Measuring the perceptual availability of phonological features during language acquisition using unsupervised binary stochastic autoencoders

Cory Shain and Micha Elsner

NAACL 2019

Background

- How do infants acquire phonological categories and features from speech?
  - No explicit supervision
  - Poor lexical and phonotactic knowledge
- **Possible answer:** By trying to remember what they perceive.
  - Limited memory → compression pressure (Baddeley and Hitch 1974)
  - Might favor language-like representations (Baddeley et al. 1998; Elsner and Shain 2017)
  - Percept modeling can provide immediate training signal

Background

- How do infants acquire phonological categories and features from speech?
  - No explicit supervision
  - Poor lexical and phonotactic knowledge
- **Possible answer:** By trying to remember what they perceive.
  - Limited memory → compression pressure (Baddeley and Hitch 1974)
  - Might favor language-like representations (Baddeley et al. 1998; Elsner and Shain 2017)
  - Percept modeling can provide immediate training signal

Background

- How do infants acquire phonological categories and features from speech?
  - No explicit supervision
  - Poor lexical and phonotactic knowledge
- **Possible answer:** By trying to remember what they perceive.
  - Limited memory → compression pressure (Baddeley and Hitch 1974)
  - Might favor language-like representations (Baddeley et al. 1998; Elsner and Shain 2017)
  - Percept modeling can provide immediate training signal

Background

- How do infants acquire phonological categories and features from speech?
  - No explicit supervision
  - Poor lexical and phonotactic knowledge
- **Possible answer:** By trying to remember what they perceive.
  - Limited memory $\rightarrow$ compression pressure (Baddeley and Hitch 1974)
  - Might favor language-like representations (Baddeley et al. 1998; Elsner and Shain 2017)
  - Percept modeling can provide immediate training signal

Background

- How do infants acquire phonological categories and features from speech?
  - No explicit supervision
  - Poor lexical and phonotactic knowledge
- **Possible answer:** By trying to remember what they perceive.
  - Limited memory → compression pressure (Baddeley and Hitch 1974)
  - Might favor language-like representations (Baddeley et al. 1998; Elsner and Shain 2017)
  - Percept modeling can provide immediate training signal

Background

- How do infants acquire phonological categories and features from speech?
  - No explicit supervision
  - Poor lexical and phonotactic knowledge
- **Possible answer:** By trying to remember what they perceive.
  - Limited memory $\rightarrow$ compression pressure (Baddeley and Hitch 1974)
  - Might favor language-like representations (Baddeley et al. 1998; Elsner and Shain 2017)
  - Percept modeling can provide immediate training signal

Background

- How do infants acquire phonological categories and features from speech?
  - No explicit supervision
  - Poor lexical and phonotactic knowledge
- **Possible answer:** By trying to remember what they perceive.
  - Limited memory $\rightarrow$ compression pressure (Baddeley and Hitch 1974)
  - Might favor language-like representations (Baddeley et al. 1998; Elsner and Shain 2017)
  - Percept modeling can provide immediate training signal

Background

- Lots of evidence for top-down influence on phoneme acquisition (Peperkamp et al. 2006; Swingley 2009; Feldman et al. 2013)
- **But** bottom-up perceptual evidence must also matter
- How **perceptually available** are phoneme categories and phonological features?

Background

- Lots of evidence for top-down influence on phoneme acquisition (Peperkamp et al. 2006; Swingley 2009; Feldman et al. 2013)
- **But** bottom-up perceptual evidence must also matter
- How **perceptually available** are phoneme categories and phonological features?

Background

- Lots of evidence for top-down influence on phoneme acquisition (Peperkamp et al. 2006; Swingley 2009; Feldman et al. 2013)
- **But** bottom-up perceptual evidence must also matter
- How **perceptually available** are phoneme categories and phonological features?

This study

- **Hypothesis:** A learner trying to remember auditory properties of segments will discover theory-driven phonological categories and features.
- **Experiment:** Implement learner as computational model and inspect representations

This study

- **Hypothesis:** A learner trying to remember auditory properties of segments will discover theory-driven phonological categories and features.
- **Experiment:** Implement learner as computational model and inspect representations

Model

- Deep neural binary stochastic autoencoder network
- Compresses auditory features of speech segments into discrete 8-bit code
- Decompresses code into original inputs
- 256 categories with which to describe perceptual world
- Model optimizes fidelity
- **Research question:** Do the optimized representations look like infants' knowledge of language?

Model

- Deep neural binary stochastic autoencoder network
- Compresses auditory features of speech segments into discrete 8-bit code
- Decompresses code into original inputs
- 256 categories with which to describe perceptual world
- Model optimizes fidelity
- **Research question:** Do the optimized representations look like infants' knowledge of language?

Model

- Deep neural binary stochastic autoencoder network
- Compresses auditory features of speech segments into discrete 8-bit code
- Decompresses code into original inputs
- 256 categories with which to describe perceptual world
- Model optimizes fidelity
- **Research question:** Do the optimized representations look like infants' knowledge of language?

Model

- Deep neural binary stochastic autoencoder network
- Compresses auditory features of speech segments into discrete 8-bit code
- Decompresses code into original inputs
- 256 categories with which to describe perceptual world
- Model optimizes fidelity
- **Research question:** Do the optimized representations look like infants' knowledge of language?

