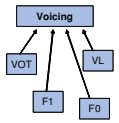
Taking statistical learning to the next level: A computational approach to the acquisition of multi-dimensional categories

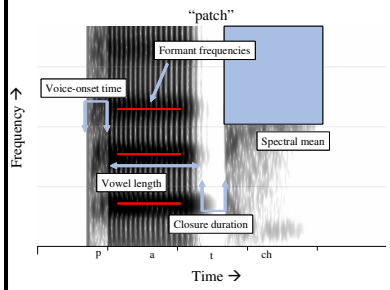
Joseph C. Toscano and Bob McMurray
University of Iowa, Dept. of Psychology

Acoustic cue integration in speech

- Multiple acoustic cues often contribute to a single phonetic categorization
- E.g. Lisker (1978) identified at least 16 cues to word-medial voicing in English
- Typically examined using *trading relations* (Repp, 1982)

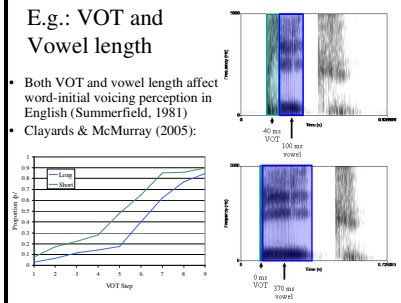


Some acoustic cues in speech



E.g.: VOT and Vowel length

- Both VOT and vowel length affect word-initial voicing perception in English (Summerfield, 1981)
- Clayards & McMurray (2005):



Question

- Identification data suggests cues are weighted differently
- *Where do cue weights come from?*
  - Possible answer: reliability of cues
  - Reliability is a natural by-product of a development mechanism, such as statistical learning

Can a developmental mechanism account for acoustic cue weights in speech?

Cue reliability and weighting

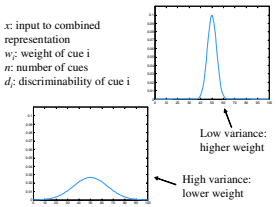
- Cues could be weighted based on their reliability
- Kalman filter: cue weights proportional to the variance of a dimension

$$d_i = \frac{1}{\sigma_i^2}$$

$$w_i = \frac{d_i}{\sum_j^n d_j}$$

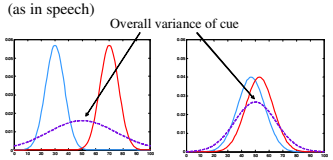
$$x = \sum_i^n w_i x_i$$

x: input to combined representation
w_i: weight of cue i
n: number of cues
d_i: discriminability of cue i



Single vs. multiple categories

- Kalman filter works for weighting a single distribution along a cue dimension
- However, it would not work for multiple categories (as in speech)



Distinct categories: high overall variance

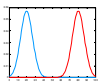
Overlapping categories: lower overall variance

Cue reliability for multiple distributions

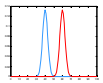
- What cues are reliable?
- Reliability depends on variance and locations of categories along a dimension

Highly reliable cues:

Widely spaced categories

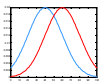


Low-variance categories

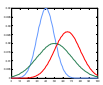


Unreliable cues:

Overlapping distributions



Numerous categories along a single dimension



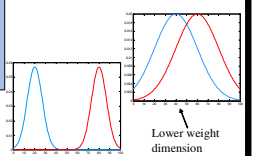
A more complicated weighting function was used that takes into account the relative frequency of each category, the distance between them, and their variability.

$$md_i = \sum_m^K \sum_l^K \frac{\phi_{il}\phi_{im}\sqrt{(\mu_{il}-\mu_{im})^2}}{\sigma_{il}\sigma_{im}}$$

$$w_i = \frac{md_i}{\sum_j^n md_j}$$

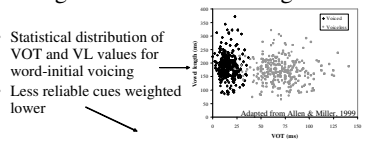
$$x = \sum_i^n m w_i x_i$$

x: input to combined representation
w_i: weight of cue i
n: number of cues
K: number of Gaussians along each dimension
d_i: discriminability of cue i

