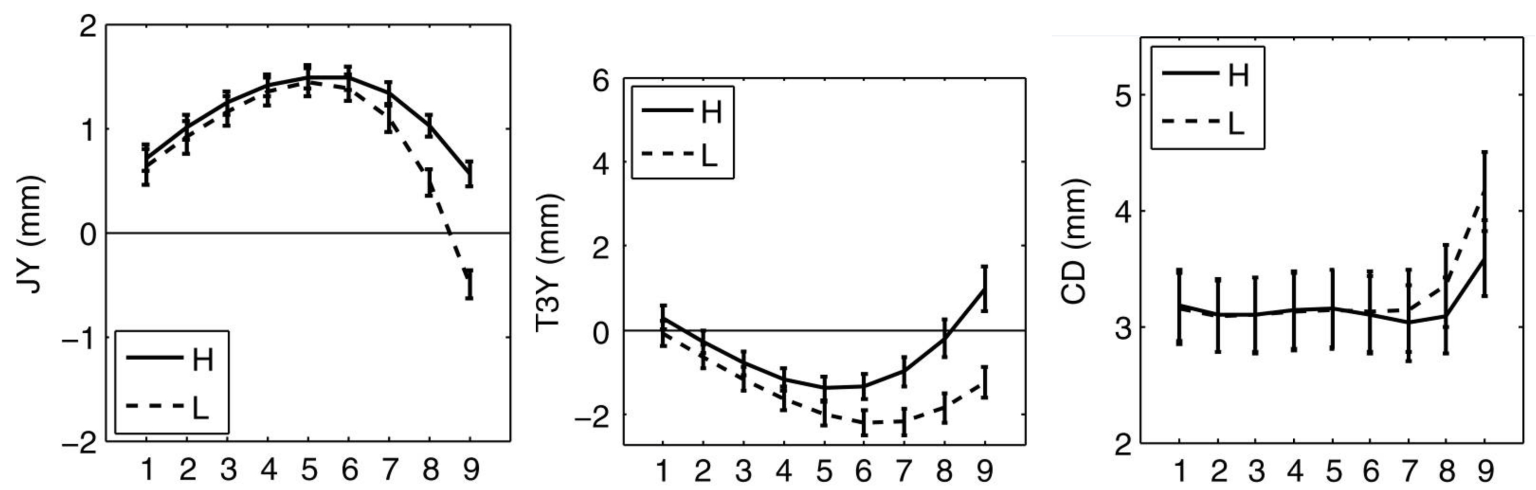
Transcriptional and acoustical methods for analyzing young children's gestural coordination patterns in sibilant fricatives

Patrick Reidy, *The Ohio State University, Dept. of Linguistics*

Background & Purpose of Study

Gestural coordination in sibilant fricatives

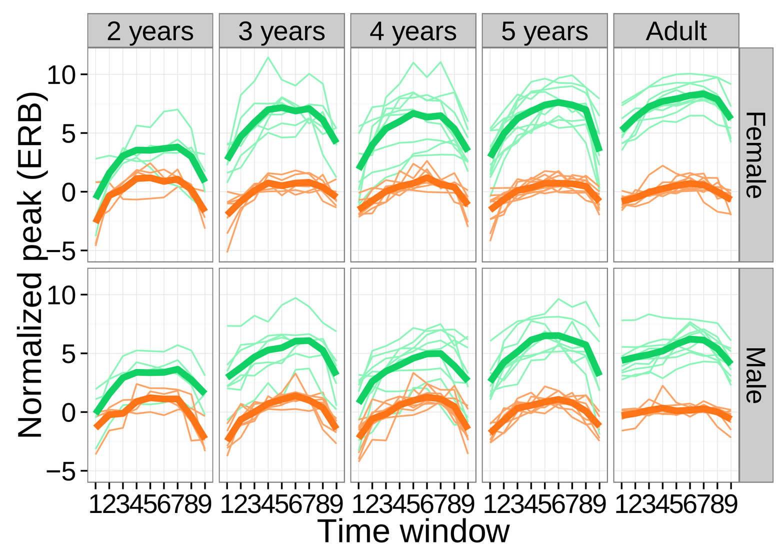
- The successful production of an adult-like /s/ or /ʃ/ requires the coordination of labial, mandibular, lingual, and laryngeal gestures.
- The formation of a narrow oral constriction and the positioning of the incisors are important to the generation of turbulence noise sources.



