Modeling age of exposure in L2 learning of vowel categories

Meghan Clayards[1], Joe Toscano[2]

[1]Centre for Research on Language Mind and Brain, McGill University, [2]Department of Psychology, University of Iowa

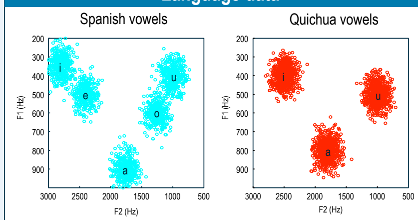
Background

- Age of exposure is a predictor of second language proficiency
- Native-like *phonological* proficiency is attained only by learners exposed at the earliest ages (simultaneous bilinguals)

Does early commitment to phonological patterns account for age of exposure effects?

- Two computational models were trained using unsupervised learning (Toscano & McMurray, 2008; McMurray et al., 2009)
  - 2D Gaussian Mixture Model
  - Hebbian Normalized Recurrence Network
- Trained on F1 and F2 tokens from two different vowel systems
  - Quichua: 3 vowels (/i/, /a/, /u/)
  - Spanish: 5 vowels (/i/, /e/, /a/, /o/, /u/)
- Compared to production data of simultaneous, early, mid, and late bilinguals from Guion (2003)

Language data



Quichua has **one** front vowel and **one** back vowel
Spanish has **two** front vowels and **two** back vowels

Guion (2003) studied four groups of Quichua-Spanish bilinguals exposed to Spanish at different ages

- Simultaneous bilinguals distinguished Spanish and Quichua vowels
- Late learners (age of first exposure > 13 yrs) produced only one front vowel and one back vowel
- Learners exposed before age 13 varied in the number of vowels they produced
- Front vowels were more often distinguished than back vowels
- Both models were trained on tokens from distributions modeled on the productions of monolingual speakers

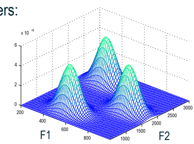
References

Guion, S. (2003) The vowel systems of Quichua-Spanish bilinguals. *Phonetica, 60*, 68-128.
McMurray, B. Aslin, R.N., Toscano, J.C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science, 12*, 369 - 378.
Toscano, J.C., McMurray, B. (2008, November). Online processing of acoustic cues in speech perception: Comparing statistical and neural network models. Poster presented at the 156th Meeting of the Acoustical Society of America, Miami, FL.

Models

2D Gaussian Mixture Model (GMM)

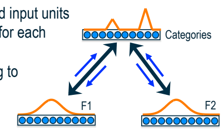
- Categories defined by 2D (F1 x F2) Gaussian distributions
- Each Gaussian defined by six parameters:
  - φ: likelihood of category
  - μ_{F1} and μ_{F2}: mean along each cue dimension
  - σ_{F1} and σ_{F2}: standard deviation along each dimension
  - ρ: correlation between F1 and F2



- Each model is initialized with 200 Gaussians with equal φ, σ, and ρ and random μ–values
- During training, model is presented with F1 and F2 values
- Parameters updated via *maximum likelihood estimation* and *winner-take-all competition* (McMurray et al., 2009)
- Trained on 80,000 training trials per run (30 runs total)

Hebbian Normalized Recurrence Network (HNRN)

- Two-layer neural network
- 200 topographically organized input units with Gaussian tuning curves for each dimension (F1 and F2)
- 20 output units corresponding to possible vowel categories
- On each training trial
  - Pair of F1 and F2 values are presented at input layer
  - Activation feeds forward to output units
  - Output units compete
  - Activation feeds back to input units
  - Weights are updated via Hebbian learning
  - Process repeats until output unit activation settles
- Trained on 40,000 training trials per run (30 runs total)



Simulation groups

- Monolingual Quichua and Spanish groups
- Simultaneous bilingual: trained on tokens from both languages
- Early bilingual: started training with Quichua and switched to input from both languages after 25% of total training time
- Mid bilingual: switched after 50% of total training
- Late bilingual: switched after 75% of total training