Model

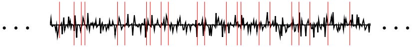
- Deep neural binary stochastic autoencoder network
- Compresses auditory features of speech segments into discrete 8-bit code
- Decompresses code into original inputs
- 256 categories with which to describe perceptual world
- Model optimizes fidelity
- **Research question:** Do the optimized representations look like infants' knowledge of language?

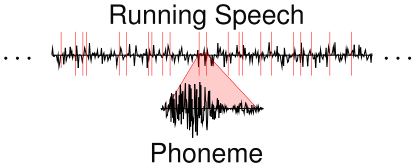
Model

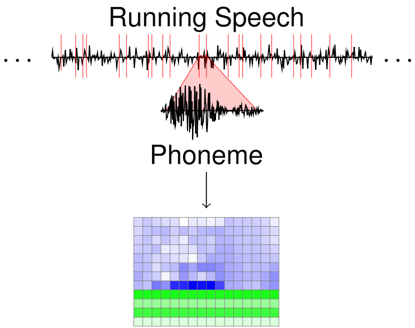
- Deep neural binary stochastic autoencoder network
- Compresses auditory features of speech segments into discrete 8-bit code
- Decompresses code into original inputs
- 256 categories with which to describe perceptual world
- Model optimizes fidelity
- **Research question:** Do the optimized representations look like infants' knowledge of language?