Mandibular and lingual gestures are coordinated so that changes in jaw height (JY, left panel) and tongue height (T3Y, center panel) yield a relatively stable degree of constriction in the oral cavity (CD, right panel), until the release of the constriction. (From Iskarous, Shadle & Proctor, 2011.)

Articulatory gestural coordination is not native in children

- /s/ is not acquired in a majority of children until age 3, and /ʃ/ is not acquired until age 4. Some children struggle with the /s/–/ʃ/ contrast until age 7 (Sander, 1972; Smit *et al.*, 1990).
- The spectral kinematic patterns of two-through five-year-old children's correct productions of /s/ and /ʃ/ differ significantly from adults' (Reidy, 2014; figure at right).

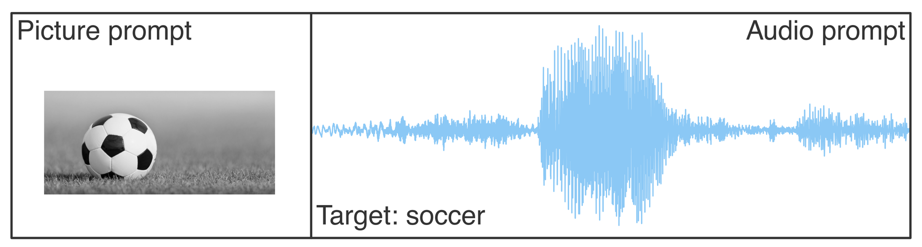


Purpose of current study

- Present methods that exploit *acoustic* speech production data, which are relatively cheap and easy to acquire, for answering questions related to the development of gestural coordination in children's sibilant fricative productions.
- Transcriptional methods that index changes in children's ability to coordinate the mandibular and lingual gestures so that an oral constriction is formed and maintained.
- Acoustical methods that help clarify changes in the dispersion of sibilant fricative categories, which indexes the development of speech motor control.

Acoustic Production Data: $\pi\alpha\iota\delta o\lambda o\gamma o\varsigma$ Project

- Participants were native English-speaking adults ($N = 20$), and two- (younger $N_a = 8$; older $N_b = 11$), three- ($N_a = 10$; $N_b = 10$), four- ($N_a = 13$; $N_b = 8$), and five-year-old children ($N_a = 11$; $N_b = 9$) who were acquiring English natively.
- Tokens of /s/ or /ʃ/ in word-initial, pre-vocalic position of real English words were elicited during an audio-prompted, picture-naming task.



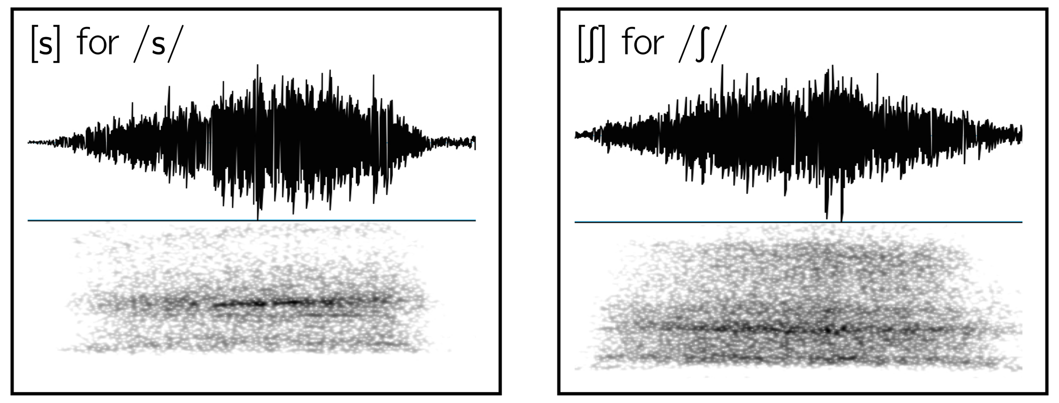
- Each production was judged for phonemic correctness by a trained phonetician.
  - These phonemic judgments were used to exclude children from the acoustic analysis: Only those children who produced at least 3 correct instances of /s/ and /ʃ/ were included.
  - This exclusion criterion left 8 two-, 14 three-, 18 four-, and 19 five-year-olds.

