Acquiring language from speech by learning to remember and predict

Cory Shain and Micha Elsner

CoNLL 2020
Two representations of the same utterance.
Two representations of the same utterance.
Two representations of the same utterance.

I just love those koala...
Two representations of the same utterance.

I just love those koala...

Which is easier to remember?
Two representations of the same utterance.

I just love those koala...

Which is easier to remember?
Which makes it easier to predict the next sound? (It’s b.)
Two hypotheses:

1. Memory Hypothesis (MH): Learning to remember speech → Learning language (e.g. Baddeley et al. 1998)
2. Prediction Hypothesis (PH): Learning to predict speech → Learning language (e.g. Apfelbaum and McMurray 2017)
Two hypotheses:

**Memory Hypothesis (MH):**
Learning to remember speech → Learning language

(e.g. Baddeley et al. 1998)
Two hypotheses:

**Memory Hypothesis (MH):**
Learning to remember speech → Learning language
(e.g. Baddeley et al. 1998)

**Prediction Hypothesis (PH):**
Learning to predict speech → Learning language
(e.g. Apfelbaum and McMurray 2017)
This Study
I just love those koalas. Do you know them? What about learning?
I just love those koalas.

That's processing. What about learning?
I just love those koala bears. That's processing. What about learning?
I just love those koalas!
I just love those koalas. That's processing. What about learning?
I just love those koala bears. What about learning?
I just love those koalas. That's processing. What about learning?
I just love those koalas.
I just love those koalas.
I just love those koalas.
I just love those koalas.

Memory Hypothesis
I just love those koalas.
I just love those koalas
I just love those koalas.
I just love those koalas.
I just love those koalas.
Q: Do language-like representations emerge from these objectives?
I just love those koalas.

Our unsupervised ANN speech processor mimics this structure.
I just love those koalas.
I just love those koala...  

Evaluate effect on representations
I just love those koalas.

Preview: Our study supports both pressures.
Model
input

time
Data

+ **Zerospeech 2015 challenge**  
  *(Versteegh et al. 2015)*
  + English (~6.5 hrs)  
  + Xitsonga (~2.5 hrs)
Data

+ Zerospeech 2015 challenge
  (Versteegh et al. 2015)
  + English (~6.5 hrs)
  + Xitsonga (~2.5 hrs)
Data

- Zerospeech 2015 challenge
  (Versteegh et al. 2015)
  - English (~6.5 hrs)
  - Xitsonga (~2.5 hrs)
Evaluation

- Phoneme segmentation (boundary P/R/F)
- Phoneme labeling (probe P/R/F)
- Phonological feature (e.g. ± voice) labeling (probe P/R/F)
- Only $L_1$ used for eval (systematically better on dev)
Evaluation

- Phoneme segmentation (boundary P/R/F)
- Phoneme labeling (probe P/R/F)
- Phonological feature (e.g. ± voice) labeling (probe P/R/F)
- Only $L_1$ used for eval (systematically better on dev)
Evaluation

+ Phoneme segmentation (boundary P/R/F)
+ Phoneme labeling (probe P/R/F)
+ Phonological feature (e.g. $\pm$ voice) labeling (probe P/R/F)
+ Only $L_1$ used for eval (systematically better on dev)
Evaluation

- Phoneme segmentation (boundary P/R/F)
- Phoneme labeling (probe P/R/F)
- Phonological feature (e.g. ± voice) labeling (probe P/R/F)
- Only $L_1$ used for eval (systematically better on dev)
Experimental Conditions

- **Memory**: $B = 0, 5, 25, 50$
- **Prediction**: $F = 0, 1, 5, 10$
- **Depth**: $L = 2, 3, 4$
Experimental Conditions

- **Memory:** $B = 0, 5, 25, 50$
- **Prediction:** $F = 0, 1, 5, 10$
- **Depth:** $L = 2, 3, 4$
Experimental Conditions

+ **Memory**: $B = 0, 5, 25, 50$
+ **Prediction**: $F = 0, 1, 5, 10$
+ **Depth**: $L = 2, 3, 4$
Results
Boundary F (English)
Similar cross-linguistic patterns
Similar cross-metric patterns
Memory ($B > 0$), prediction ($F > 0$) and depth ($L > 2$) generally help
Significant effects on performance of
- Memory ($p < 0.006$)
- Prediction ($p < 0.001$)
- Multiscale encoding ($p < 0.001$)
In a cognitively constrained speech processing model, memory and prediction pressures support phoneme acquisition in complementary ways.
In a cognitively constrained speech processing model, memory and prediction pressures support phoneme acquisition in complementary ways.