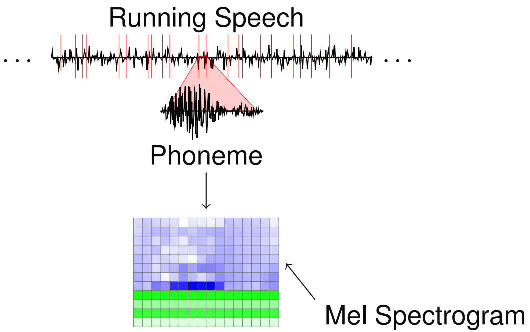
Running Speech

Running Speech

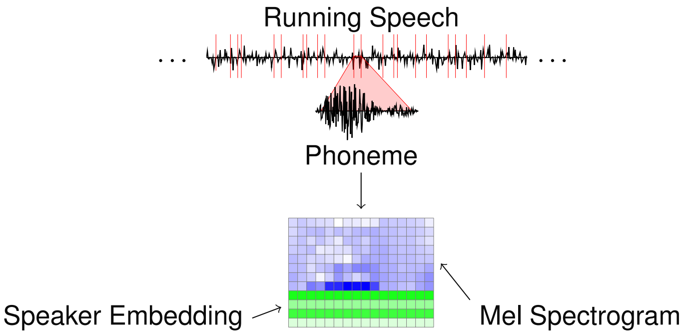


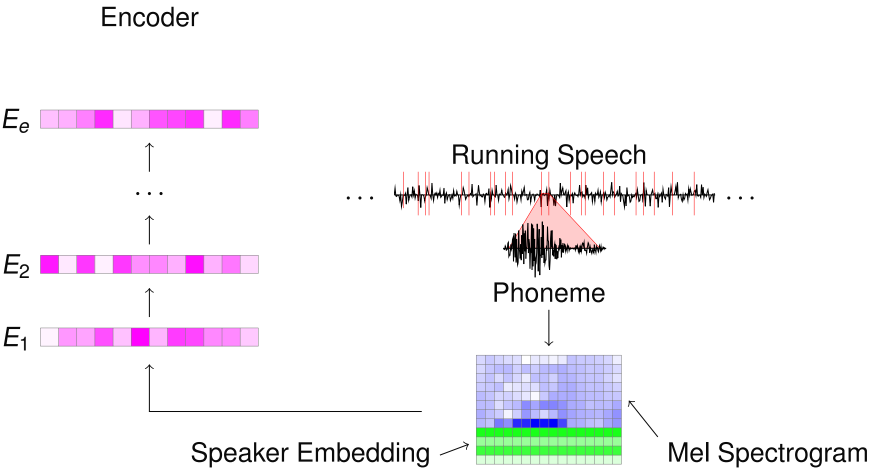

Running Speech
Phoneme


Running Speech
Phoneme

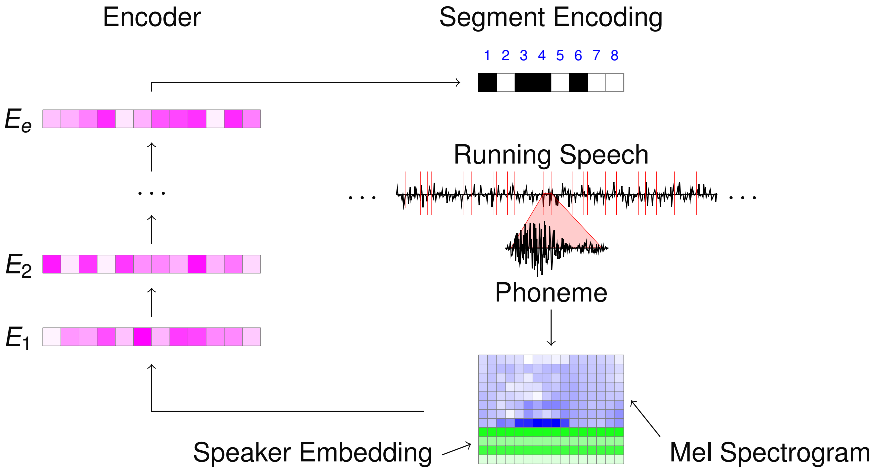


Running Speech
. . .
. . .
Phoneme
Mel Spectrogram

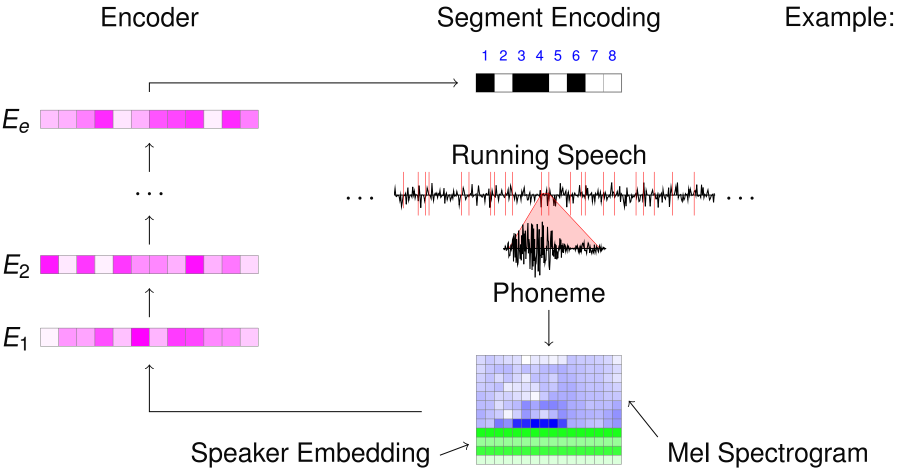

Running Speech
. . .
. . .
Phoneme
Speaker Embedding
Mel Spectrogram

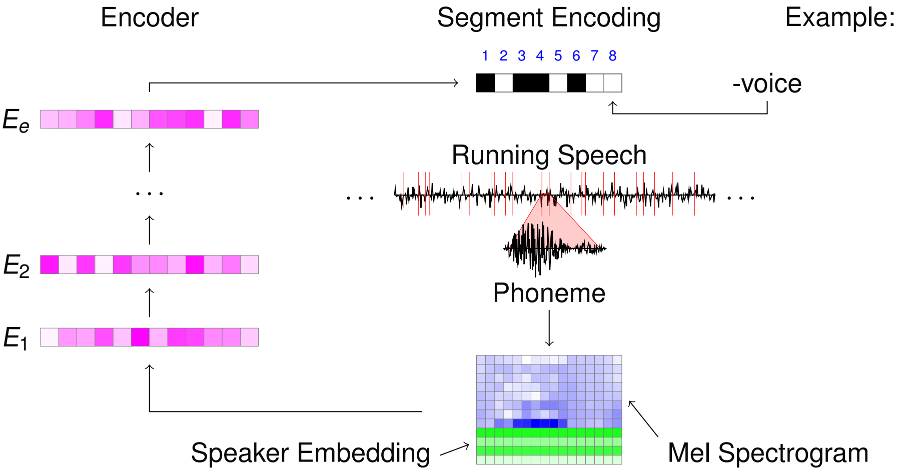


Encoder
E_e
E_2
E_1
Running Speech
Phoneme
Speaker Embedding
Mel Spectrogram

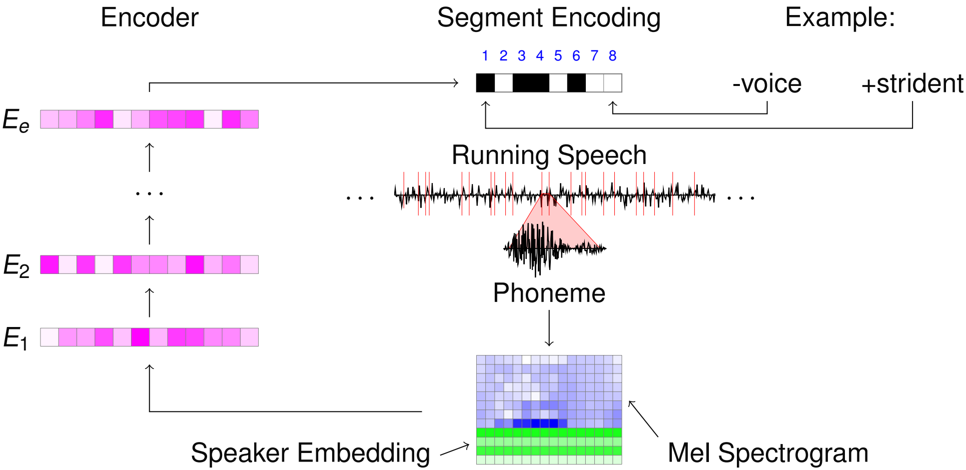


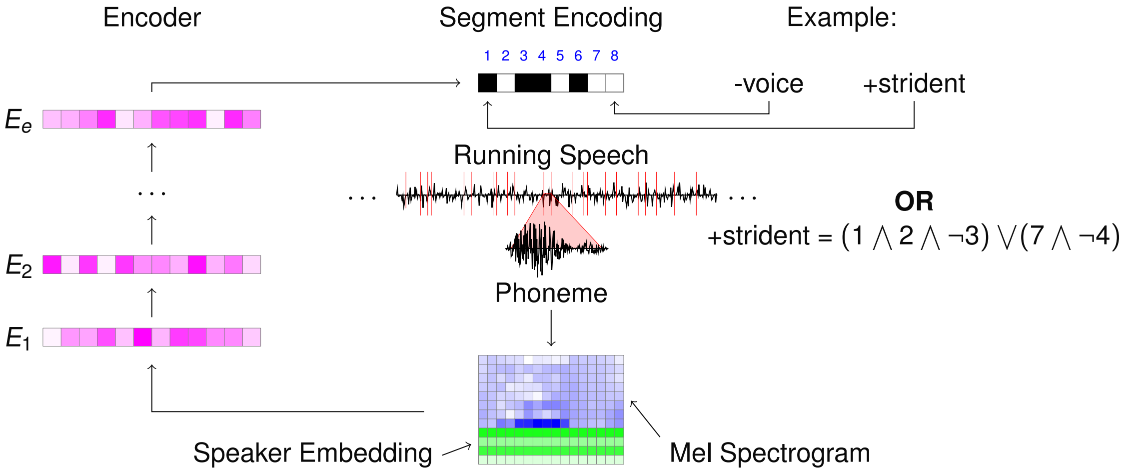
Encoder
Segment Encoding
1 2 3 4 5 6 7 8
E_e
Running Speech
. . .
. . .
. . .
E_2
Phoneme
E_1
Speaker Embedding
Mel Spectrogram


Encoder
Segment Encoding
Example:
1 2 3 4 5 6 7 8
E_e
Running Speech
. . .
. . .
. . .
E_2
Phoneme
E_1
Speaker Embedding
Mel Spectrogram

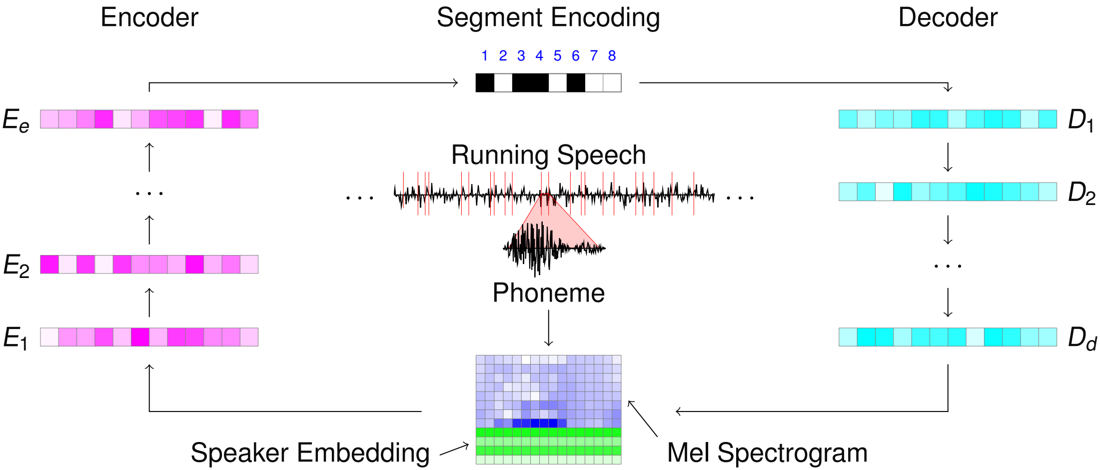


Encoder
Segment Encoding
Example:
1 2 3 4 5 6 7 8
-voice
E_e
Running Speech
. . .
Phoneme
E_2
E_1
Speaker Embedding
Mel Spectrogram


Encoder
Segment Encoding
Example:
1 2 3 4 5 6 7 8
-voice
+strident
E_e
Running Speech
E_2
E_1
Phoneme
Speaker Embedding
Mel Spectrogram



Encoder
Segment Encoding
Example:
1 2 3 4 5 6 7 8
-voice
+strident
E_e
Running Speech
. . .
OR
