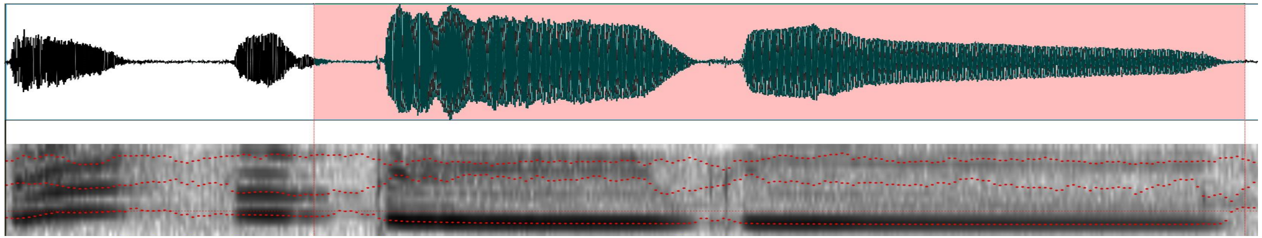
Speech segmentation with a neural encoder model of working memory

Micha Elsner and Cory Shain

The Ohio State University

What is unsupervised segmentation?

you**want**toseethebook lookthere’saboywithhishat andadoggie
you**want**tolookatthis lookatthis haveadrink takeitout you**want**itin putthaton that
yes okay openitup takethedoggieout ithinkitwillcomeout what daddy



- The infant hears a stream of utterances
- And has to pick out lexical units

What can the infant do?

- Learn some words as early as 6 months (Bergelson+ 12)
- Rarely produce partial words, but do run words together (Peters 83)
- Distinguish function words from non-words by 12 months (Shi+ 06)

"Word knowledge" in this sense may be very partial and incomplete

Models of word segmentation

- Phonotactic: Fleck 08, Rytting+ 07, Daland+ 11 and others
  Track transitional probabilities between phones
- Bayesian: Brent 98, Goldwater+ 09, Boerschinger+ 14 and others
  Balance predictive power with innate bias against rare words
- Feature-based unigram: Berg-Kirkpatrick+ 10
  Generative maxent model with features like #vowels per word
- Process-oriented: Lignos+ 11
  Subtractive segmentation removes known words from beginning of utterance

Hard to adapt these to speech

Separately trained acoustic units:

- External phone recognizer: de Marcken 96, Rytting 07 and others
- Hybrid neural-Bayesian: Kamper+ 16

Learn their own acoustics, but less flexible:

- Gaussian-HMMs: Lee+ 12, 15, see also Jansen 11
- Syllable discovery and clustering: Räsänen 15

Our model

Audio *or* character-based input
Multilevel autoencoder
Constrained by memory capacity
(*But not state-of-the-art results)

Why a new model?

- Explain learning biases using memory mechanism
  - Links biases in previous work to memory
  - Lower-level basis for Bayesian “small lexicon”-type priors?
  - “Phonological loop” (Baddeley+ 74) as modeling device
- Cope with variable input
- Explore unsupervised learning in neural framework

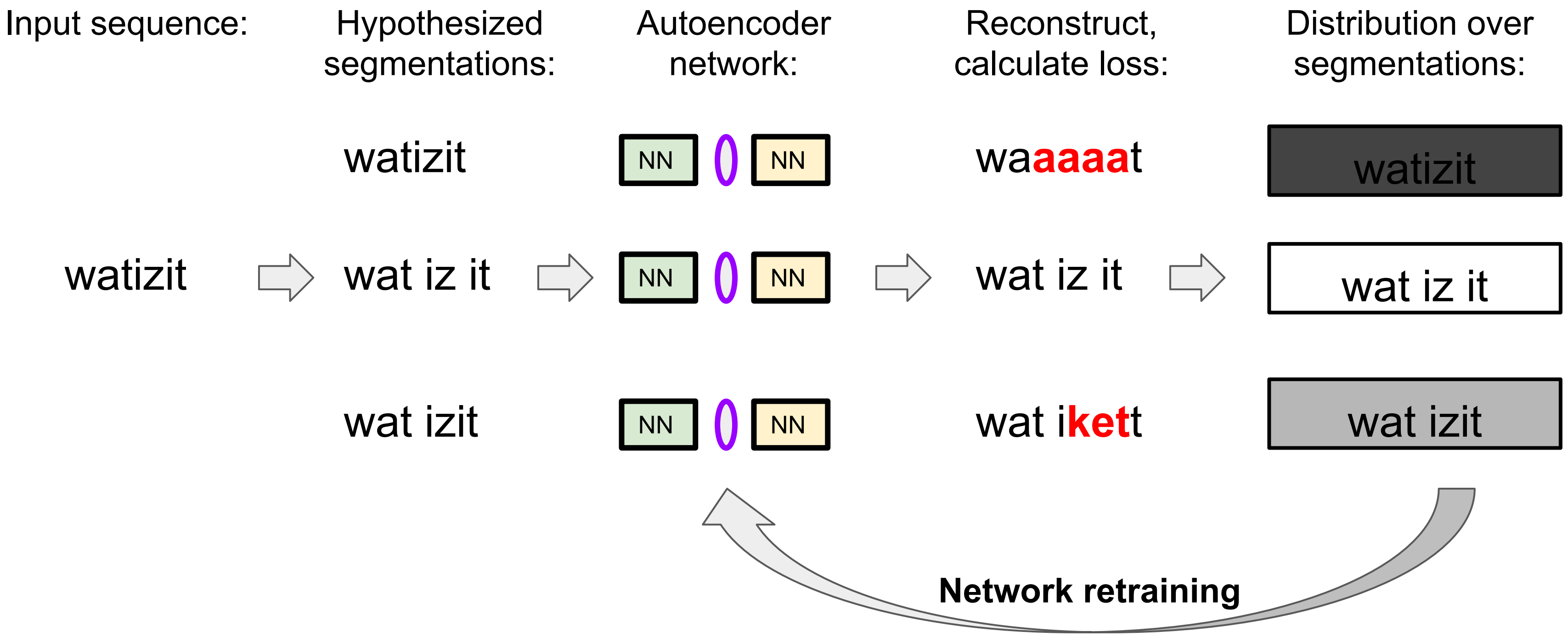
Why a new model?

