Short-term, not long-term, average spectra of preceding sentences bias consonant categorization

Christian E. Stilp (christian.stilp@louisville.edu)
Department of Psychological and Brain Sciences, University of Louisville

2pSC2

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)	/d/ or /g/
Sentence with /d/-like (high F_3) frequencies emphasized	/g/ (low F_3)
Sentence with /g/-like (low F_3) frequencies emphasized	/d/ (high F_3)

Stilp and Assgari (2017b; 2018; under review) showed that the **natural signal statistics (NSS)** of sentences (inherent spectral properties without any filtering) were sufficient to bias speech categorization. In particular, Stilp and Assgari (2018) showed that the last 500 ms of context sentence spectra were most important for biasing subsequent /d/-/g/ categorization.

If the last 500 ms of the context is most important for producing SCEs, can spectral properties of earlier in the context nullify that influence? Here we selected and constructed sentences where the Early (everything before the last 500 ms) and Late (last 500 ms) portions of the context made different predictions for performance.

STIMULI

Sentence Contexts

1. Unfiltered

- Drawn from the HINT database (Nilsson *et al.*, 1994)
- **Mean Spectral Differences (MSDs)** were measured
  - MSD = difference in long-term average energy across low-F_3 (1700-2700 Hz) and high-F_3 (2700-3700 Hz) regions, in dB
  - Positive MSDs indicate Low F_3 energy > High F_3 energy, negative MSDs indicate Low F_3 energy < High F_3 energy
  - Measured separately in Early (everything before last 500 ms) and Late (last 500 ms) portions of context sentences
- Stimuli possessed spectra with two different patterns
  - Early vs. Late: MSDs strongly biased in opposite directions
  - Nothing vs. Late: Early spectrum had MSD ≈ 0, Late spectrum had large MSD
- Sentence content, duration, and other acoustic parameters freely varied

2. Filtered

- "Correct execution of my instructions is crucial" (2200 ms) from TIMIT (Garofolo *et al.*, 1990), the same stimulus as used in Stilp and Assgari (2017a)
- Processed by FIR filters to amplify one spectral region (1700-2700 Hz or 2700-3700 Hz) in order to match the MSD of each part of each unfiltered sentence (Early and Late)
- Filtering conducted separately for Early and Late sentence segments

Consonant Targets

- Series of 10 natural CVs interpolated from [da] to [ga] (365 ms) from Stephens and Holt (2011); the same stimuli as used in Stilp and Assgari (2018; under review)

METHODS

Participants

- 17 native English speakers with no known hearing impairments

Procedure

- *Practice:* 20 sentences from the AzBio corpus (Spahr *et al.*, 2012) paired with endpoint consonants; >80% categorization accuracy needed to continue to test
- *Test:* 160 trials (in random order) in each of four blocks (illustrated below; presented in counterbalanced orders)
  - Two blocks presented unfiltered context sentences; the other two blocks presented filtered contexts with MSDs that matched unfiltered sentence MSDs
  - Trial structure: sentence, 50-ms ISI, then target CV which listeners identified as "da" or "ga" (see schematic in Introduction)

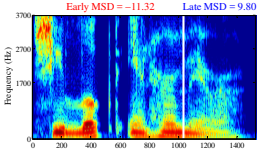
SCE

- For each block, measured as the mean number of stimulus steps separating 50% points on logistic regressions fit to responses following each context sentence
- Group data are shown below, which are consistent with the mean SCEs listed in each figure title (all of which were significantly greater than zero)

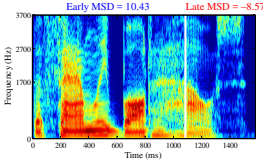
RESULTS

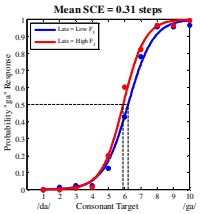
Early vs. Late, Unfiltered

"She looked in her mirror"
Early MSD = −11.32 Late MSD = 9.80



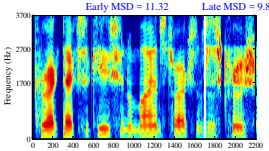
"The family bought a house"
Early MSD = 10.43 Late MSD = −8.57





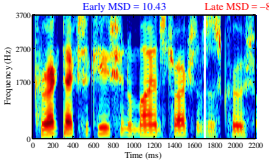
Early vs. Late, Filtered

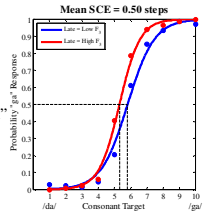
"Correct execution of my instructions is crucial"
Early MSD = 11.32 Late MSD = 9.80



(signal processing error: Early MSD should be −11.32, but results still show the strong influence of the Late window)