+strident = $(1 \wedge 2 \wedge \neg 3) \vee (7 \wedge \neg 4)$
E_2
Phoneme
E_1
Speaker Embedding
Mel Spectrogram



Encoder
Segment Encoding
Decoder
1 2 3 4 5 6 7 8
E_e
D_1
Running Speech
E_2
D_2
Phoneme
E_1
D_d
Speaker Embedding
Mel Spectrogram

Prior work

- Zerospeech challenge on unsupervised speech processing (Versteegh et al. 2015)
- Unsupervised phone discovery (Vallabha et al. 2007; Lee and Glass 2012; Feldman et al. 2013; Antetomaso et al. 2017)
- Word-level modeling
- Not featural (phonemes are atomic)

Prior work

- Zerospeech challenge on unsupervised speech processing (Versteegh et al. 2015)
- Unsupervised phone discovery (Vallabha et al. 2007; Lee and Glass 2012; Feldman et al. 2013; Antetomaso et al. 2017)
- Word-level modeling
- Not featural (phonemes are atomic)

Prior work

- Zerospeech challenge on unsupervised speech processing (Versteegh et al. 2015)
- Unsupervised phone discovery (Vallabha et al. 2007; Lee and Glass 2012; Feldman et al. 2013; Antetomaso et al. 2017)
- Word-level modeling
- Not featural (phonemes are atomic)

Prior work

- Zerospeech challenge on unsupervised speech processing (Versteegh et al. 2015)
- Unsupervised phone discovery (Vallabha et al. 2007; Lee and Glass 2012; Feldman et al. 2013; Antetomaso et al. 2017)
- Word-level modeling
- Not featural (phonemes are atomic)