- Explain learning biases using memory mechanism
- Cope with variable input
    - No need for a separate phone recognizer
    - Neural nets can extract features from audio
    - Latent numeric word representations robustly represent variation
- Explore unsupervised learning in neural framework

Why a new model?

- Explain learning biases using memory mechanism
- Cope with variable input
- Explore unsupervised learning in neural framework
  - Modern neural net technology still isn’t dominant in unsupervised learning
  - Previous neural segmenters (Elman 90, Christiansen+ 98, Rytting+ 07) use distant supervision/SRNs
  - Other current efforts (Kamper+ 16) use hybrid neural-Bayesian mechanisms
  - We use autoencoders (cf. Socher’s latent tree models)
    - Another new model (Chung+ 17) use latent neural segmentation for different tasks

Idea: words are chunks you can remember

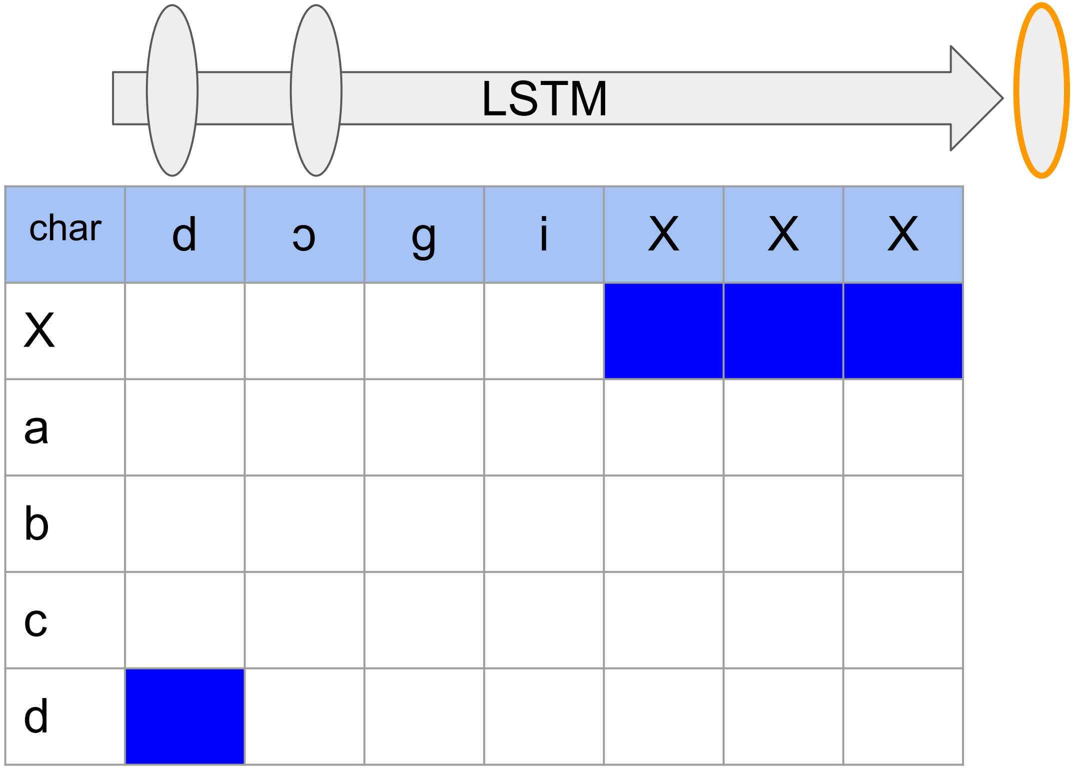


Key ideas:

- Autoencoder doesn’t predict segmentation directly
    - But provides a **loss function** for segmentation
- Need different imperfect reconstructions based on segmentation
    - Due to limited memory capacity
    - Model shouldn’t be at ceiling
- Assumption: **real words** are easier to remember

see Cho+ 14, Vinyals+ 15, etc.