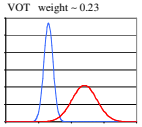


E.g.: VOT and vowel length

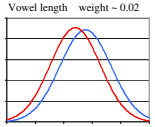
- Statistical distribution of VOT and VL values for word-initial voicing
- Less reliable cues weighted lower



VOT weight ~ 0.23



Vowel length weight ~ 0.02



Reliability is developmental

- Statistical distributions can determine reliability, which can be used to weight cues
- Reliability is determined with long-term experience over the course of development
- Developmental mechanism for measuring reliability: ***statistical learning***
  - Already being used to learn speech categories
  - Could also be used to learn reliability – can track both the prototype and variance of a cue

Can the process of category acquisition *also* yield information about reliability?

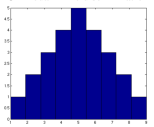
Statistical learning

Statistical learning provides a mechanism for learning the entire distribution of a speech cue

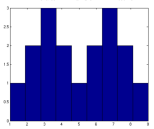
Maye et al. (2002):

- Infants exposed to a unimodal distribution of speech sounds will not discriminate different tokens
- Infants exposed to a bimodal distribution will discriminate

Unimodal → no discrimination



Bimodal → discrimination

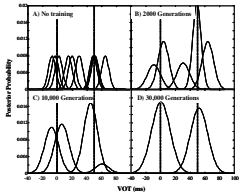


Approach

- Use a computational model of speech category learning to answer two questions:
  - Does reliability give the correct weights?
  - Does the learning process match the developmental trajectory of children?
- Our model extends an earlier model that learns categories along a single dimension (McMurray, Aslin, Toscano, in press)

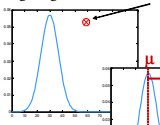
Model architecture

- Mixture of Gaussians (MOG) model
- Each Gaussian distribution represents a possible acoustic category
- Model is trained on acoustic cue-values from speech production measurements
- The model begins with many distributions and eliminates unnecessary ones
- During learning, the model adjusts the parameters of the distributions to match the input

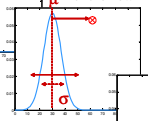


Learning

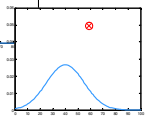
e.g. single Gaussian; receives new input



Parameters are adjusted to account for the new input

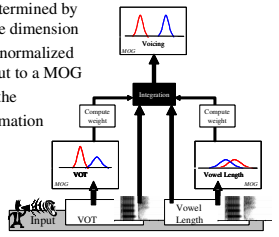


Gaussian after learning



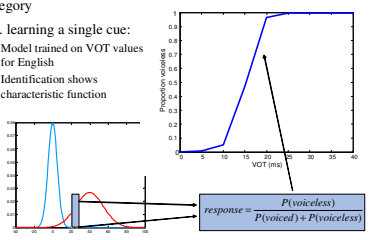
Cue integration in the model

- Cue weights determined by reliability of cue dimension
- Cue-values are normalized and used as input to a MOG that represents the combined information from each cue



Testing the model

- Identification measured by computing posterior probability for different stimuli for the closest matching Gaussians to each category
- e.g. learning a single cue:
  - Model trained on VOT values for English
  - Identification shows characteristic function



$$response = \frac{P(voiceless)}{P(voiced) + P(voiceless)}$$

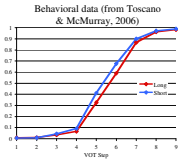
Simulations

- Question 1: *Does the statistical reliability of a cue determine how much it is weighted?*
  - English word-initial voicing (/b/-/p/)
  - Japanese fricative-affricate (/ʃ/-/tʃ/)
- Question 2: *Does the developmental trajectory match that of children?*
  - English fricative place (/s/-/ʃ/)