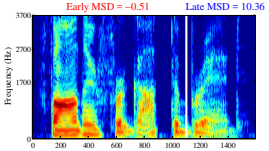
"Correct execution of my instructions is crucial"
Early MSD = 10.43 Late MSD = −8.57



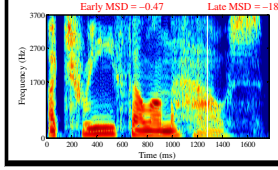


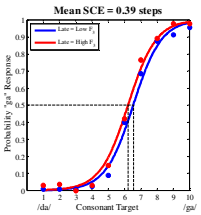
Nothing vs. Late, Unfiltered

"Father forgot the bread"
Early MSD = −0.51 Late MSD = 10.36



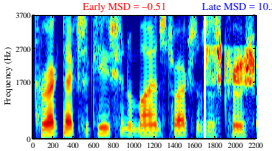
"A tree fell on the house"
Early MSD = −0.47 Late MSD = −18.08



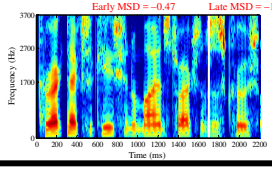


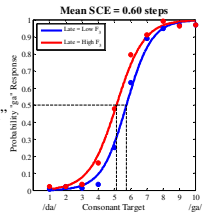
Nothing vs. Late, Filtered

"Correct execution of my instructions is crucial"
Early MSD = −0.51 Late MSD = 10.36



"Correct execution of my instructions is crucial"
Early MSD = −0.47 Late MSD = −18.08





DISCUSSION

- Unfiltered context sentences produced SCEs that biased consonant categorization, consistent with Stilp and Assgari (2018)

- Late MSDs predicted performance
  - Positive Late MSD = more high-F_3 /d/ responses
  - Negative Late MSD = more low-F_3 /g/ responses
- Early MSDs had no influence on performance, whether they exceeded ±10 dB (Early vs. Late) or were ≈ 0 (Nothing vs. Late)
  - Early vs. Late Unfiltered tested against Nothing vs. Late Unfiltered in paired-*t* test: $t_{16} = 0.44$, $p = 0.67$
  - Early vs. Late Filtered tested against Nothing vs. Late Filtered in paired-*t* test: $t_{16} = 0.76$, $p = 0.46$

- MSDs of entire unfiltered sentences cannot predict these results
  - Early vs. Late: entire-sentence MSDs were large but of the opposite sign of Late MSDs ("She looked in the mirror" MSD = −10.69; "The family bought a house" MSD = 8.39)
  - Nothing vs. Late: entire-sentence MSDs were extremely small ("Father forgot the bread" MSD = 0.60, "A tree fell on the house" MSD = −2.46), yet these materials biased performance to a similar degree as the Early vs. Late Unfiltered stimuli

- SCEs were numerically smaller in Unfiltered conditions than in Filtered Conditions, but these differences were not statistically significant (*t*-tests ≈ 1, $p \approx 0.33$)
  - This difference was significant in Stilp and Assgari (under review), but that was across eight vowel categorization experiments. The comparison here is likely underpowered
  - Variability in duration, phonetic content, and many other properties across unfiltered sentences likely contribute

- Results deviate from timecourse work by Holt (2005; 2006), particularly with Early MSDs failing to nullify the influence of Late MSDs
  - Sizeable differences in how speech versus nonspeech (tone) contexts sample frequency regions over time

- MSDs in the last 500 ms of context sentences were a poor predictor of vowel categorization in Stilp and Assgari (under review)
  - Are the present results specific to consonant (/d/-/g/) categorization? Parallel research examining Early/Late windows and vowel (/ɪ/-/ɛ/) categorization needed

REFERENCES

1. Garofolo J *et al.* (1990) "DARPA TIMIT acoustic-phonetic continuous speech corpus." National Institute of Standards and Technology, Gaithersburg, MD
2. Holt LL (2005) *Psych Sci, 16(4)*, 305-312
3. Holt LL (2006) *JASA, 120(5)*, 2801-2817
4. Nilsson M, Soli SD, Sullivan JA (1994) *JASA, 95(2),* 1085-1099
5. Spahr AJ *et al.* (2012) *Ear Hear, 33(1)*, 112-117
6. Stephens JDW, Holt LL (2011) *Sp Comm, 53(6)*, 877-888
7. Stilp CE, Assgari AA (under review) *Atten Percept Psychophys*
8. Stilp CE, Assgari AA (2017a) *JASA, 141(2)*, EL153-EL158
9. Stilp CE, Assgari AA (2017b) *JASA, 142*, 2707
10. Stilp CE, Assgari AA (2018) *JASA, 143*, 1944