Transcriptional Methods

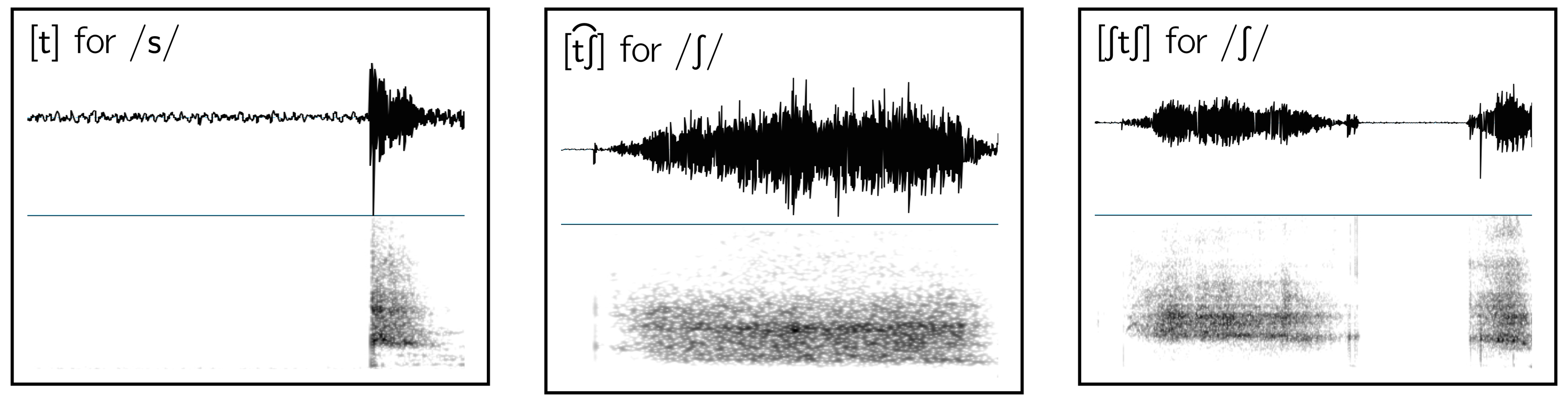
Transcription symbol set

- WorldBet symbols, which could be combined in series to denote a production whose quality changed over time, or with a colon (:) to denote a sound intermediate between two others.
- Transcriptions were then pooled into four classes that reflect differences in gestural coordination.

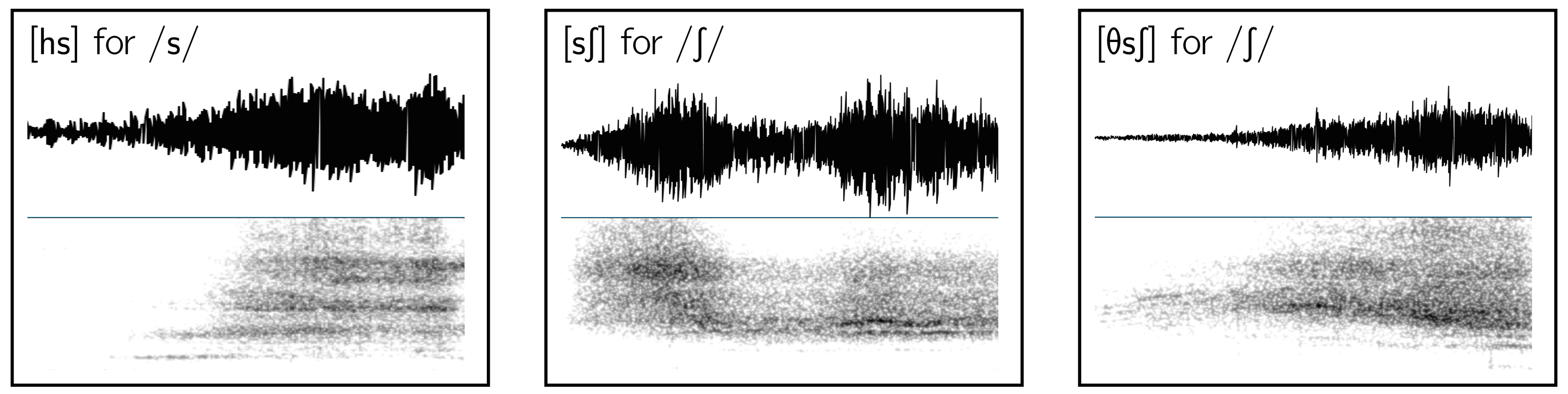
1) *Correct productions*: Phonetic match to target.