Model part 1: phonological encoding

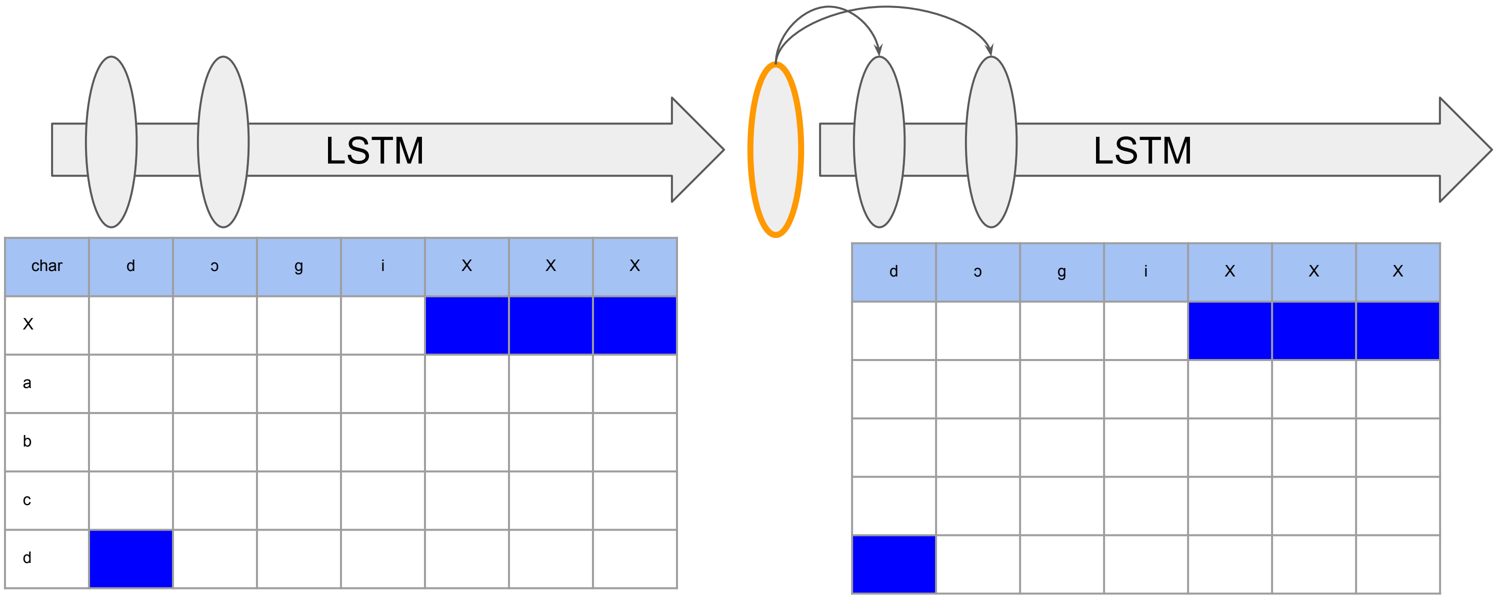


w-dimensional latent word representation

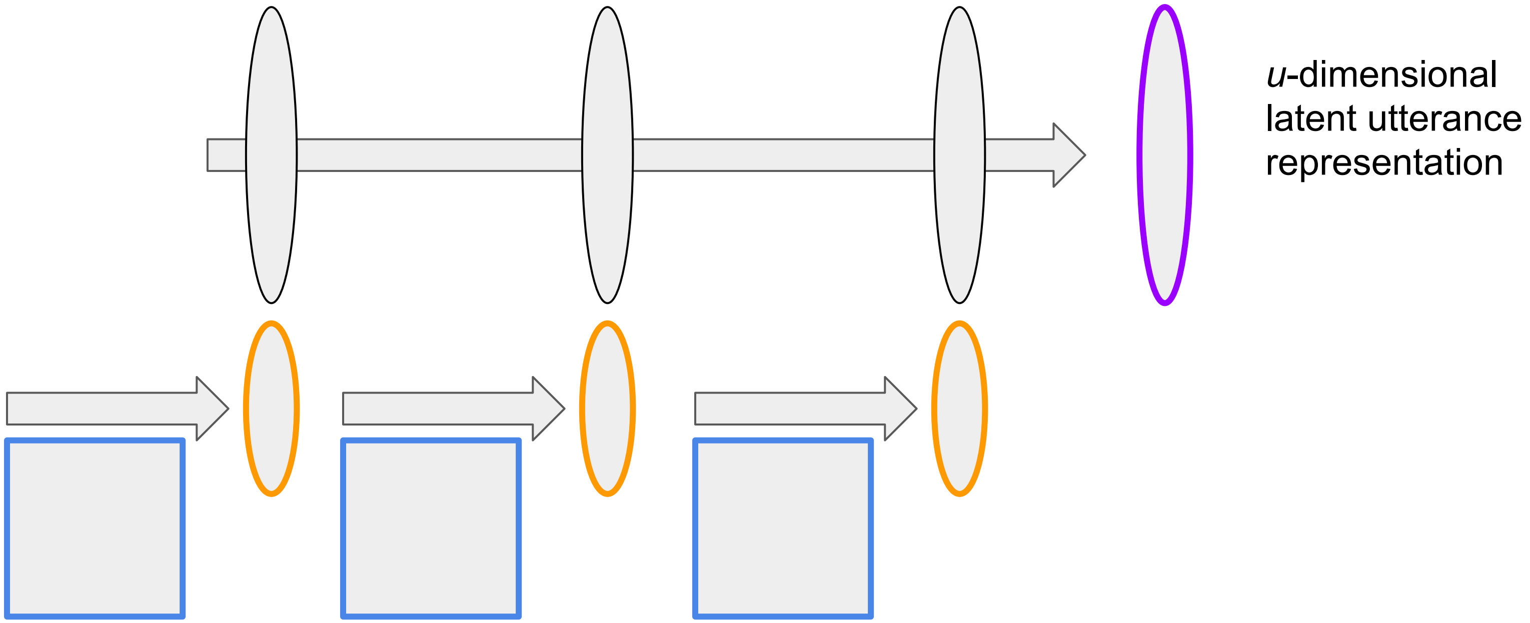
one-hot characters / MFCCs for each frame

