

Multi-Task and Transfer Learning in Low-Resource Speech Recognition

Joshua Meyer

PhD Candidate

Department of Linguistics

University of Arizona

Roadmap

- Overview of Transfer Learning
 - Multi-Task Learning
 - Copy-Paste Transfer
- Multi-Task Learning Studies
 - Linguistic Tasks
 - Engineered Tasks
 - Discovered Tasks
- Copy-Paste Transfer Studies
 - Multilingual Transfer
 - Model Interpretability
- Conclusion

Introduction

Motivation

Current training methods
for automatic speech recognition
require massive collections of data.

However, most use-cases have
little — if any — available data.

Motivation

Current training methods
for automatic speech recognition
require massive collections of data.

However, most use-cases have
little — if any — available data.

Motivation

Current training methods
for automatic speech recognition
require massive collections of data.

However, most use-cases have
little — if any — available data.

But we can exploit similar use-cases!

Transferring Bias

Useful bias comes from a source
domain

Transferring Bias

Useful bias comes from a source
domain

source **Dataset**

Transferring Bias

Useful bias comes from a source
domain

source **Dataset**

-or-

source **Model**

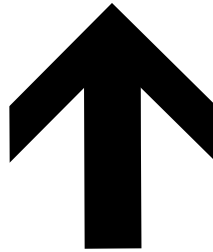
Example of Domain

source Dataset	→	English Speech Dataset
-or-		-or-
source Model	→	Trained English Model

Automatic Speech Recognition?

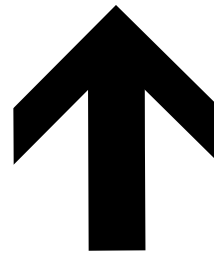
Automatic Speech Recognition

"THE DOG"



Automatic Speech Recognition

"THE DOG"



HARD



Automatic Speech Recognition

"THE DOG"



T H E D O G



EASIER

Automatic Speech Recognition

"THE DOG"



T H E D O G ← *"Phoneme-like" units*



Automatic Speech Recognition

"THE DOG"



T H E D O G



← *Acoustic Model*



Automatic Speech Recognition

"THE DOG"