Data

- Zerospeech 2015 challenge datasets
  - **Xitsonga:** ~2.5 hrs read speech, 24 speakers
  - **English:** ~5 hrs spontaneous speech, 12 speakers

Data

- Zerospeech 2015 challenge datasets
  - **Xitsonga:** ~2.5 hrs read speech, 24 speakers
  - **English:** ~5 hrs spontaneous speech, 12 speakers

Data

- Zerospeech 2015 challenge datasets
  - **Xitsonga:** ~2.5 hrs read speech, 24 speakers
  - **English:** ~5 hrs spontaneous speech, 12 speakers

Results: Unsupervised Phoneme Classification

- Homogeneity (H), Completeness (C), V-measure (V) (Rosenberg and Hirschberg 2007)

Results: Unsupervised Phoneme Classification

	Xitsonga			English		
Model	H	C	V	H	C	V
Baseline	0.023	0.013	0.016	0.006	0.004	0.005
-discrete,-speaker	0.281	0.191	0.227	0.246	0.166	0.198
-discrete,+speaker	0.302	0.185	0.230	0.205	0.180	0.192
+discrete,-speaker	0.360	0.206	0.262	0.240	0.161	0.193
Full model (+discrete,+speaker)	**0.462**	**0.268**	**0.339**	**0.270**	**0.180**	**0.216**

(H)omogeneity, (C)ompleteness, (V)-measure

Results: Unsupervised Phoneme Classification

	Xitsonga			English		
Model	H	C	V	H	C	V
Baseline	0.023	0.013	0.016	0.006	0.004	0.005
-discrete,-speaker	0.281	0.191	0.227	0.246	0.166	0.198
-discrete,+speaker	0.302	0.185	0.230	0.205	0.180	0.192
+discrete,-speaker	0.360	0.206	0.262	0.240	0.161	0.193
Full model (+discrete,+speaker)	**0.462**	**0.268**	**0.339**	**0.270**	**0.180**	**0.216**

Large relative improvement (20-40x) over random demonstrates clear learning signal.

(H)omogeneity, (C)ompleteness, (V)-measure

Results: Unsupervised Phoneme Classification

	Xitsonga			English		
Model	H	C	V	H	C	V
Baseline	0.023	0.013	0.016	0.006	0.004	0.005
-discrete,-speaker	0.281	0.191	0.227	0.246	0.166	0.198
-discrete,+speaker	0.302	0.185	0.230	0.205	0.180	0.192
+discrete,-speaker	0.360	0.206	0.262	0.240	0.161	0.193
Full model (+discrete,+speaker)	**0.462**	**0.268**	**0.339**	**0.270**	**0.180**	**0.216**

Large relative improvement (20-40x) over random demonstrates clear learning signal.
Speaker embeddings and binary neurons improve clustering.

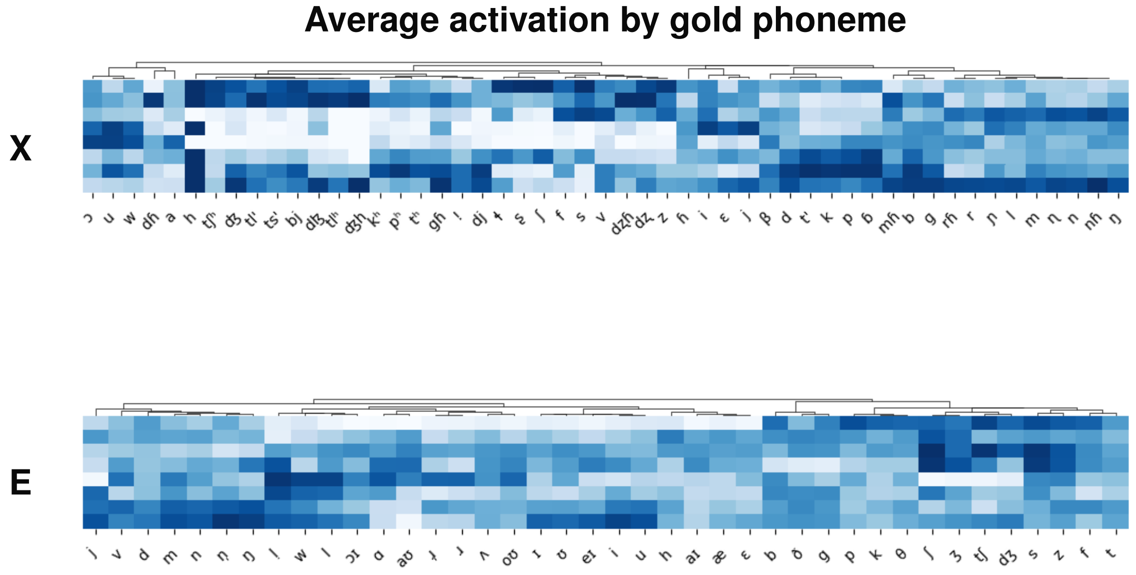
(H)omogeneity, (C)ompleteness, (V)-measure