Analyses

- Total number of categories
- Mean F1 and F2 values of model categories
- Number of front vs. back vowel discriminations

Number of categories

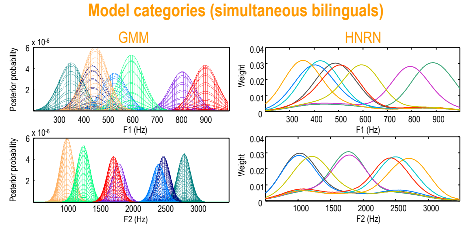
- GMM: components with ϕ > mean
- HNRN: output units with weights > mean (averaged across F1 and F2 weights)

Category means

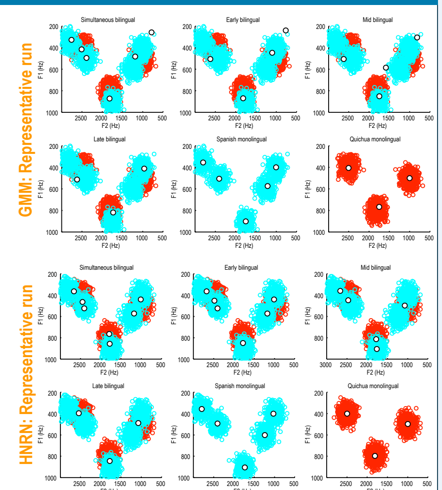
- GMM: μ_{F1} and μ_{F2}
- HNRN: input unit with greatest weight

Results

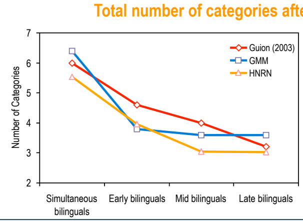
Model categories (simultaneous bilinguals)



- Models were able to learn the vowel systems of each language
- Categories reflect distribution of data in the input



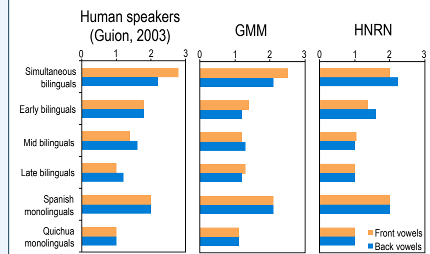
Total number of categories after learning



- Number of categories decreases with later exposure to L2
- Model categories clustered near vowel means for each language
- Fits with results from Guion (2003)

Results (continued)

Front and back vowel distinctions



- As with bilingual speakers, both models show fewer front and back vowel distinctions with later exposure

Discussion

Conclusions

- Both models learn the vowel systems of Quichua and Spanish when trained on monolingual data
- When exposed to bilingual data, both models distinguish different numbers of categories depending on when L2 (Spanish) is introduced, paralleling data from bilingual speakers with different ages of acquisition
- Age of exposure effect in the models arises as a consequence of learning without any changes in model plasticity (learning rates)
- Language users may be restricted in learning a second language not because of a critical period, but by the commitments that the system has already made to the first language; a separate mechanism is not necessary

Future directions

- GMM categories are sometimes unstable and with too much training data, the model may end with fewer categories than in the language (unlike human learners); further exploring the parameter space of the model may lead to a more stable set of parameters
- HNRN can be given additional types of input, such as context information or labels for different lexical items; could allow the network to maintain distinctions between more categories
- Look at whether it is more or less difficult to learn an L2 with more categories than L1 (Quichua->Spanish) or to learn one with fewer categories (Spanish->Quichua)

Acknowledgements

We would like to thank Bob McMurray for helpful discussions about these simulations. This work was supported by a postdoctoral fellowship awarded by the CRLMB to M.C. and NIH DC008089 awarded to Bob McMurray.

meghan.clayards@mcgill.ca, joseph-toscano@uiowa.edu