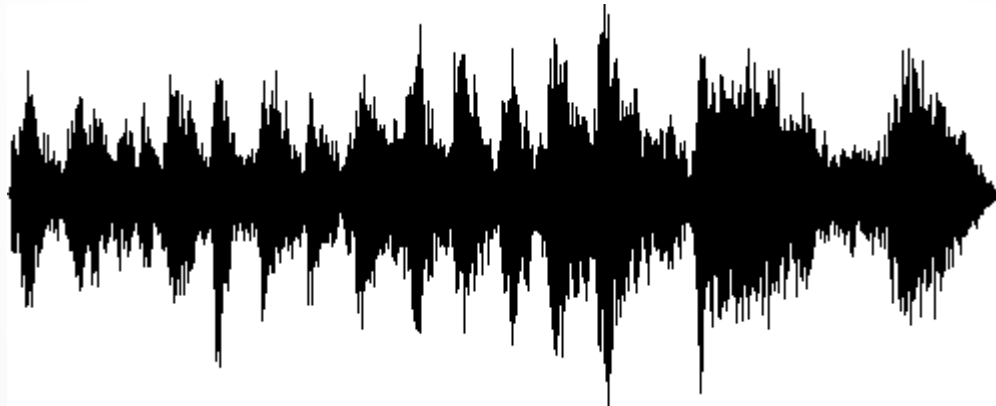


← *Language Model*

T H E D O G



← *Acoustic Model*

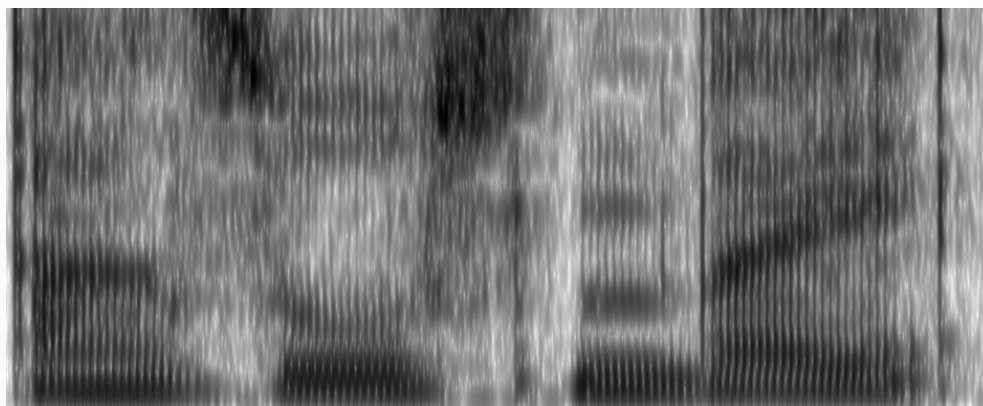


ASR Acoustic Modeling

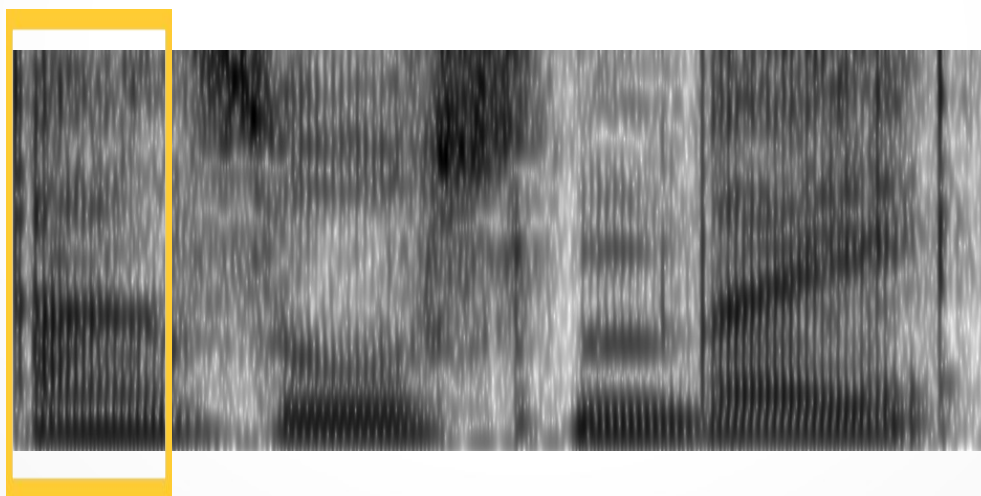