Fixed-length with padding

Model part 1: phonological encoder-decoder

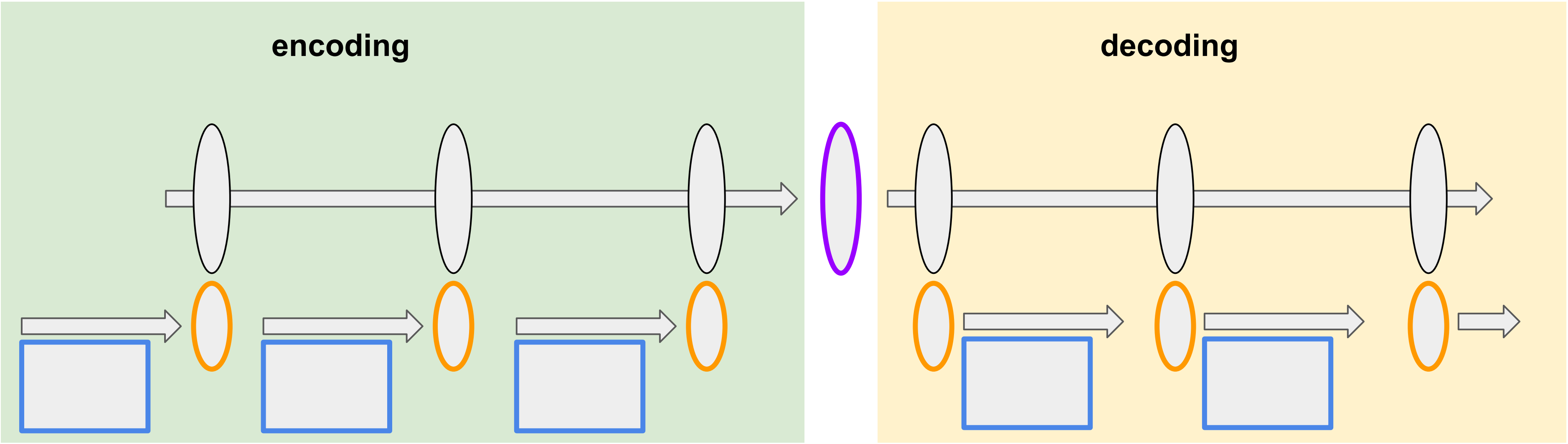


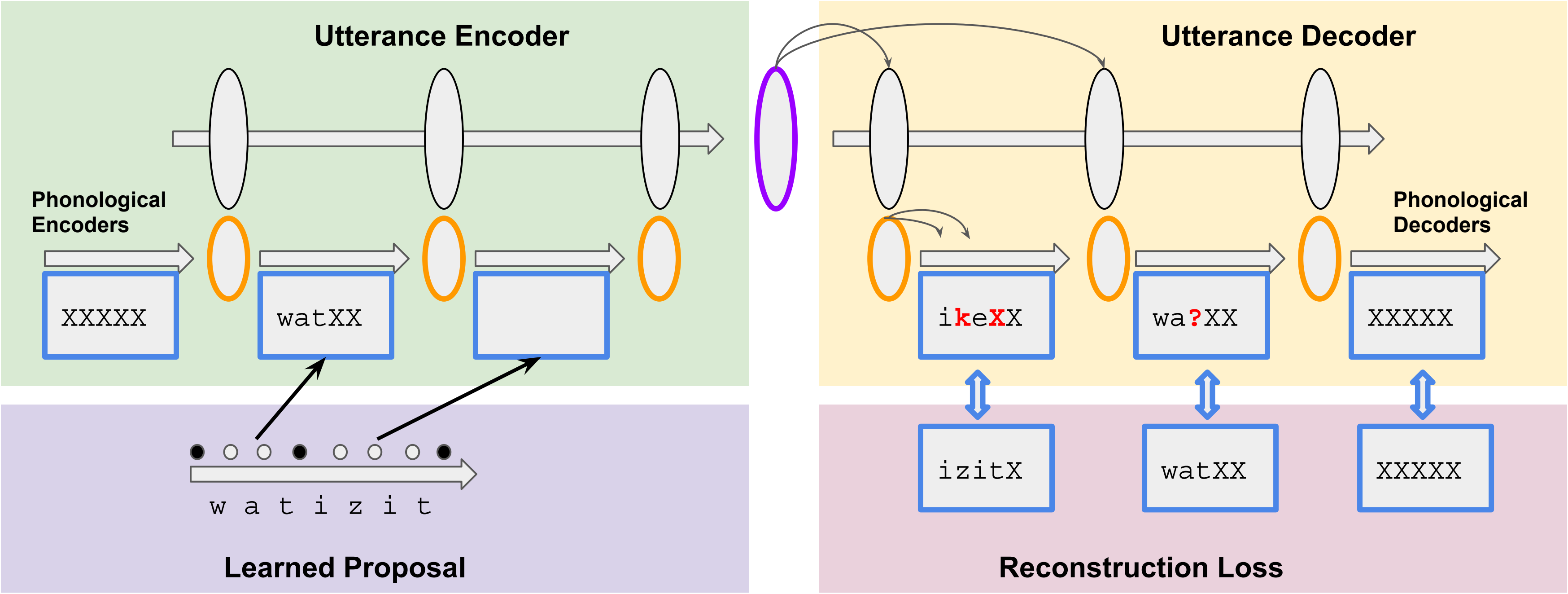