Results: Unsupervised Phoneme Classification

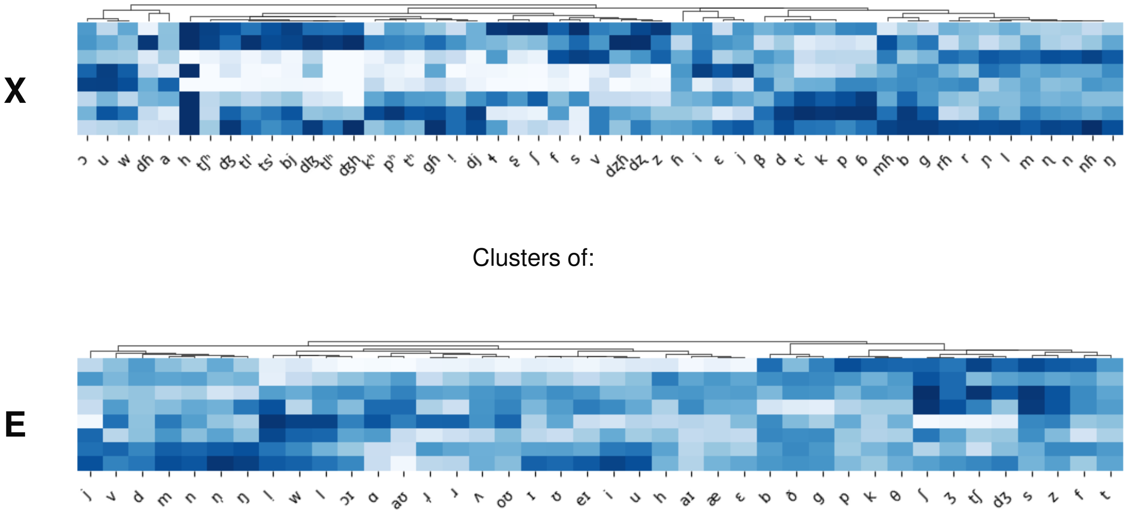
Model	Xitsonga H	Xitsonga C	Xitsonga V	English H	English C	English V
Baseline	0.023	0.013	0.016	0.006	0.004	0.005
-discrete,-speaker	0.281	0.191	0.227	0.246	0.166	0.198
-discrete,+speaker	0.302	0.185	0.230	0.205	0.180	0.192
+discrete,-speaker	0.360	0.206	0.262	0.240	0.161	0.193
Full model (+discrete,+speaker)	**0.462**	**0.268**	**0.339**	**0.270**	**0.180**	**0.216**

Large relative improvement (20-40x) over random demonstrates clear learning signal.
Speaker embeddings and binary neurons improve clustering.
Top-down guidance likely needed to refine representations.