Acoustic Model



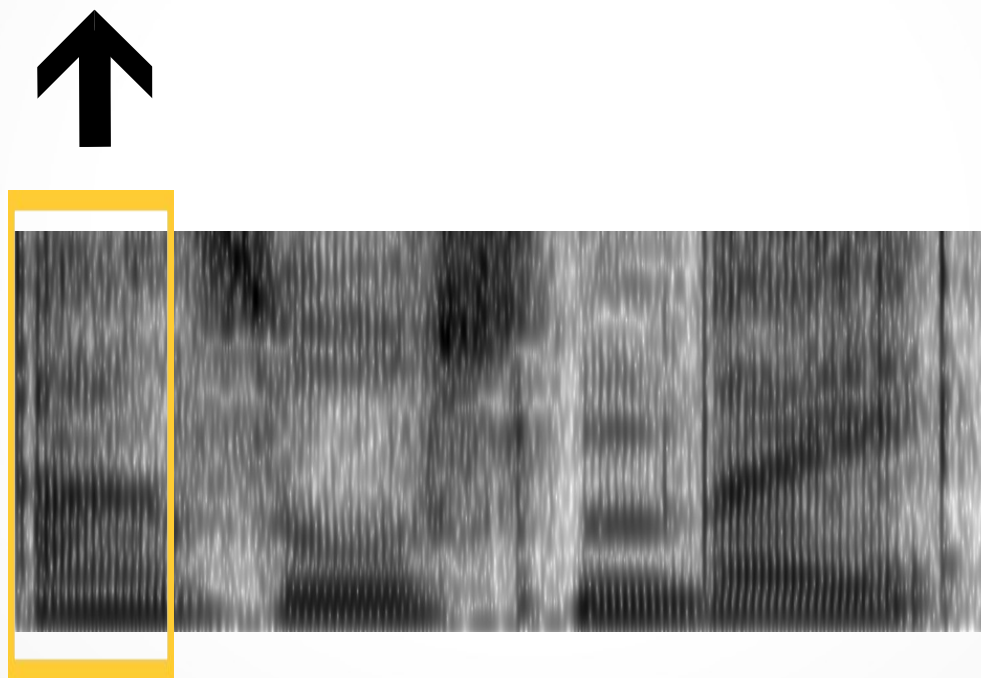
Acoustic Model



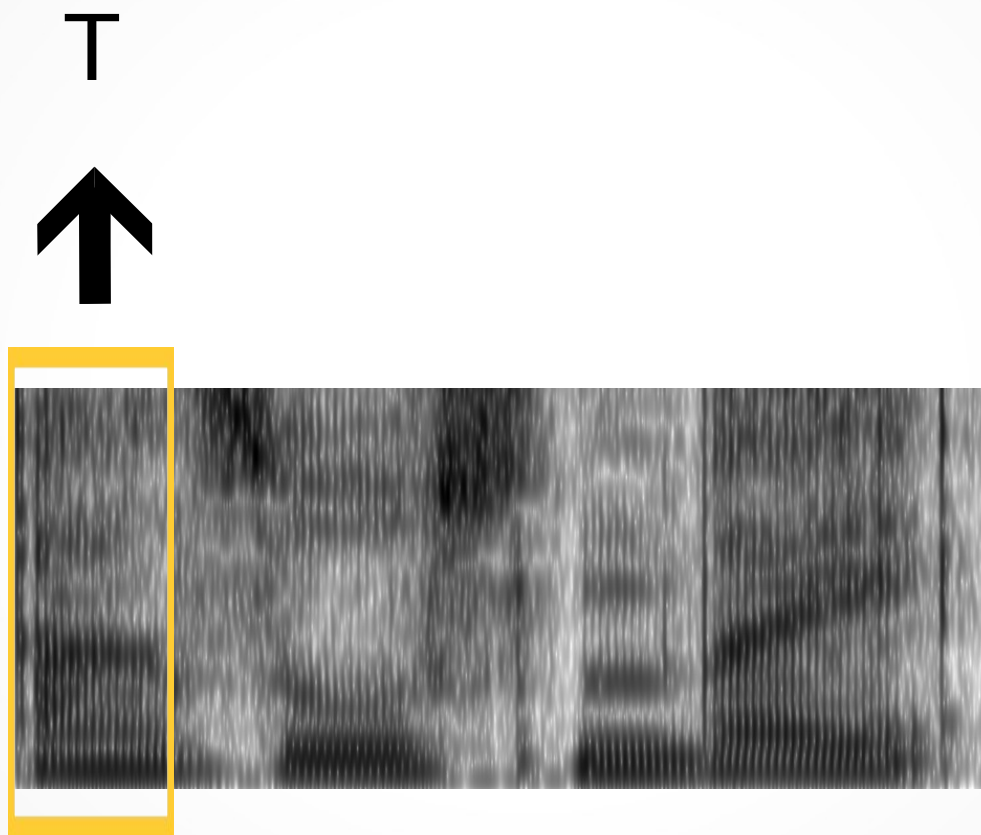
Acoustic Model



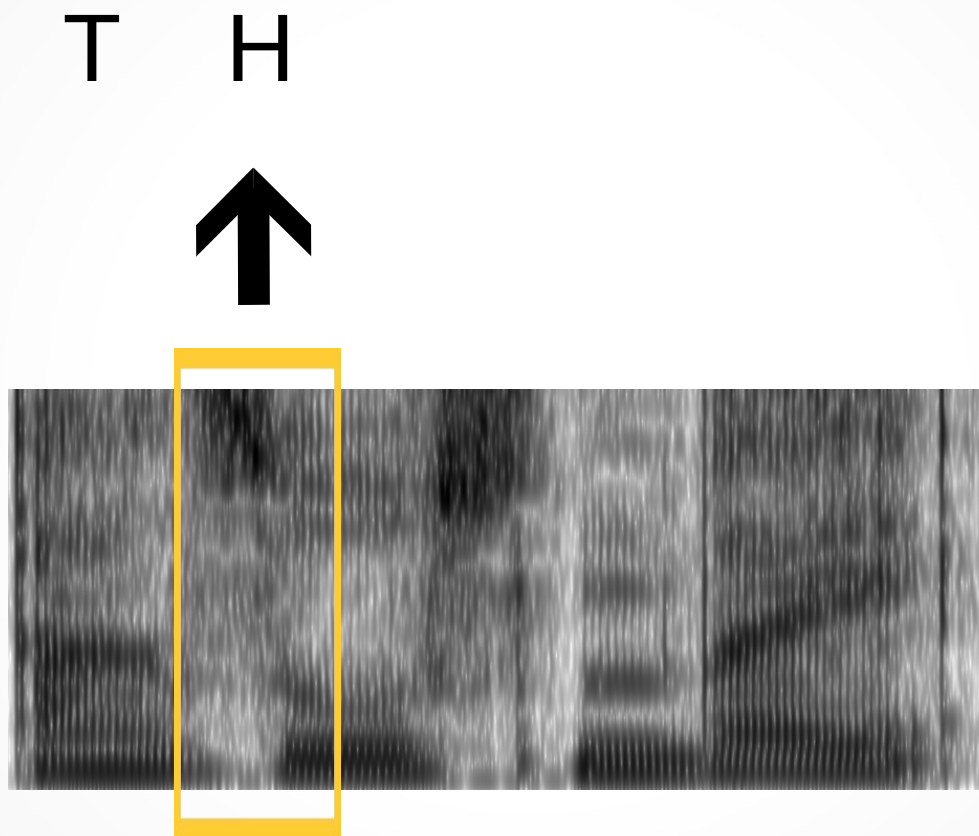