See paper for add’l analyses:

+ Plausibility wrt human echoic memory limits
+ Controls for effects of inductive bias
+ Effect of memory and prediction on boundary P/R trade-off
Thank you!

https://github.com/coryshain/dnnseg

Funded in part by a Google Faculty Research Award to Micha Elsner.
Apfelbaum, Keith S and Bob McMurray (2017). “Learning during processing: Word learning doesn’t wait for word recognition to finish”. In: *Cognitive science* 41, pp. 706–747.

Baddeley, Alan, Susan Gathercole, and Costanza Papagno (Jan. 1998). “The Phonological Loop as a Language Learning Device”. In: *Psychological Review* 105.1, pp. 158–173.

Lee, Chia-ying and James Glass (2012). “A Nonparametric {Bayesian} Approach to Acoustic Model Discovery”. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp. 40–49.

Michel, Paul et al. (2017). “Blind Phoneme Segmentation With Temporal Prediction Errors”. In: Proceedings of ACL 2017, Student Research Workshop, pp. 62–68.

Versteegh, Maarten et al. (2015). “The zero resource speech challenge 2015”. In: Sixteenth Annual Conference of the International Speech Communication Association.
Baselines

- **U**: (untrained) architecturally matched untrained model
- **X**: (cross-language) architecturally matched model trained on opposite language
Baselines

- **U**: (untrained) architecturally matched untrained model
- **X**: (cross-language) architecturally matched model trained on opposite language
### Results: Best Dev Model

| Model      | English |     |     | Xitsonga |     |     |
|------------|---------|-----|-----|----------|-----|-----|
|            | Bd      | Pc  | Fc  | Bd       | Pc  | Fc  |
| Full       | 65.3    | 22.9| 49.3| 39.3     | 28.6| 53.8|
| Baseline U | 30.4    | 12.3| 42.2| 22.1     | 15.4| 46.2|
| Baseline X | 52.4    | 20.5| 47.1| **44.8** | 27.8| 53.2|

Condition: $B = 25$, $F = 1$, $L = 3$
Results: Effect of Learning (vs. Baseline U)

Boundary F

Phoneme F

Feature F
Consistent improvements with memory ($B > 0$), prediction ($F > 0$), and depth ($L > 2$)
Similar pattern to overall, suggesting that gains are driven by learning
Results: Effect of Language (vs. Baseline X)

Boundary F

Phoneme F

Feature F
Speech in one language highly informative about speech in another
Consistent improvements with memory, prediction, and depth ($L > 2$), except Bd in X
Results: Analysis

+ Memory and prediction impose competing pressures
Results: Analysis

Boundary P

Boundary R
**Results: Analysis**

\[ B = 0: \text{Higher recall, lower precision} \]
Results: Analysis

$F = 0$: Higher precision, lower recall
## Results: Hypothesis Tests

| Predictor                | $\beta$ | $t$   | $p$       |
|--------------------------|---------|-------|-----------|
| Intercept                | -1.22   | -7.73 | 3.89e-14***|
| Memory                   | 0.247   | 2.75  | 0.006**   |
| Prediction               | 0.959   | 9.86  | 2.0e-16***|
| Multiscale               | 0.305   | 4.10  | 4.58e-5***|
| Comparison=Full          | 0.037   | 0.453 | 0.651     |
| Comparison=BaselineX     | -0.064  | -0.709| 0.479     |
| Metric=Phoneme           | 0.021   | 0.240 | 0.810     |
| Metric=Feature           | 0.022   | 0.250 | 0.803     |

**Linear Regression**
No word evaluation.

- None of our designs improved on dumb word seg baseline
- Got some ideas, future work...
Baselines

- No SOTA comparison
  - Previous unsupervised phone segmenters use monolingual data
  - Unrealistically simple
    (Michel et al. 2017)
  - Unavailable
    (Lee and Glass 2012)