Filtered and unfiltered sentences produce different spectral context effects in vowel categorization

UNIVERSITY OF LOUISVILLE

Christian E. Stilp and Ashley A. Assgari
Department of Psychological and Brain Sciences, University of Louisville

4pSC21

INTRODUCTION

Perception of a given speech sound is heavily influenced by surrounding sounds. When spectral properties differ between earlier (context) and later (target) sounds, this can produce **spectral contrast effects (SCEs)** that bias categorization of later sounds.

Context	More likely to perceive
Sentence (unmodified)	/ɪ/ or /ɛ/ vowel target
Sentence with /ɛ/-like (high F_1) frequencies emphasized	/ɪ/ (low F_1)
Sentence with /ɪ/-like (low F_1) frequencies emphasized	/ɛ/ (high F_1)

Most often, investigators studied SCEs by filtering one context sentence two ways (*e.g.,* low-F_1 emphasis, high-F_1 emphasis). This approach is agnostic to the **natural signal statistics (NSS)** of speech, as some speech samples might possess the desired spectral properties without any filtering.

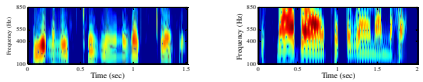
Here, listeners categorized vowels (/ɪ/-/ɛ/) following context sentences that naturally possessed peaks in their long-term spectra without any filtering. Filtered sentences with equivalent spectral peaks were included as controls. This permitted comparisons of how context shapes speech categorization in circumstances that were highly controlled (filtered sentences) versus more variable (and more like everyday speech; unfiltered sentences).

STIMULI

Sentence Contexts

1. Unfiltered

- Drawn from TIMIT (Garofolo *et al*., 1990) and HINT (Nilsson *et al.*, 1994) databases
- **Mean Spectral Differences (MSDs)** were measured
  - Difference in long-term average energy across low-F_1 (100-400 Hz) and high-F_1 (550-850 Hz) frequency regions, in dB



MSD = 11.23 dB (favoring low F_1)

MSD = 15.11 dB (favoring high F_1)

- Selected for having low-F_1-biased MSD or high-F_1-biased MSD
  - Talker, sentence content, and other acoustic parameters freely varied

2. Filtered

- "Please say what this vowel is" spoken by CS (2174 ms), the same stimulus as used in Stilp *et al.* (2015) and other reports
- Processed by FIR filters to amplify one spectral region (100-400 Hz or 550-850 Hz) in order to match the MSD of each unfiltered sentence

Vowels

- Series of 10 natural vowels interpolated from [ɪ] to [ɛ] using PRAAT (246 ms), the same stimuli as used in Stilp *et al.* (2015) and other reports

Trial Structure

- Sentence, 50-ms ISI, then a target vowel which listeners identified as "ih" as in 'bit' or "eh" as in 'bet' (see schematic in Introduction)

METHODS

Participants

- 147 normal-hearing native English speakers (approximately 18 in each of the 8 experiments)
- Required to pass Practice (see below) and maintain 80% accuracy on vowel endpoints throughout the experiment (n=3 removed)

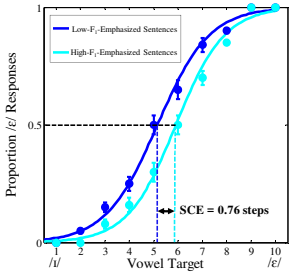
Procedure

- *Practice:* 20 sentences from the AzBio corpus (Spahr *et al.*, 2012) paired with endpoint vowels; >80% accuracy needed to continue to test session
  - Participants allowed to repeat practice session up to 2 more times in order to meet criterion
- *Test:* 160 trials (in random order) in each of 4 blocks (presented in counterbalanced orders)
  - 2 blocks presented unfiltered context sentences; the other 2 blocks presented filtered contexts with MSDs that matched those in the unfiltered sentences
  - Each block presented vowels preceded by either a low-F_1-emphasized sentence or high-F_1-emphasized sentence
  - The net MSD tested in a given block was the average of the MSDs for the two sentence contexts (*e.g.*, for the two sentences shown in Stimuli, 11.23 dB low-F_1 MSD & 15.11 dB high-F_1 MSD = 13.17 net MSD)

RESULTS

Measuring SCEs

- In each condition, logistic regressions were fit to each listener's responses to low-F_1-emphasized and high-F_1-emphasized sentences
- 50% points were calculated from each regression equation and converted into stimulus step numbers (1-10, interpolated as needed; see figure at right)
- SCE = the number of stimulus steps separating the 50% points for the low-F_1-emphasized and high-F_1-emphasized functions
- SCEs were calculated for each listener, then averaged across listeners for each experimental condition
- Each participant group contributed 4 SCEs to the overall analysis (2 SCEs following filtered sentences, 2 SCEs following unfiltered sentences)



Filtered Sentences

- As spectral peak magnitude (*i.e.*, filter gain) increased, SCE magnitudes increased linearly. This relationship was quite strong ($r_{filtered}$ = 0.88, p < .0001)
- Previous studies using this sentence revealed a linear relationship between SCE magnitudes and filter gains of +5, +10, +15, and +20 dB (Stilp *et al.*, 2015; Stilp & Alexander, 2016). Here, this linear relationship interpolates between these values, replicating and extending this linear scaling of SCEs