Acoustic Model



Acoustic Model

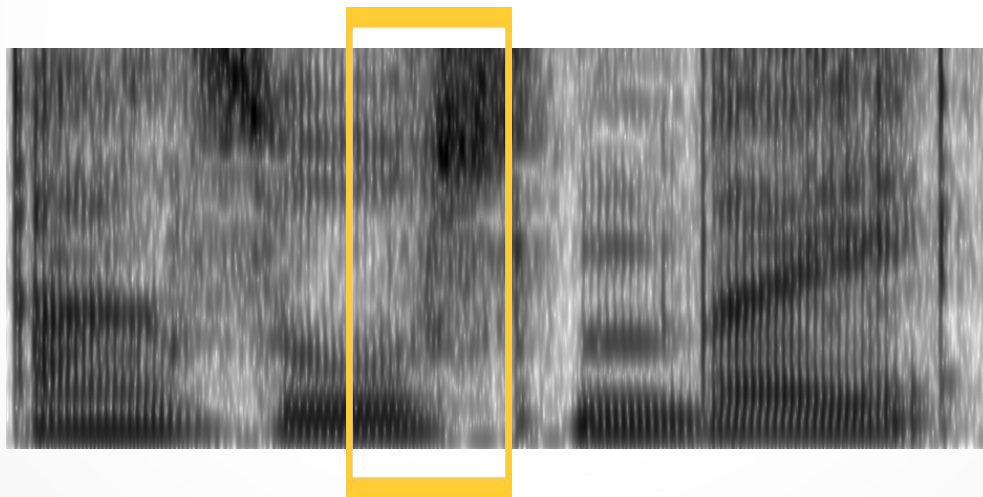


Acoustic Model



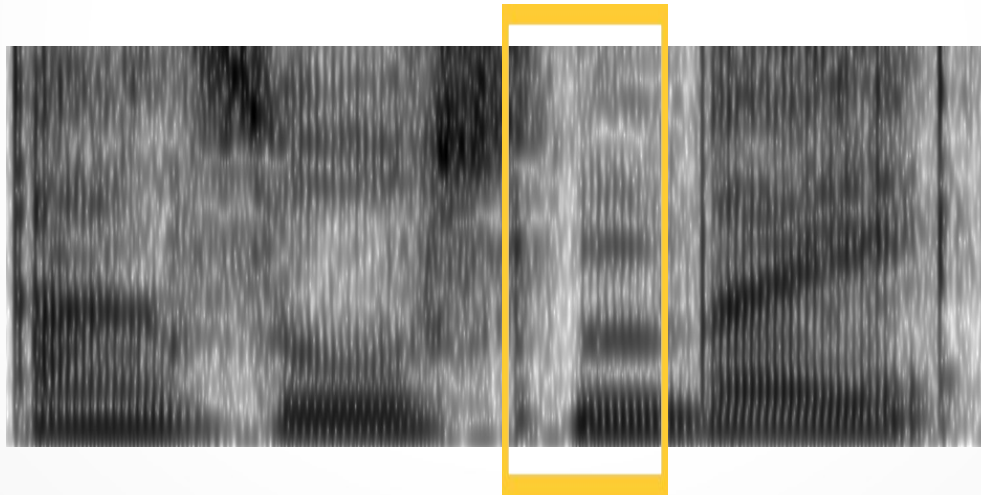
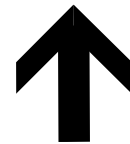
Acoustic Model

