**1994**). "The internal representation of spectral contrast in hearing-impaired listeners," J. Acoust. Soc. Am. **95**(6), 3518–3528.

Turner, C. W., and Van Tasell, D. J. (**1984**). "Sensorineural hearing loss and the discrimination of vowel-like stimuli," J. Acoust. Soc. Am. **75**(2), 562–565.

Wainwright, M. J. (**1999**). "Visual adaptation as optimal information transmission," Vis. Res. **39**(23), 3960–3974.

Wilcox, R. R. (**2005**). *Introduction to Robust Estimation and Hypothesis Testing*, 2nd ed. (Elsevier Academic Press, London).