Model part 2: utterance encoding



Model part 2: utterance encoder-decoder

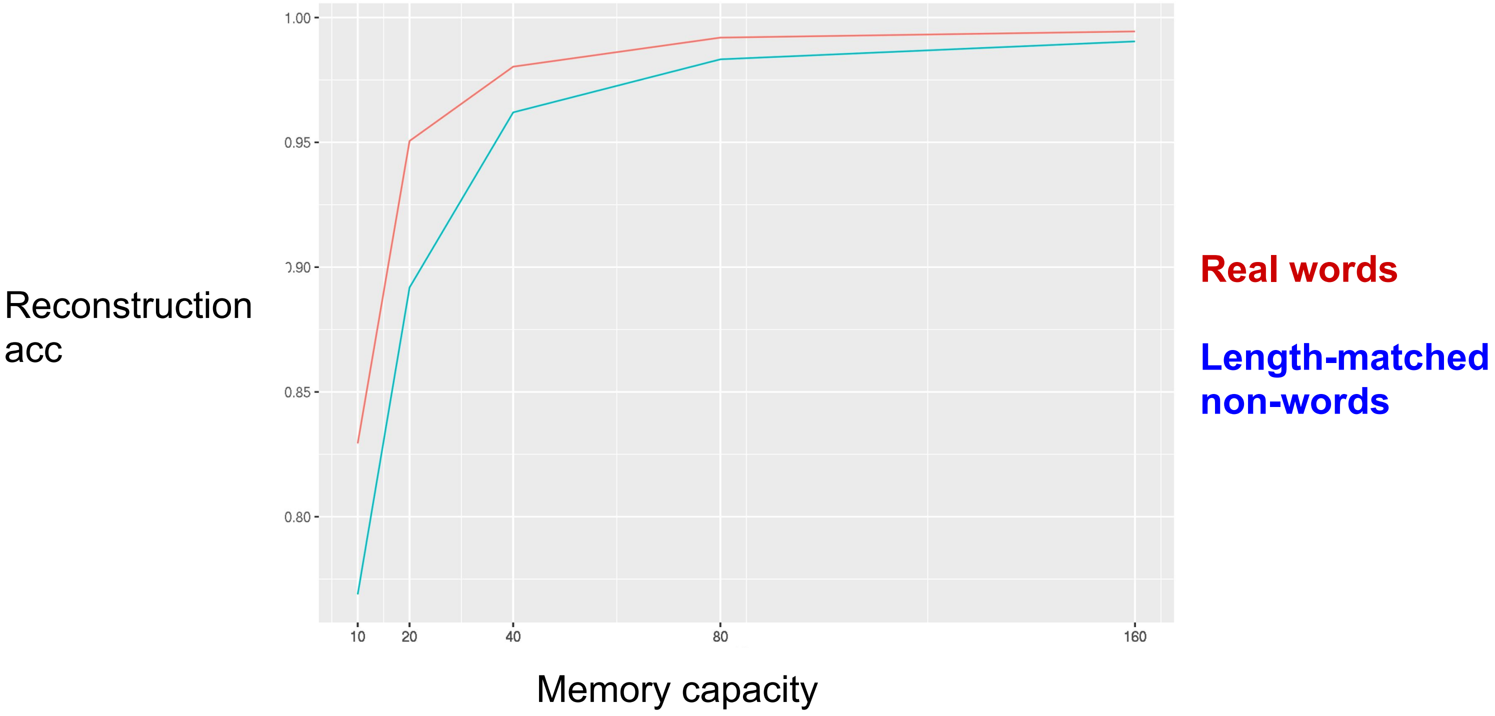
Autoencoder loss: reconstruction of the original sequence