T H E



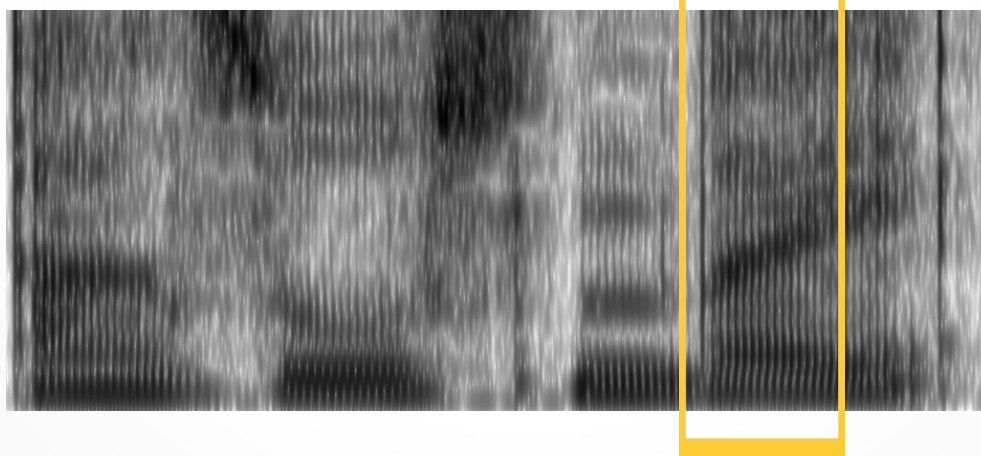
Acoustic Model

T H E D



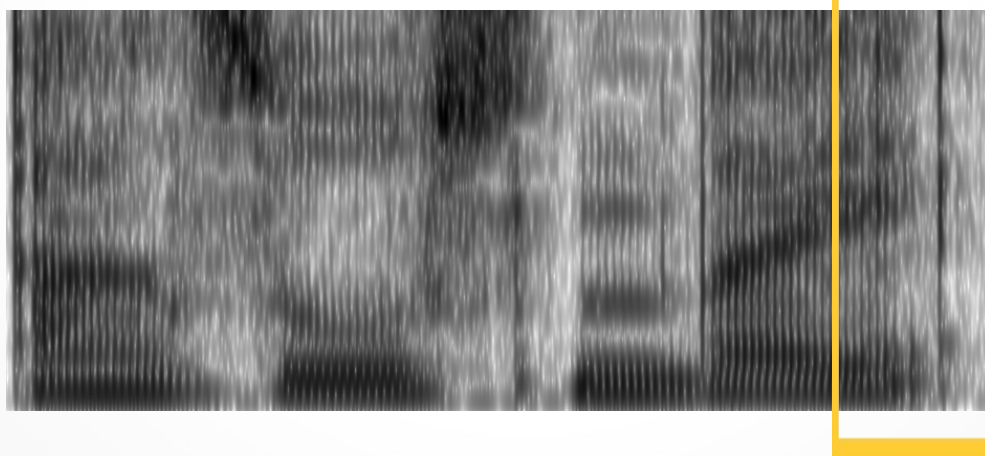
Acoustic Model

T H E D O



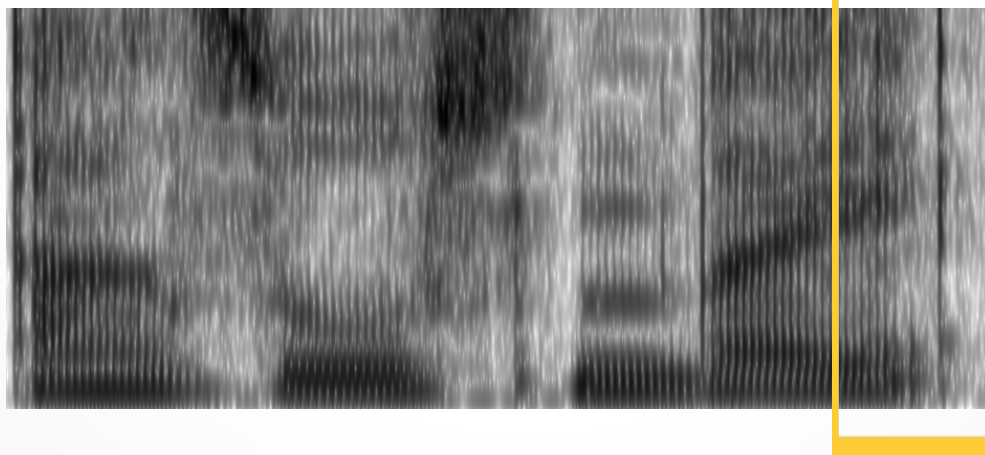
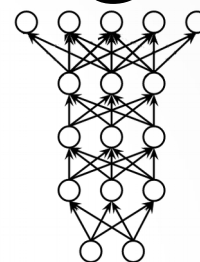
Acoustic Model