(H)omogeneity, (C)ompleteness, (V)-measure

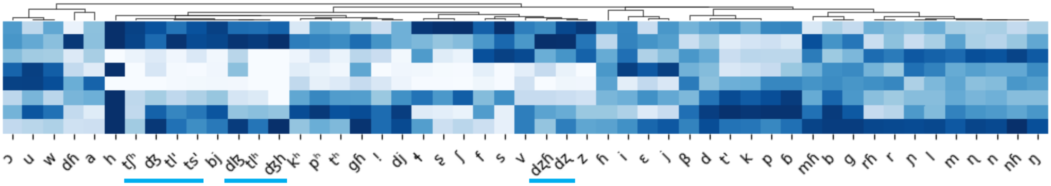

Average activation by gold phoneme
X
E

Average activation by gold phoneme



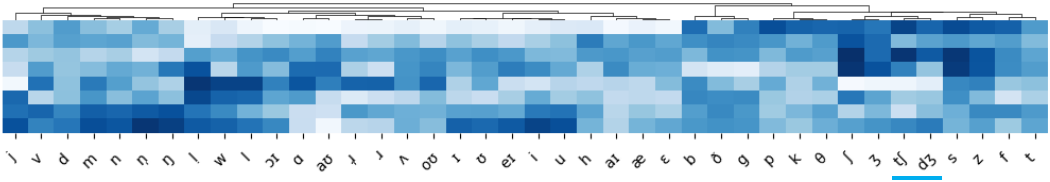
Average activation by gold phoneme

X



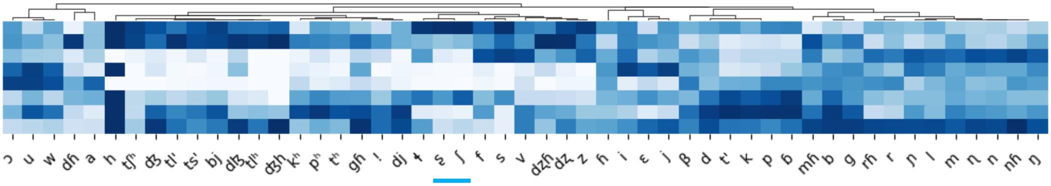
Clusters of: **Affricates**

E



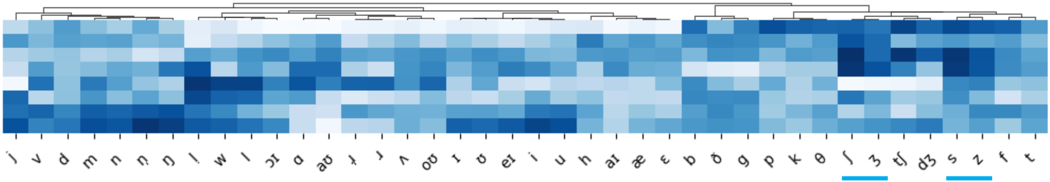
Average activation by gold phoneme

X



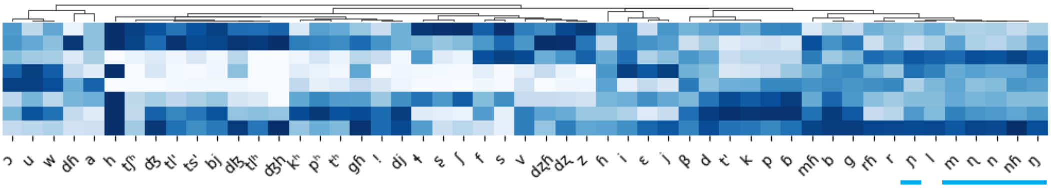
Clusters of: **Sibilant Fricatives**

E



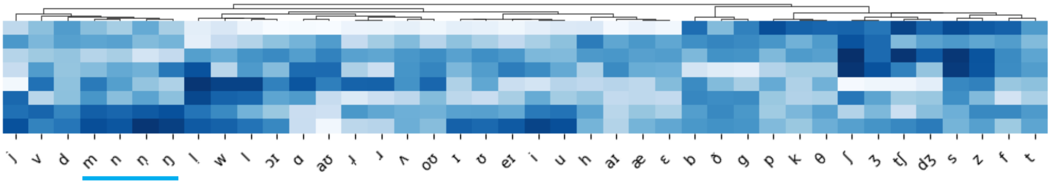
Average activation by gold phoneme

X



Clusters of: **Nasals**

E



Results: Feature Recovery (Quantitative)

- Decompose phonemes into phonological features (Hayes 2011; Hall et al. 2016)
- Fit random forest classifiers: latent bits → distinctive features
- Quantifies feature recoverability from logical statement on latent bits

Results: Feature Recovery (Quantitative)

- Decompose phonemes into phonological features (Hayes 2011; Hall et al. 2016)
- Fit random forest classifiers: latent bits → distinctive features
- Quantifies feature recoverability from logical statement on latent bits

Results: Feature Recovery (Quantitative)

- Decompose phonemes into phonological features (Hayes 2011; Hall et al. 2016)
- Fit random forest classifiers: latent bits $\rightarrow$ distinctive features
- Quantifies feature recoverability from logical statement on latent bits

Feature	F-score
voice	0.94
sonorant	0.92
continuant	0.86
consonantal	0.86
approximant	0.86
syllabic	0.84
dorsal	0.83
strident	0.81
low	0.80
front	0.73
high	0.67
back	0.66
round	0.66
labial	0.65
coronal	0.65
tense	0.63
delayed release	0.62
anterior	0.55
nasal	0.51
distributed	0.38
constricted glottis	0.29
lateral	0.26
labiodental	0.17
trill	0.15
spread glottis	0.12
implosive	0.01

Xitsonga feature recovery

Feature	F-score
voice	0.89
sonorant	0.87
approximant	0.82
continuant	0.81
consonantal	0.78
syllabic	0.74
dorsal	0.71
strident	0.68
coronal	0.63
anterior	0.61
delayed release	0.55
front	0.55
high	0.49
tense	0.45
back	0.44
nasal	0.41
labial	0.37
low	0.37
distributed	0.33
stress	0.33
diphthong	0.33
round	0.27
lateral	0.25
labiodental	0.14
spread glottis	0.07

English feature recovery

Feature	F-score
voice	0.94
sonorant	0.92
continuant	0.86
consonantal	0.86
approximant	0.86
syllabic	0.84
dorsal	0.83
strident	0.81
low	0.80
front	0.73
high	0.67
back	0.66
round	0.66
labial	0.65
coronal	0.65
tense	0.63
delayed release	0.62
anterior	0.55
nasal	0.51
distributed	0.38
constricted glottis	0.29
lateral	0.26
labiodental	0.17
trill	0.15
spread glottis	0.12
implosive	0.01

Xitsonga feature recovery

Example:

p
[-voice]

b
[+voice]

Feature	F-score
voice	0.89
sonorant	0.87
approximant	0.82
continuant	0.81
consonantal	0.78
syllabic	0.74
dorsal	0.71
strident	0.68
coronal	0.63
anterior	0.61
delayed release	0.55
front	0.55
high	0.49
tense	0.45
back	0.44
nasal	0.41
labial	0.37
low	0.37
distributed	0.33
stress	0.33
diphthong	0.33
round	0.27
lateral	0.25
labiodental	0.14
spread glottis	0.07

English feature recovery

Feature	F-score
voice	0.94
sonorant	0.92
continuant	0.86
consonantal	0.86
approximant	0.86
syllabic	0.84
dorsal	0.83
strident	0.81
low	0.80
front	0.73
high	0.67
back	0.66
round	0.66
labial	0.65
coronal	0.65
tense	0.63
delayed release	0.62
anterior	0.55
nasal	0.51
distributed	0.38
constricted glottis	0.29
lateral	0.26
labiodental	0.17
trill	0.15
spread glottis	0.12
implosive	0.01

Xitsonga feature recovery

Example:

t
[-sonorant]

a
[+sonorant]

Feature	F-score
voice	0.89
sonorant	0.87
approximant	0.82
continuant	0.81
consonantal	0.78
syllabic	0.74
dorsal	0.71
strident	0.68
coronal	0.63
anterior	0.61
delayed release	0.55
front	0.55
high	0.49
tense	0.45
back	0.44
nasal	0.41
labial	0.37
low	0.37
distributed	0.33
stress	0.33
diphthong	0.33
round	0.27
lateral	0.25
labiodental	0.14
spread glottis	0.07