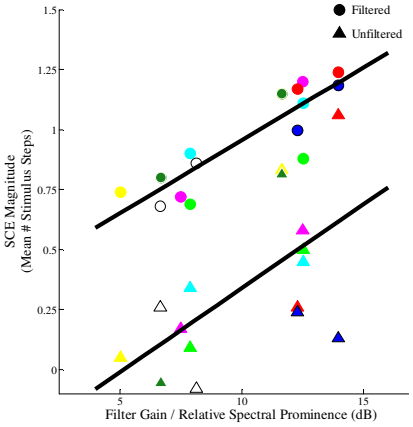
Unfiltered Sentences

Similarities to Filtered Sentences Results

- As spectral peak magnitude (*i.e.*, MSDs) increased, SCE magnitudes again increased linearly ($r_{unfiltered}$ = 0.62, p < .01). This extends the linear relationship between context spectra and SCEs to unfiltered sentence contexts
- The rate at which SCEs increased as a function of MSD was very similar across unfiltered sentences (linear regression slope = 0.07) and filtered sentences (slope = .06)

Differences from Filtered Sentences Results

- SCEs produced by unfiltered sentences (grand mean = 0.35 stimulus steps) are much smaller than SCEs produced by filtered renditions of a single sentence with equivalent MSDs (grand mean = 0.97 steps; t_{15} = 9.89, p < .0001)
- This was also evident in the much smaller intercept to the linear regression fit to unfiltered-sentence SCEs (intercept = –0.36) compared to filtered-sentence SCEs (intercept = 0.35)
- SCEs were more variable following unfiltered sentences (range = 1.14, SD = 0.33) compared to those produced by filtered sentences (range = 0.56, SD = 0.20)



DISCUSSION

- Acoustic contexts need not be highly acoustically controlled to bias subsequent speech categorization. Natural unfiltered sentences with inherent spectral peaks (measured through MSDs) still biased speech categorization. This speaks to the generality of SCEs and their likely widespread influence on everyday speech perception (Stilp *et al.*, 2015).
- However, these approaches are not equivalent. SCEs produced by unfiltered sentences were significantly smaller than those produced by filtered renditions of a single sentence. This is an extremely important consideration in terms of how closely highly controlled studies (*i.e.*, filtered conditions) model everyday speech perception and its considerable acoustic variability (*i.e.*, unfiltered conditions).
- Watkins (1991) suggested that spectral context effects were a means of compensating for systematic distortion in the communication channel. Using unfiltered stimuli removed the intermediary effects of a distorting communication channel, and SCEs were still observed. Thus, spectral context effects are not merely "channel effects" but a general means of emphasizing informative changes in the listening environment.
- Results validate an NSS approach to speech perception. NSS have been highly profitable for understanding vision (*e.g.*, Field, 1987; Schwartz & Simoncelli, 2001; Simoncelli, 2003), but this approach has been difficult to extend to speech given its rampant acoustic variability. By examining sentence-length segments, perceptually relevant statistical regularities emerged (MSDs). This offers great promise for future research tying speech perception to the statistics of the natural acoustic environment.
- Recent research argued that spectral context effects are driven by the long-term average spectrum (LTAS) of the context (Laing *et al.*, 2012). Yet, SCEs were significantly smaller when produced by unfiltered sentences than filtered sentences with equivalent MSDs in relevant frequency regions. Other sources of information beyond the LTAS likely influence the presence and magnitudes of SCEs. For example, sentence-to-sentence variability in talkers' fundamental frequencies restrains the magnitudes of SCEs (Assgari & Stilp, 2015; Assgari *et al.*, 2016). This can help guide stimulus selection and predictions in future research.

REFERENCES

1. Assgari AA, Stilp CE (2015) *JASA, 138(5),* 3023-3032
2. Assgari AA, Mohiuddin A, Theodore RM, Stilp CE (2016) *JASA, 139(4),* 2124
3. Field (1987) *J Opt Soc Am, 4(12),* 2379-2394
4. Garofolo J *et al.* (1990) "DARPA TIMIT acoustic-phonetic continuous speech corpus." National Institute of Standards and Technology, Gaithersburg, MD
5. Laing EJC, Liu R, Lotto AJ, Holt LL (2012) *Front Psych., 3,* 1-9. doi: 10.3389/fpsyg.2012.00203
6. Nilsson M, Soli SD, Sullivan JA (1994) *JASA, 95(2),* 1085-1099
7. Schwartz O, Simoncelli EP (2001) *Nature Neurosci, 4(8),* 819-825
8. Simoncelli EP (2003) *Curr Op Neurobio, 13(2),* 144-149
9. Spahr AJ *et al.* (2012) *Ear Hear, 33(1),* 112-117
10. Stilp CE, Alexander JM (2016) *POMA, 26,* doi:10.1121/2.0000233
11. Stilp CE, Anderson PW, Winn MB (2015) *JASA, 137(6),* 3466-3476
12. Watkins AJ (1991) *JASA, 90(6),* 2942-2955