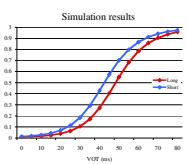
Simulation 1: English voicing

- 50 models trained on English VOT and VL distributions for voicing categories
- Identification results show a trading relation that corresponds to behavioral data from human listeners

Behavioral data (from Toscano & McMurray, 2006)

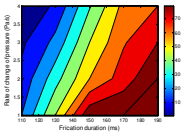


Simulation results

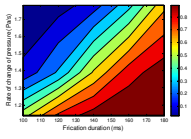


Simulation 2: Japanese fricative-affricate

Japanese listeners (adapted from Mitani et al. (2006)



Simulation results



- Model and human listeners' proportion /ʃ/ identification for different frication duration and rise time values
- 50 models trained on production data from Japanese speakers and tested on stimuli similar to those used with human listeners
- The model can successfully weight and combine multiple cues

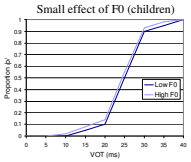
Statistical learning is sufficient for weighting cues

- Initial simulations show that reliability metric provides appropriate cue weights
  - Metric is based on statistical properties of the input – information learned over development
  - Statistical learning provides sufficient information for learning categories and weighting cues
- Remaining question: *Does the model follow the same developmental trajectory as children?*

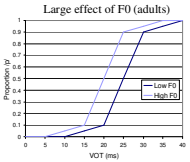
Changes over development

- Weight of individual cues changes over development
- e.g. Bernstein (1983): children do not use F0 for word-initial voicing, but adults do

Small effect of F0 (children)



Large effect of F0 (adults)



- Nittrouer (2002):
  - Tested children's identification of /s/-/ʃ/ distinction and varied two cues: spectral mean & formant frequencies
  - Found that children were more biased towards using formant cues
  - Children perceptually biased toward using formant transitions because they attend to general movements of the vocal tract when perceiving speech
- *However…* Mayo & Turk (2004):
  - Tested other phonetic contrasts and continua
  - Found that children's cue weighting depends on context
  - Children are not always biased towards transitions

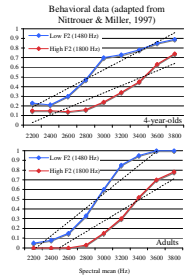
Does the model show developmental patterns that are similar to children?

Simulation 3: English fricative place

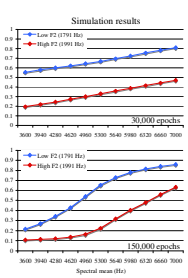
- Nittrouer (2002):
  - /s/-/ʃ/ distinction
  - Varied formant onset frequency and spectral mean
  - Found children relied on formant cues more
- Testing:
  - 50 models tested
  - Measured cue weights over time
  - Tested identification during learning (every 10,000 epochs) to assess developmental trajectory
- Results:
  - Similar developmental trajectory to children

Identification results

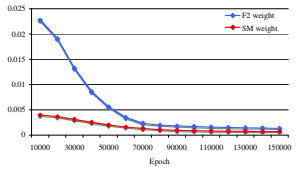
Behavioral data (adapted from Nittrouer & Miller, 1997)



Simulation results



Cue weights over time



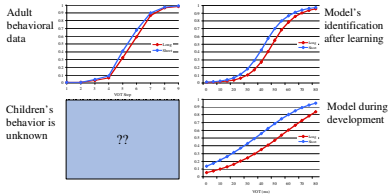
- F2 initially weighted higher
- Model learns that the cue is less reliable and downweights it after more learning

Discussion

- Model demonstrates that:
  - Reliability can be used to weight cues, and
  - It can be determined via statistical learning
- **Cue weighting can be accomplished without any additional learning mechanisms beyond what infants and children can already use**
  - The same information used to learn categories is applied to the problem of cue weighting
- Statistical learning serves a dual purpose:
  - Learning distributional information
  - Determining the reliability of perceptual cues in order to weight and integrate them

Future work

- Can the model be used to predict children's behavior for other sets of cues?
- VOT and vowel length:



References

Allen, J. S. & Miller, J. L. (1999). Effect of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *Journal of the Acoustical Society of America, 106*, 2031-2039.