Utterance Encoder
Phonological Encoders
XXXXX
watXX
w a t i z i t
Learned Proposal
Utterance Decoder
Phonological Decoders
ikeXX
wa?XX
XXXXX
izitX
watXX
XXXXX
Reconstruction Loss

Real words are easier to memorize

(using the phonological network alone)



Cognitive architecture simulates memory

- Memory separated into **phonological** and **lexical** units
  - *Phonological loop* vs *episodic memory*
- Levels must work together to reconstruct the sequence
  - Utterance level wants few words with predictable order
  - Word level wants short words with phonotactic regularities…
- Balancing these demands leads to good segmentations

see Mnih+ 14 and others

Training: gradient estimates with sampling

Network gives reconstruction loss for any segmentation

Search the space of segmentations for good options

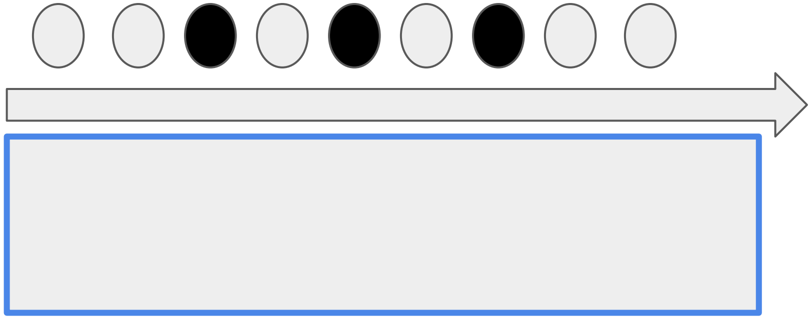
1. Sample some segmentations
2. Score them with the network
3. Compute importance weights
4. Sample posterior segmentation, update network parameters

Learn the proposal distribution

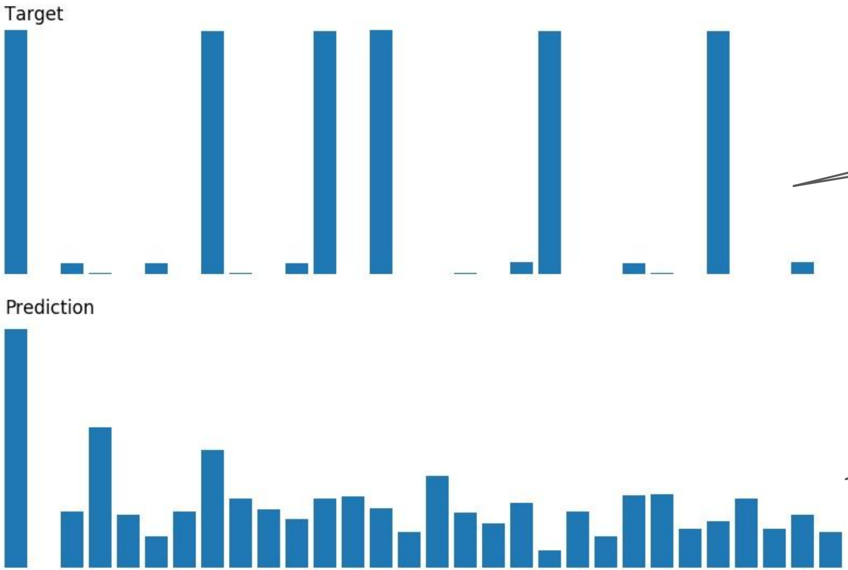
Train another LSTM on the whole sequence to produce the proposal:

WAtIzIt

W 7.6e-05 A 0.002 **t 0.30** I 0.004 **z 1.0** I 2.1e-05 **t 1.0** | X 6.9e-06



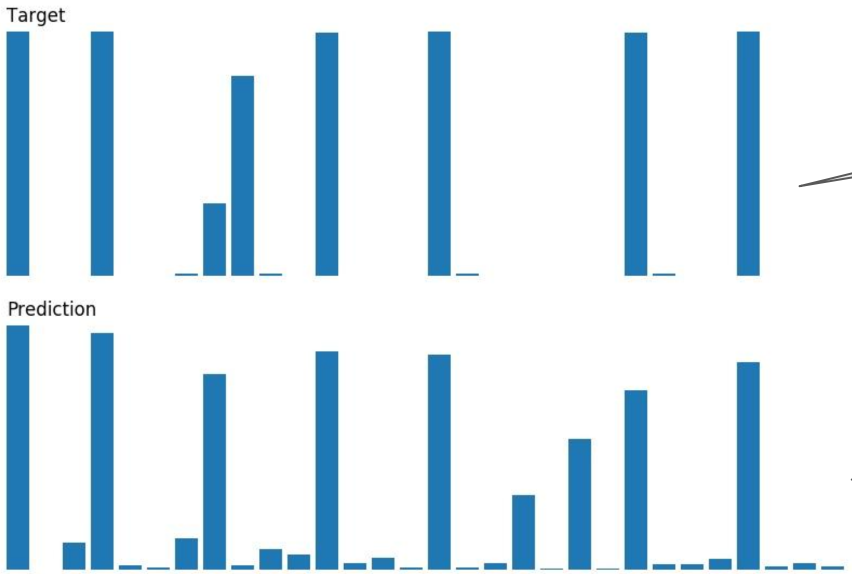
Increasing confidence over time: iteration 1



Distribution over segment boundaries after encode/decode

Proposed segment boundaries

Increasing confidence over time: iteration 12



Distribution over segment boundaries after encode/decode

Proposed segment boundaries

Characters (Brent 9k utterances)

Phonemically transcribed child-directed speech

	Breakpoint F	Token F
Goldwater bigrams	87	74
Johnson syllable-collocation		87
Berg-Kirkpatrick maxent		88
Fleck phonotatic	83	71
This work: neural	83	72

Our results: comparable to Fleck+ 08

Sample segmentations

yu want tu si **D6bUk**
lUk D*z **6b7** wIT hIz h&t
&nd 6**d Ogi**
yu want tu **lUk&t** DIs
lUk&t DIs
h&v 6**d rINk**
oke nQ
WAts DIs
WAts D&t
WAt Iz It

lUk k&n yu tek It Qt
tek It Qt
yu want It In
pUt D&t an
D&t
yEs
oke
op~ It Ap
tek D6 dOgi Qt
9**T INk** It wIl kAm Qt

Versteegh+ 15

Acoustic input: Zerospeech 2015

English casual conversation (also provides Xitsonga: future work!)
Important limitation: not child-directed

Few alterations from character mode…

- Dense input: MFCCs, deltas, double-deltas
- Mean squared error loss function
- No utterance boundaries (some hacky estimates)
- Initial proposal from voice activity detection
- Simplified one-best sampling (ask later!)

Acoustics (Zerospeech ‘15 English)

	Breakpoint F	**Token F**
Lyzinski+ 15	29	2
Räsänen+ 15	47	10
Räsänen+ 15 (corrected)	55	12
Kamper+ 16	62	21
This work	51	10

Our results: comparable to Räsänen et al

Conclusions

- Unsupervised neural model for character and acoustic input
- Performance driven by memory limitations
- Supports cognitive theories of memory-driven learning

Future work

- Search problems: importance sampling is bad!
- Better architecture: beyond frame-by-frame LSTMs
- More levels of representation, more tasks
  - Phones vs words
  - Clustering and grounding representations
- Multilingual (Xitsonga and others)

Thank you!

Thanks also to OSU Clippers, Mark Pitt and Sharon Goldwater for comments.
This work was supported by NSF 1422987.
Computational resources provided by the Ohio Supercomputer Center and NVIDIA corporation.

Memory

Working memory has multiple components:

- *Phonological loop:* limited recall of acoustics (nonword repetition)
- *Episodic memory:* syntactic/semantic encoding