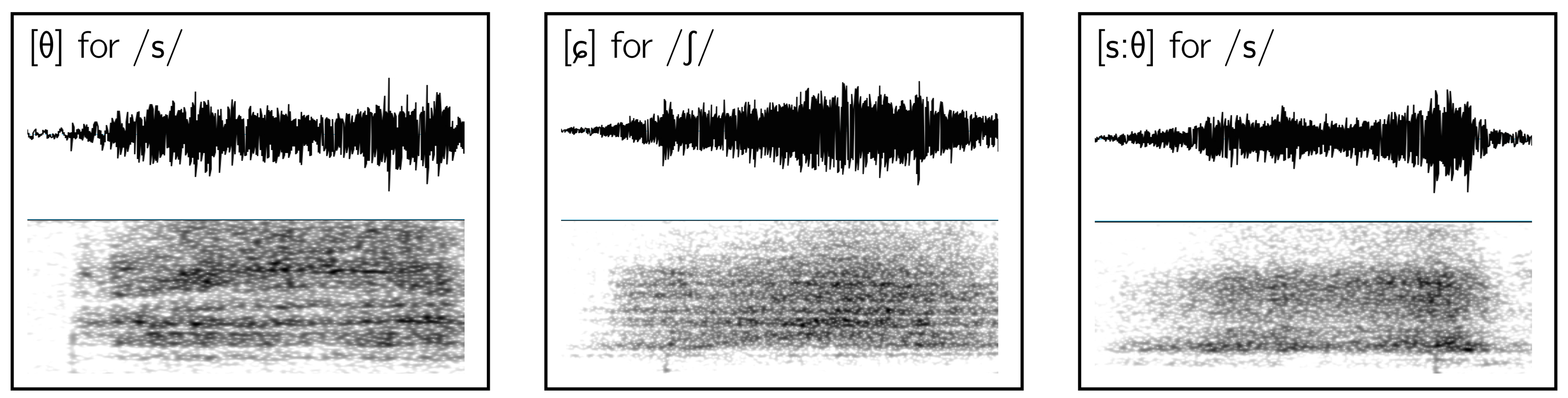
2) *Fortition errors*: Comprised a stop or an affricate.



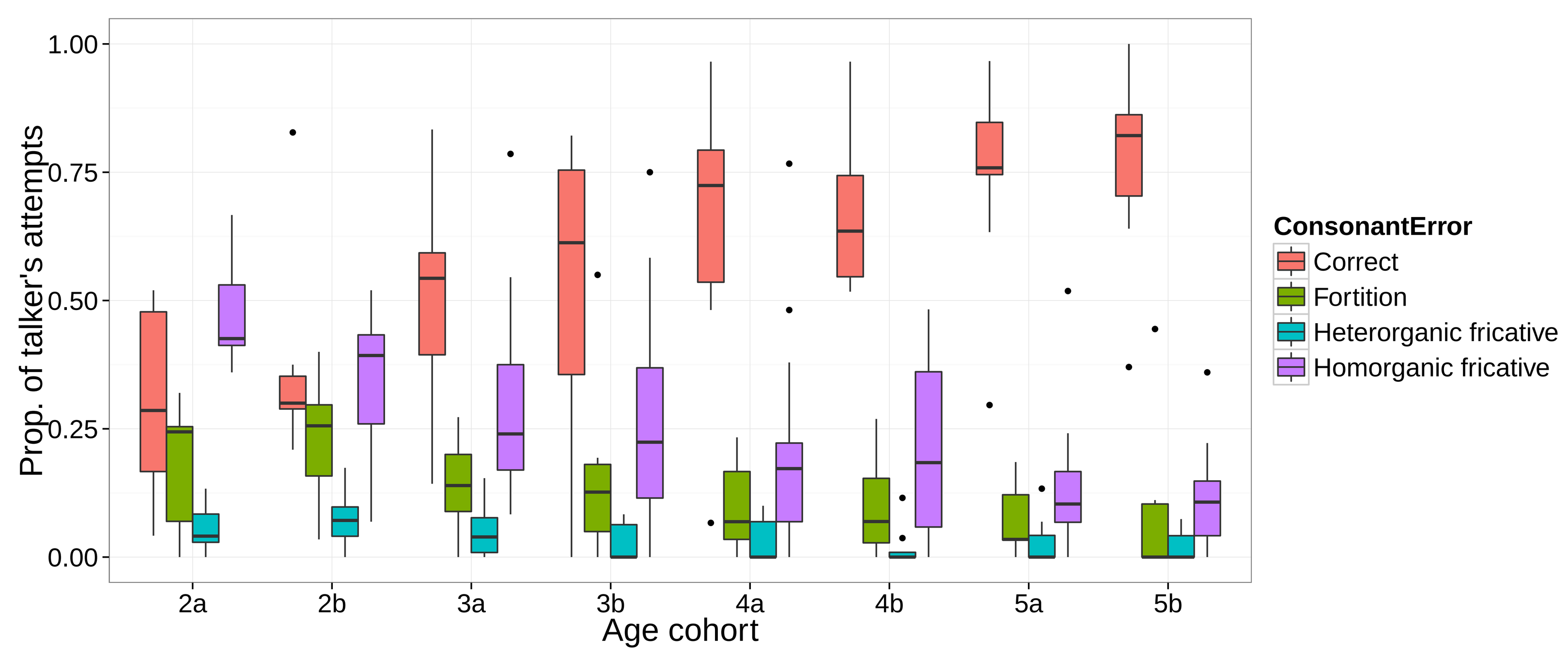
3) *Heterorganic fricative errors*: Sequences of fricative symbols.