T H E D O G



Acoustic Model

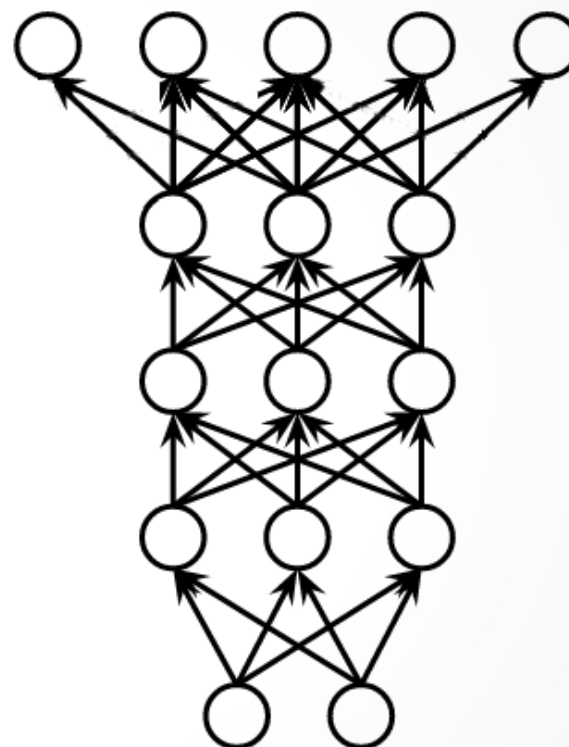
T H E D O G



Acoustic Model

Phonetic Labels

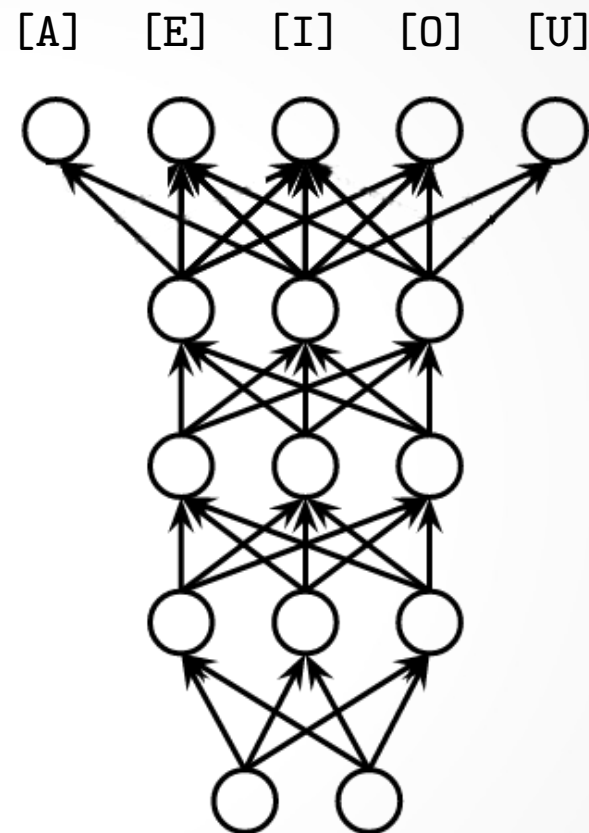
Audio Features



Acoustic Model

Phonetic Labels

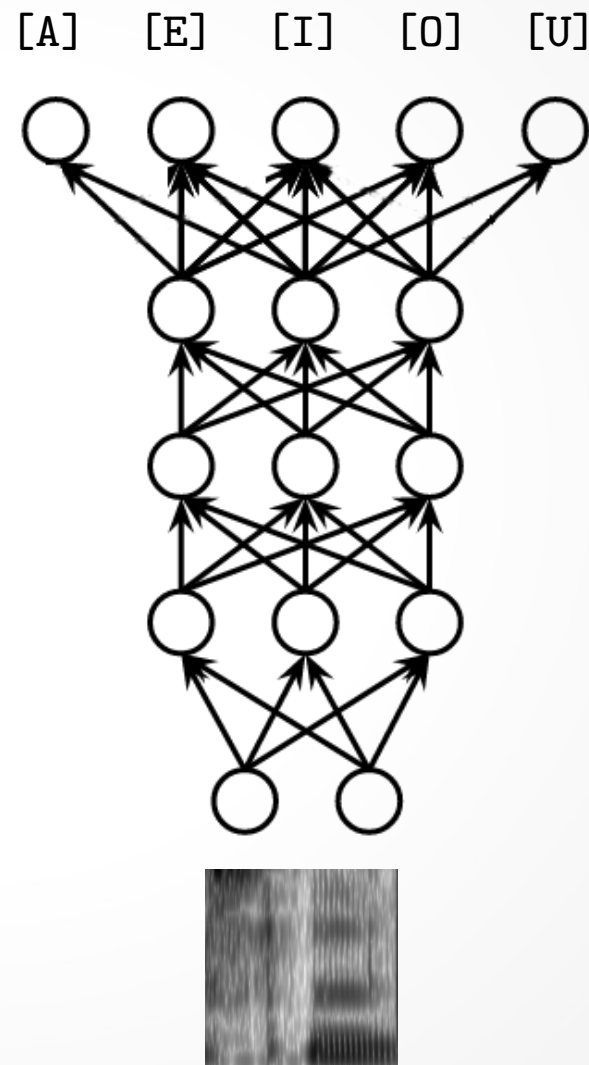
Audio Features



Acoustic Model

Phonetic Labels

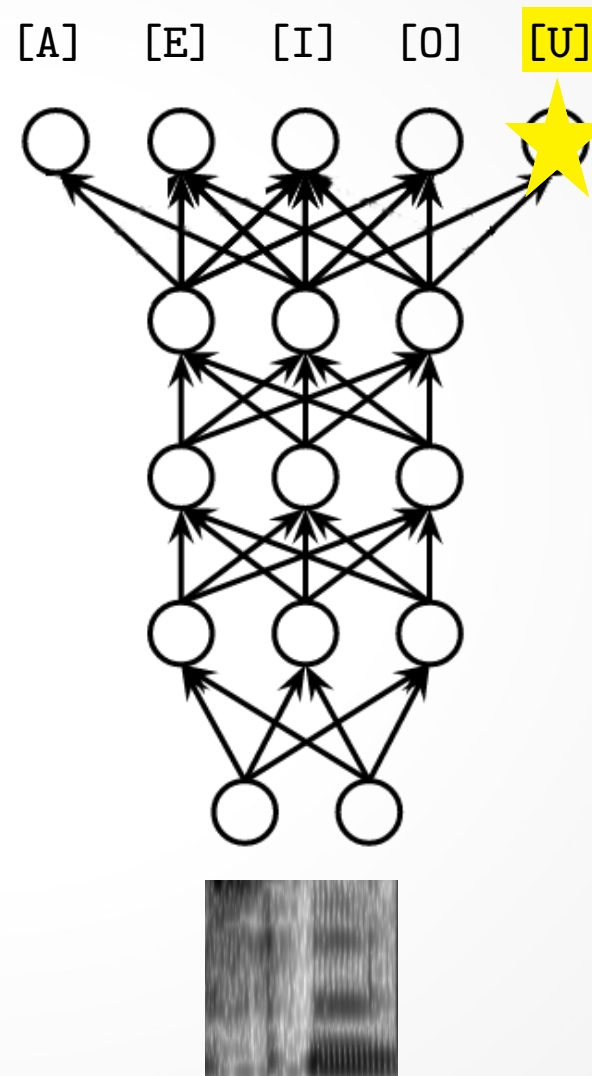