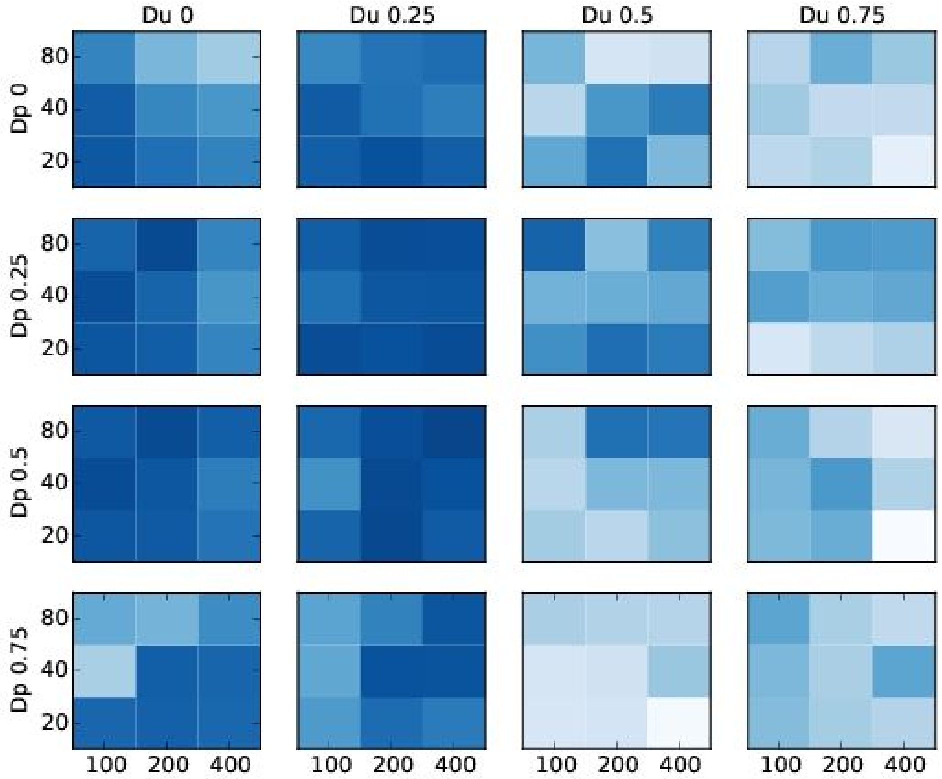
Baddeley+ (98): phonological loop is critical for word learning
Ability to remember plausible non-words correlates with vocabulary

As in our model, words that are hard to remember are harder to learn

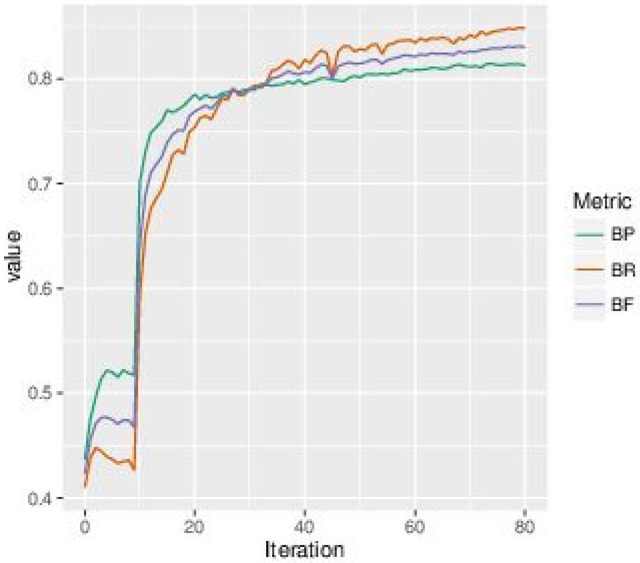
Annoying technical details

- Memory capacity and dropout:
  - Two **capacity** parameters (character and word)
  - Two **dropout** layers (delete characters and words)
- Fixed-length padding (for implementational tractability):
  - Requires an estimate of number of words per utterance
- Some additional parameters:
  - Penalty for one-letter words; otherwise lexical layer can learn phonology
  - Penalty for deleting chars by creating super-long words; functions as a max word length

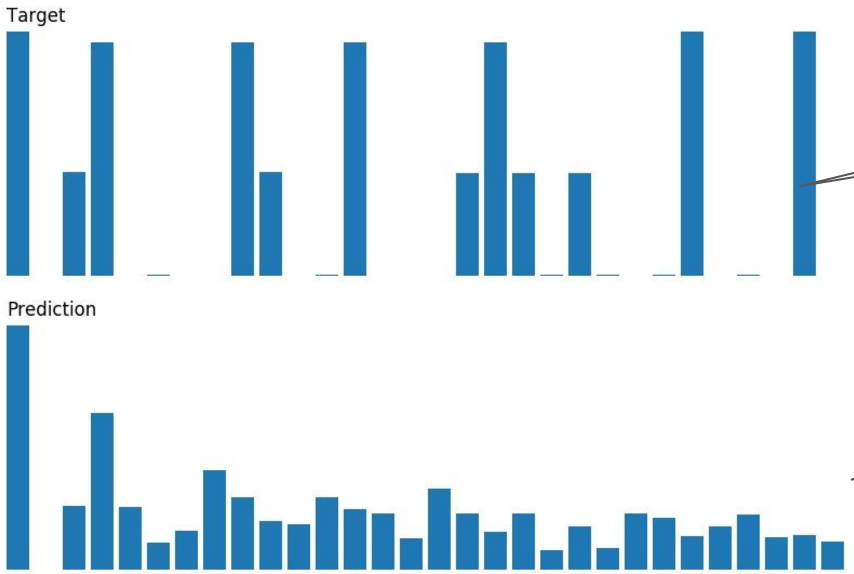
Tuning on Brent



Learning curves



Increasing confidence over time: iteration 4



Distribution over segment boundaries after encode/decode

Proposed segment boundaries

Increasing confidence over time: iteration 8