4) *Homorganic fricative errors*: A single fricative or intermediate between to fricatives.



Cross-sectional error patterns

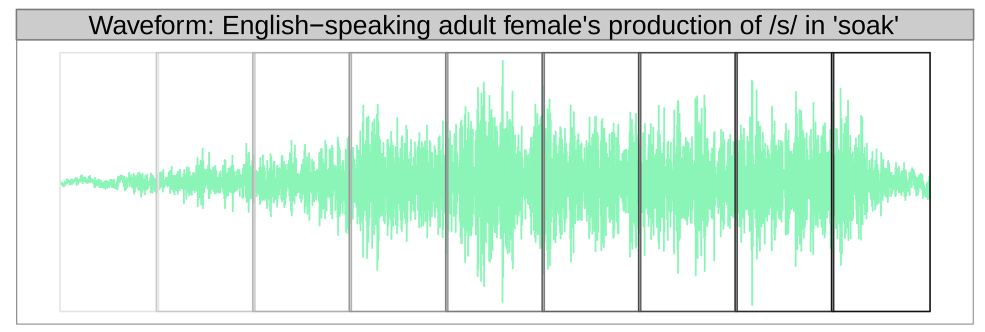


Acoustical Methods: Peak ERB Trajectory

Overview & motivation

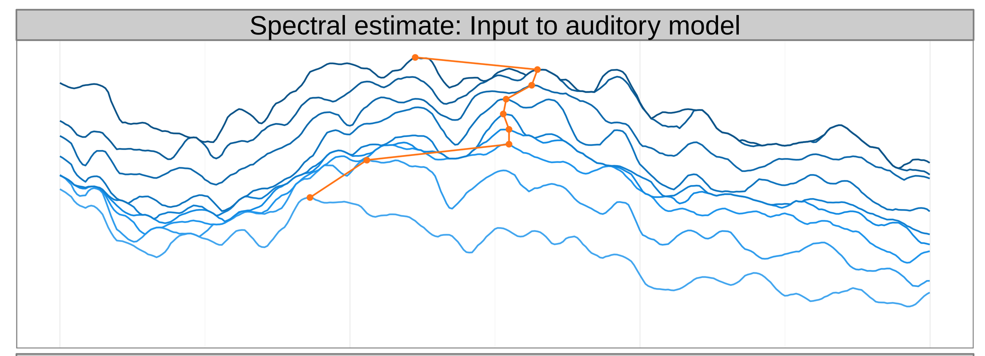
- Peak ERB is a measure of the most prominent *psychoacoustic* frequency, rather than physical frequency. The auditory model employed in the computation of the psychoacoustic spectrum offers a physiologically plausible (and relevant) way to smooth the "noisy" spectrum of a sibilant.
- Gestures are *dynamic*, not static; so, a trajectory, rather than a point measure is likely to be more informative here.