Audio Features



Acoustic Model

Phonetic Labels

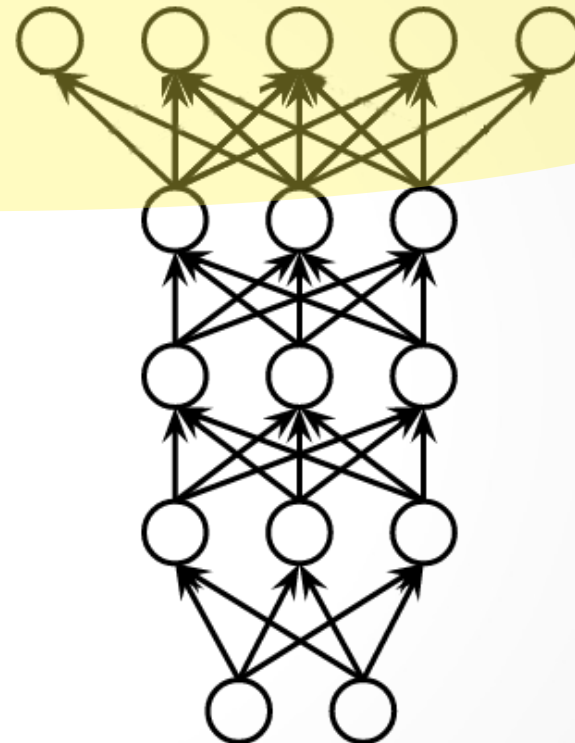
Audio Features



Acoustic Model

Phonetic Labels

[A] [E] [I] [O] [U]



Audio Features

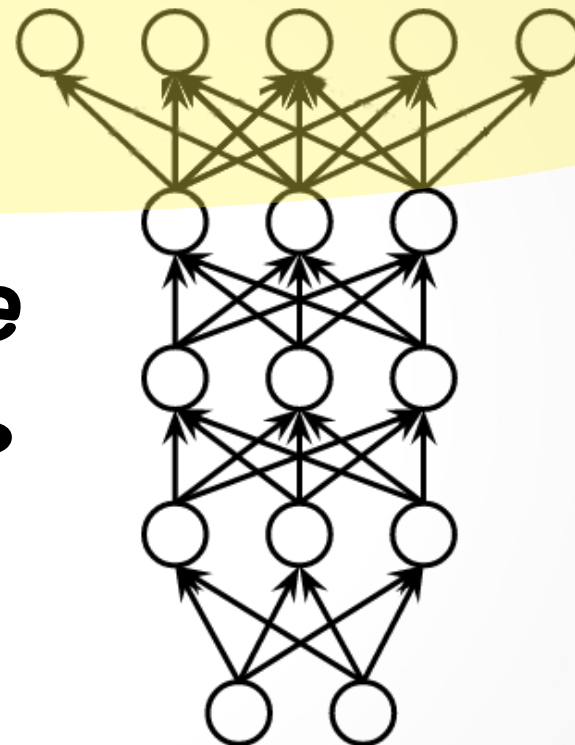
Acoustic Model

Phonetic Labels

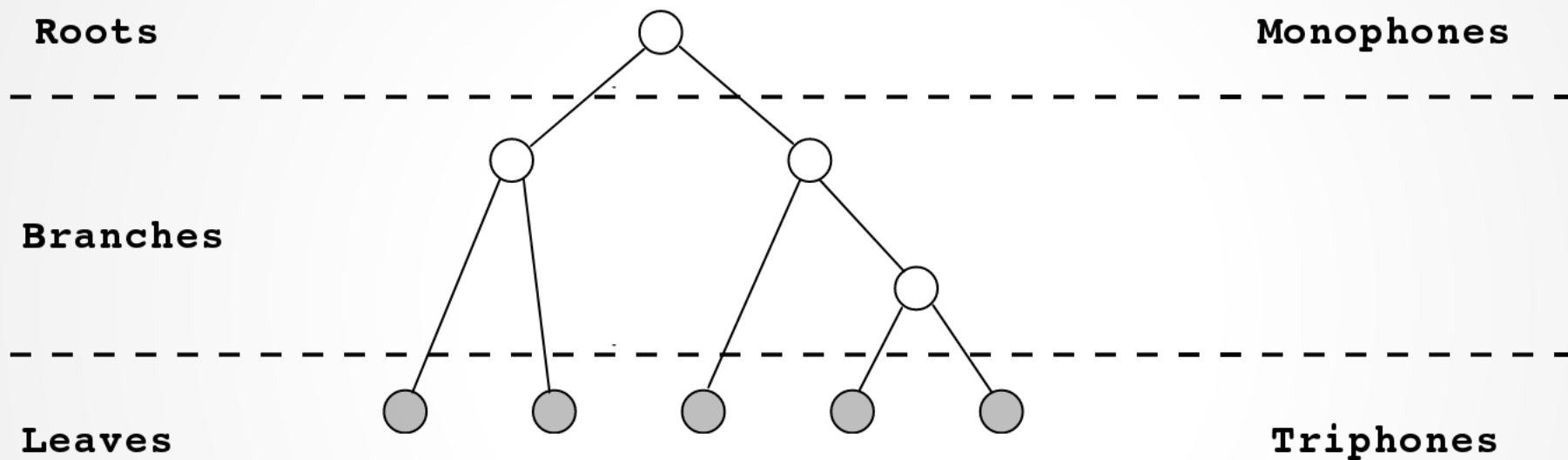
[A] [E] [I] [O] [U]

