Graph alignment and cross-modal learning during early infancy

Andrew R. Plummer
The Ohio State University, Columbus, OH, USA

www: http://www.learningtotalk.org
email: plummer@ling.ohio-state.edu

Introduction

• Results of decades of research on vowels support the conclusion that perception and production of language-specific vowel categories cannot be based on invariant targets that are represented directly in either the auditory domain or the articulatory (sensorimotor) domain.

• For example, an infant and an adult female’s productions of the neutral vowel [e] differ when they are schematically represented (a) in the articulatory (sensorimotor) domain using VXLAM (visual and language-acoustic model) (left), (b) in the acoustic domain as points in the F1:F2 formant space (middle), or (c) in the auditory domain using ERB-transformed patterns (right).

• This raises a number of questions about how an infant can acquire the cognitive representations relevant for learning the vowels of the ambient language.

• Some models of acquisition assume a fixed auditory transform to normalize for talker vocal tract size (e.g., Callan et al., 2000), ignoring evidence that normalization must be culture-specific (e.g., Johnson, 2005).

• Others assume that learning can be based on statistical regularities solely within the auditory domain (e.g., Assmann and Neavey, 2008), ignoring evidence that articulatory experience also shapes vowel category learning (e.g., Kamen and Watson, 1991).

• More recent models assume that learning is based primarily on statistical regularities within both the auditory domain (e.g., Bulthoff et al., 2009, Ananthakrishnan and Savin, 2011) or articulatory domain (e.g., Raaijmakers et al., 2013), as reviewed by interaction with a caretaker, ignoring developmental complexities of internal representation of the interaction, including:

• the gradual development of representation of the self and others (e.g., Mead, 1909; Hsu et al., 2019);

• the creation of internal modal representations (Mateeff and Kuhl, 1994) and multisensory peripersonal knowing (Lewkowitz and Ghazanfar, 2009).

Cross-modal Learning as Graph Alignment

• We outline an alternative approach that models cross-modal learning using graph structures, called “mandalas,” which organize sensory information in the auditory and in the articulatory domain, so that information can be linked across the two domains via graph alignment.

• Graph alignment is guided by perceptual targets that are internalized in early infancy through social vocal interaction with caretakers, so that vowel categories (c) can be identified with the abstractness that mediates between the two domains during alignment, rather than with domain-internal representations (a), (b), or (e).

Corresponding author: Andrew R. Plummer - Department of Linguistics - The Ohio State University - Columbus, Ohio, USA

Internalization and Pairing Computations

• Internalization is the process by which sensorimotor coordinates are transferred to the auditory domain to establish vowel categories. Each of these processes is thought to involve standard algorithms (with different weights) for vowel classification.

Acoustic and Social Signals

• Goodness values are modeled using a statistical model of acoustic and social signals from infant-directed speech experiments (Bulthoff et al., 2009). The vowel stimuli were generated by the stochastic linear auditory model (SLAM), for each of six ages, including 6 months and 10 years (left).

• Each set of stimulus was extracted within speaker and articulatory domain (middle).

• Sensorimotor alignments are represented in the infant’s creation of a provisional representation of a social agent in the infant vocal learning environment.

Sensormotor Alignment Computations

• Forward (sensorimotor to auditory) and backward (auditory to sensorimotor) alignments are performed using the weighted sensorimotor pairings.

• The weighted sensorimotor pairings are generated by multiplying the goodness values of the acoustic model (Section A.4) with a speaker’s goodness value for the stimulus.

• This produces a weighted sensorimotor pair for each age (6 months to 10 years) that can be offset from the weighted internal model.

Multisensory Representational Output

• The response pairs are internalized by the infant to create a weighted internal model, which is used to generate a category response output (right).

• The weighted internal model pairs are then used to create weighted sensorimotor pairings, which are input into the auditory domain.

• Multisensory forward and backward calculations are performed using weighted internal models, which are input into the auditory domain.

Summary

• On our approach vowel normalization is a generative procedure, rather than a reductive innateness computation or statistical summary.

• Acoustic and social signals derived from interaction with a caretaker provide the raw material for the normalization computation.

• Auditory representations are computed over the acoustic and social signals, providing the input for sensorimotor alignment and the computation of rich representations of the self and the caretaker.

• Higher-order intermodal representations of the self and caretaker are computed from the sensorimotor alignments, which reflect multisensory perceptual narrowing.

• The intermodal representations are then aligned to yield a commensuration structure that provides the basis for vowel categorization, and other cognitive computations.

Acknowledgments

The perceptual categorization data are from a study by Benjamin Munson, using stimuli provided by Lucie Michaud and texts recorded by Catherine McBride-Chang, Changlee Mays, Asimina Synta, Kiyoko Yoneyama, and Hyunjoo Chung. Work supported by NSF grants BCS 0728277 (to Benjamin Munson) and BCS 0728106 (to Mark Badzak).

www: http://www.learningtotalk.org
email: plummer@ling.ohio-state.edu