Bernstein, L. (1983). Perceptual development for labeling words varying in voice onset time and fundamental frequency. *Journal of Phonetics, 11*, 383-393.

Clayards, M. & McMurray, B. (2005). Can the temporal encoding of gradient lexical activation reveal the grain of phonological units? Paper presented at the 11th Midcontinental Workshop on Phonology, Ann Arbor, MI.

Lisker, L. (1978). *Rapid* vs. *rabid*: A catalogue of acoustic features that may cue the distinction. *Haskins Laboratories Status Report on Speech Research, SR-54*, 127-132.

Maye, J., Weker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*, B101-B111.

Mayo, C. & Turk, A. (2004). Adult-child differences in acoustic cue weighting are influenced by segmental context: Children are not always perceptually biased toward transitions. *Journal of the Acoustical Society of America, 115*, 3184-3194.

McMurray, B., Aslin, R. N., & Toscano, J. C. (in press). Statistical learning of phonetic categories: Insights from a computational approach. Invited submission for a special section of *Developmental Science*.

Nittrouer, S. & Miller, M. E. (1997). Predicting developmental shifts in perceptual weighting schemes. *Journal of the Acoustical Society of America, 101*, 2253-2266.

Nittrouer, S. (2002). Learning to perceive speech: How fricative perception changes, and how it stays the same. *Journal of the Acoustical Society of America, 112*, 711-719.

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin, 92*, 81-110.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 1074-1095.

Toscano, J. C. & McMurray, B. (2006). A lexical locus for the integration of asynchronous cues to voicing: An investigation with natural stimuli. Poster presented at the 151st Meeting of the Acoustical Society of America, Providence, RI.

Acknowledgements

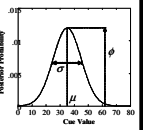
- This research was supported by a University of Iowa Student Government Research Grant to JCT and NIDCD 1R01DC008089-01A1 to BM.
- We would like to thank Joanne Miller, J. Sean Allen, and Allard Jongman for providing phonetic data used to train the model.

Additional details about the model:

- Each category defined by a Gaussian distribution:

$$G_i(x) = \phi_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{(x-\mu_i)^2}{2\sigma_i^2}\right)$$

- Each Gaussian has three parameters:
  - μ: centroid of the distribution
  - σ: spread (variance) of the distribution
  - ϕ: height (frequency) of the distribution



- Likelihood of a particular cue-value is the sum of the posteriors of each Gaussian:

$$M(x) = \sum_i^K G_i(x)$$

Learning algorithm

- Gaussian parameters are updated using maximum likelihood estimation by gradient descent
- To compute learning rules, take the partial derivative of the likelihood function w.r.t. each Gaussian parameter
- Use derivatives to update parameters based on each input

$$\Delta\mu_i = \eta_\mu \cdot \left(\frac{G_i(x)}{M(x)}\right)\frac{(x-\mu_i)}{\sigma_i^2}$$

$$\Delta\sigma_i = \eta_\sigma \cdot \left(\frac{G_i(x)}{M(x)}\right)\left(\sigma_i^{-3}(x-\mu_i)^2 - \sigma_i^{-1}\right)$$

$$\Delta\phi_i = \eta_\phi \frac{G_i(x)}{M(x)}$$

Input to the model

- The model received individual cue-values as input along each dimension and updated its parameters on each trial
- Data used to train the model was sampled from distributions based on acoustic measurements of the various cues used

Learning parameters and starting states

- Learning rate parameters:
  - η_μ
  - η_σ
  - η_ϕ

  } Learning rates for each Gaussian parameter
- Starting state parameters:
  - μs: randomly distributed in the range of cue-values along that dimension
  - σs: each Gaussian has the same starting σ
  - ϕs: 1/K (K = number of Gaussians [constant])