Joint word segmentation and phonetic category induction

Micha Elsner, Stephanie Antetomaso and Naomi H. Feldman

The Ohio State University
University of Maryland

Early language learning: why is it hard?

Many sources of variability:

Different pronunciations (you vs ya)
Sounds vary in context (ɪ as [ɪn] vs [ɪt])
Overlap in phonetic categories (æ vs ɛ)
Uncertain lexicon; weak top-down signal

Two modeling approaches

Unsupervised speech recognition

Lee et al 2015, Jansen and Church 2011, Varadarajan et al 2008

Facing full complexity of natural data...
These models learn too many sound categories
Do these capture contextual variants?
Or result from other shortcomings of the model?

Controlled cognitive models

Daland and Pierrehumbert 2011, Rytting et al 2010, Neubig et al 2010,

Limited datasets and tasks distinguish effects of different kinds of variability
Previous work: can learn phones without contextual variation... given known word boundaries

What if we take away these boundaries?

Segmentation and vowel clustering

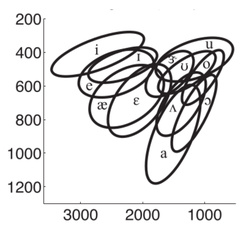
Categorical consonants, continuous vowels [F1, F2]

y[380.53 1251.69]w[811.88 1431.96]nt[532.91 1094.14]

"you want to"

⬇

yu • wɑn • tu

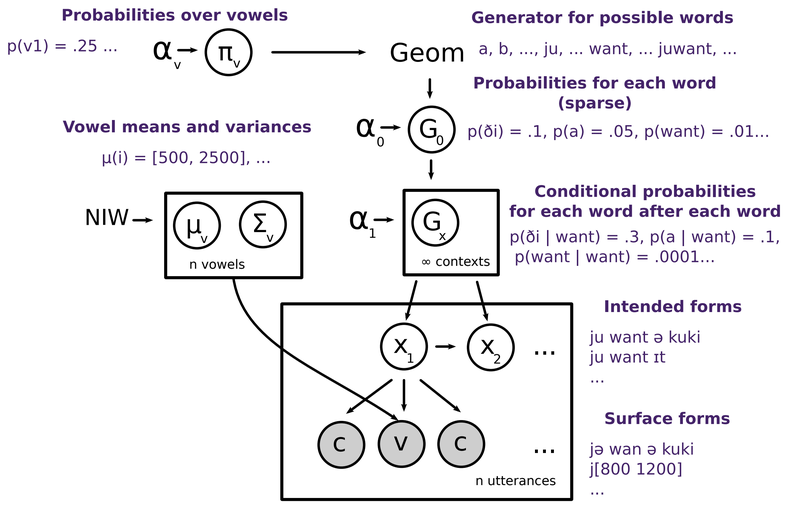


Word strings: Brent 1999
Vowel formants: Hillenbrand et al 1995 lab dataset

Dataset vowels vary, but not contextually!

Model architecture

Follows Feldman et al 2013, Elsner et al 2013, Goldwater et al 2009



Inference

Uses finite-state transducer encoding and beam sampling van Gael et al 2008, Huggins and Wood 2014
Standard annealing schedule for convergence
Plus **block moves** to reanalyze vowels in lexical entries

Number of components mixes very poorly

Runs here used fixed number of vowels
Posterior probabilities peak near correct n=12, but infinite-mixture samplers don't find this solution

Analysis

Joint segmentation/recognition errors relatively few, phonologically implausible:

milk: me + lk; sit: say + t; should I: shoe + d + l

Errors of this type probably uncommon for real infants
Harder with multiple languages or dialects?

Task scores

	Segmentation			Vowels
	P	R	F	F
Goldwater DPSEG	76	72	74	
Feldman LexDist				76
Our joint sampler, n=12	64	69	67	83

Changes relatively small...
Segmentation scores drop somewhat...
Vowel categorization improves a bit (perhaps due to bigram model)

Vowel confusion

		X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
ɑ	aa	1530	0	73	51	1	1	0	0	0	6	0	1
æ	ae	0	2581	1	0	251	10	10	14	0	0	4	1
ə	ah	88	1	6760	403	4	10	0	0	0	104	0	14
ɔ	ao	46	0	5	1043	0	0	0	0	0	43	0	2
ɛ	eh	0	50	5	0	2459	17	7	3	0	0	3	1
ɚ	er	0	6	0	0	19	2012	3	1	0	0	7	56
eɪ	ey	0	60	0	0	8	8	1257	132	66	0	103	0
ɪ	ih	0	10	0	0	9	2	526	4182	24	0	1197	0
i	iy	0	0	0	0	0	0	7	27	3802	0	10	0
oʊ	ow	1	0	39	13	0	1	0	0	0	1951	0	17
ʊ	uh	1	0	2	2	0	13	0	0	0	7	0	1353
u	uw	0	0	2	0	0	33	0	0	0	44	1	2947

Interpreting results

Categorization: phonetic category overlap doesn't matter much...
Context-sensitive variation is the major problem
Still working on separating contributions of pronunciation variation and coarticulatory processes

Funded by NSF grants 1422987 and 1421695. We thank Sharon Goldwater for her advice and software.