Computation of peak ERB trajectory



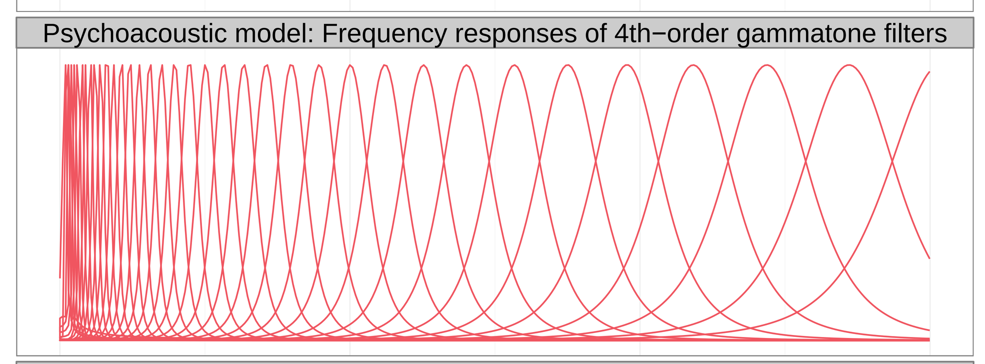
Sibilant waveform pre-processing

- Onset and offset of frication marked manually by a trained phonetician.
- Waveform was not pre-emphasized.
- Nine 20-ms windows spaced evenly across duration of the sibilant.



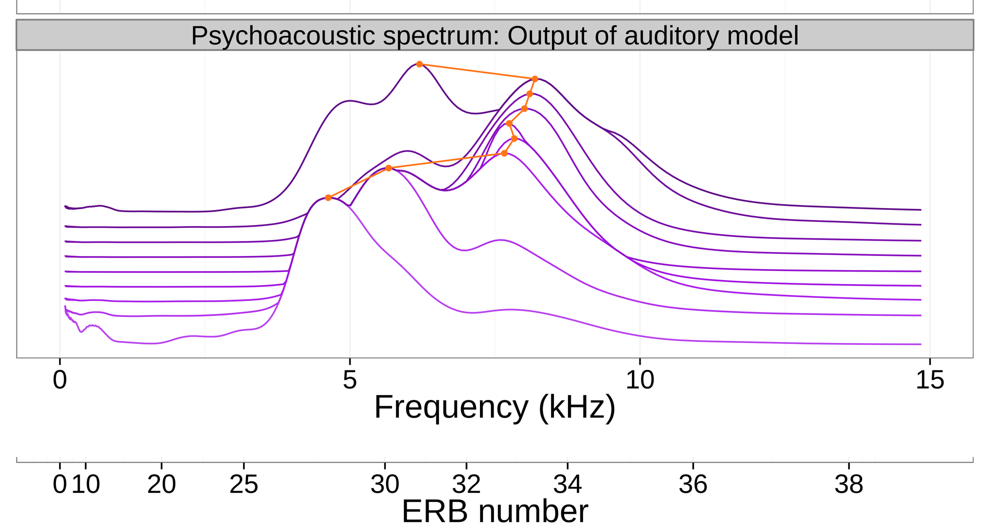
Multitaper spectrum ($K = 8$; $NW = 4$)

- MTS is similar to DFT, but estimates ordinate values with less error.
- Spectra were estimated from the nine windows (light to dark). Spectral peak trajectory is shown as orange path.



Gammatone filterbank (361 channels)

- Center frequencies spaced every 0.1 ERB; bandwidths proportional to CFs.
- Models cochlear differential frequency selectivity, with respect to notched-noise masking.



Peak ERB trajectory

- The amplitude of the psychoacoustic spectrum at a given frequency ω is the total energy output by the filter, whose CF is ω, in response to a given input spectrum.
- In the analysis of category dispersion below, a talker's sibilant category (/s/ or /ʃ/) is represented by a collection of peak ERB trajectories.

Demonstration: Dispersion of sibilant categories

- The dispersion of a phone category has been argued to index speech motor control (cf. Smith & Goffman, 1998).
- Adults tend to produce /s, ʃ/ with less dispersion than young children (Munson, 2004) or adolescents (Romeo, Hazan & Pettinato, 2013).
- We examined whether this age difference in dispersion is drawn into sharper relief when computed from peak ERB trajectories, rather than univariate, midpoint values.
  - Dispersion of a category $\{\sigma_n\}$ (a collection of peak ERB trajectories or midpoint values), with mean $\overline{\sigma}$, is:

$$\text{Dispersion}(\{\sigma_n\}) = N^{-1} \sum_n M^2(\sigma_n, \overline{\sigma})$$

  - Manhattan distance: $M(\sigma_1, \sigma_2) = \sum_t |\sigma_1[t] - \sigma_2[t]|$
  - The expected age difference in category dispersion is indeed revealed more clearly when computed from trajectories, rather than point values.