English feature recovery

Feature	F-score
voice	0.94
sonorant	0.92
continuant	0.86
consonantal	0.86
approximant	0.86
syllabic	0.84
dorsal	0.83
strident	0.81
low	0.80
front	0.73
high	0.67
back	0.66
round	0.66
labial	0.65
coronal	0.65
tense	0.63
delayed release	0.62
anterior	0.55
nasal	0.51
distributed	0.38
constricted glottis	0.29
lateral	0.26
labiodental	0.17
trill	0.15
spread glottis	0.12
implosive	0.01

Xitsonga feature recovery

Example:

i
[-back]

u
[+back]

Feature	F-score
voice	0.89
sonorant	0.87
approximant	0.82
continuant	0.81
consonantal	0.78
syllabic	0.74
dorsal	0.71
strident	0.68
coronal	0.63
anterior	0.61
delayed release	0.55
front	0.55
high	0.49
tense	0.45
back	0.44
nasal	0.41
labial	0.37
low	0.37
distributed	0.33
stress	0.33
diphthong	0.33
round	0.27
lateral	0.25
labiodental	0.14
spread glottis	0.07

English feature recovery

Feature	F-score
voice	0.94
sonorant	0.92
continuant	0.86
consonantal	0.86
approximant	0.86
syllabic	0.84
dorsal	0.83
strident	0.81
low	0.80
front	0.73
high	0.67
back	0.66
round	0.66
labial	0.65
coronal	0.65
tense	0.63
delayed release	0.62
anterior	0.55
nasal	0.51
distributed	0.38
constricted glottis	0.29
lateral	0.26
labiodental	0.17
trill	0.15
spread glottis	0.12
implosive	0.01

Xitsonga feature recovery

Example:

g
[-anterior]

d
[+anterior]

Feature	F-score
voice	0.89
sonorant	0.87
approximant	0.82
continuant	0.81
consonantal	0.78
syllabic	0.74
dorsal	0.71
strident	0.68
coronal	0.63
anterior	0.61
delayed release	0.55
front	0.55
high	0.49
tense	0.45
back	0.44
nasal	0.41
labial	0.37
low	0.37
distributed	0.33
stress	0.33
diphthong	0.33
round	0.27
lateral	0.25
labiodental	0.14
spread glottis	0.07

English feature recovery

Conclusion

- Percept modeling can support phonology learning
- Top-down constraints likely needed for adult-like performance
- Asymmetries in model performance mimic those of human infants
- Graded feature recovery statistics suggest testable hypotheses about infant speech processing

Conclusion

- Percept modeling can support phonology learning
- Top-down constraints likely needed for adult-like performance
- Asymmetries in model performance mimic those of human infants
- Graded feature recovery statistics suggest testable hypotheses about infant speech processing

Conclusion

- Percept modeling can support phonology learning
- Top-down constraints likely needed for adult-like performance
- Asymmetries in model performance mimic those of human infants
- Graded feature recovery statistics suggest testable hypotheses about infant speech processing

Conclusion

- Percept modeling can support phonology learning
- Top-down constraints likely needed for adult-like performance
- Asymmetries in model performance mimic those of human infants
- Graded feature recovery statistics suggest testable hypotheses about infant speech processing

Thank you!

Code:

`https://github.com/coryshain/dnnseg`

Acknowledgements:

National Science Foundation grant #1422987

Anonymous NAACL 2019 reviewers

References

- Antetomaso, Stephanie et al. (2017). "Modeling phonetic category learning from natural acoustic data". In: Proceedings of the annual Boston University Conference on Language Development.
- Baddeley, Alan, Susan Gathercole, and Costanza Papagno (1998). "The Phonological Loop as a Language Learning Device". In: Psychological Review 105.1, pp. 158–173.
- Baddeley, Alan D and Graham Hitch (1974). Working Memory. Stirling, Scotland: University of Stirling.
- Elsner, Micha and Cory Shain (2017). "Speech segmentation with a neural encoder model of working memory". In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1070–1080.
- Feldman, Naomi H et al. (2013). "A role for the developing lexicon in phonetic category acquisition.". In: Psychological review 120.4, p. 751.
- Hall, Kathleen Currie et al. (2016). "Phonological CorpusTools: A free, open-source tool for phonological analysis". In: 14th Conference for Laboratory Phonology. Vol. 543.

References

- Hayes, Bruce (2011). Introductory phonology. Vol. 32. Hoboken: John Wiley \& Sons.
- 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.
- Peperkamp, Sharon et al. (2006). “The acquisition of allophonic rules: Statistical learning with linguistic constraints”. In: Cognition 101.3, B31–B41.
- Rosenberg, Andrew and Julia Hirschberg (2007). “V-measure: A conditional entropy-based external cluster evaluation measure”. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing
- Swingley, Daniel (2009). “Contributions of infant word learning to language development”. In: Philosophical Transactions of the Royal Society of London B: Biological Sciences 364.1536, pp. 3617–3632.

References

Vallabha, Gautam K et al. (2007). “Unsupervised learning of vowel categories from infant-directed speech”. In: Proceedings of the National Academy of Sciences 104.33, pp. 13273–13278.

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