*Where do we
get labels?*

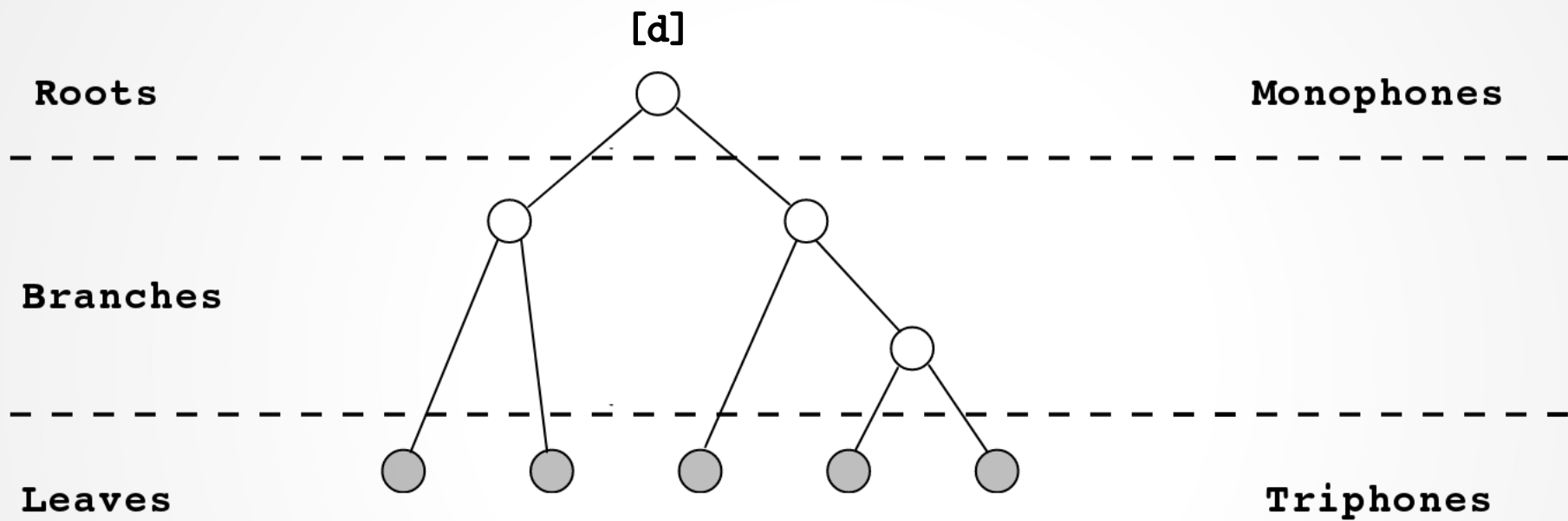
Audio Features



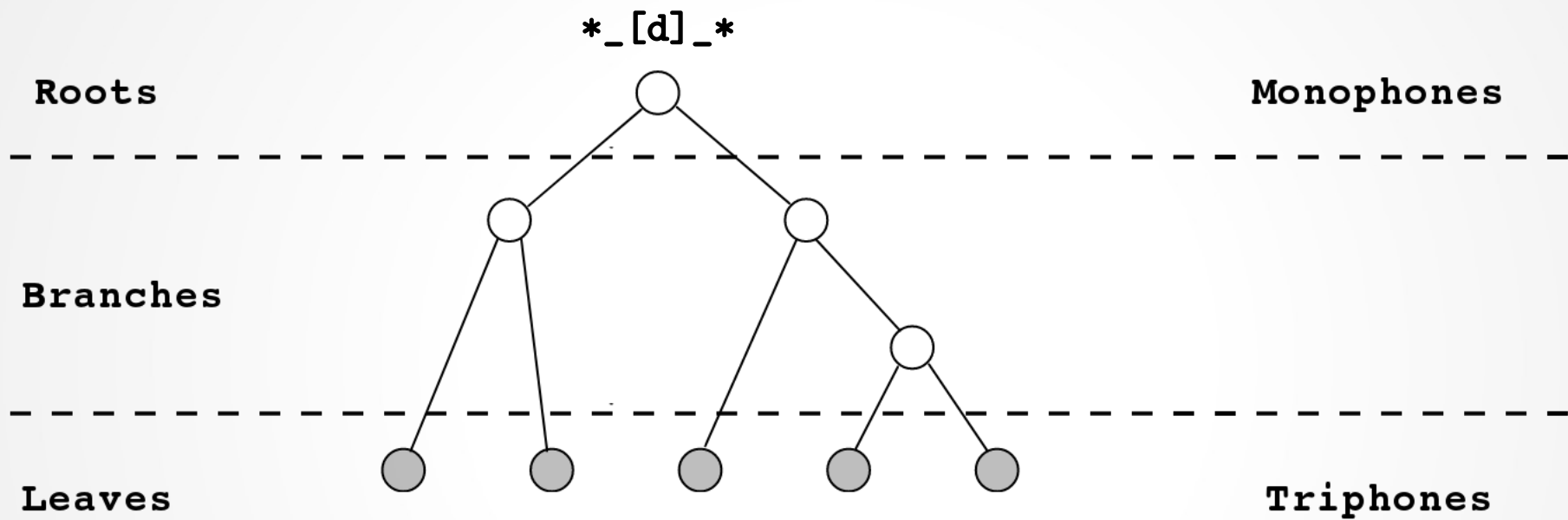
Phonetic Decision Tree



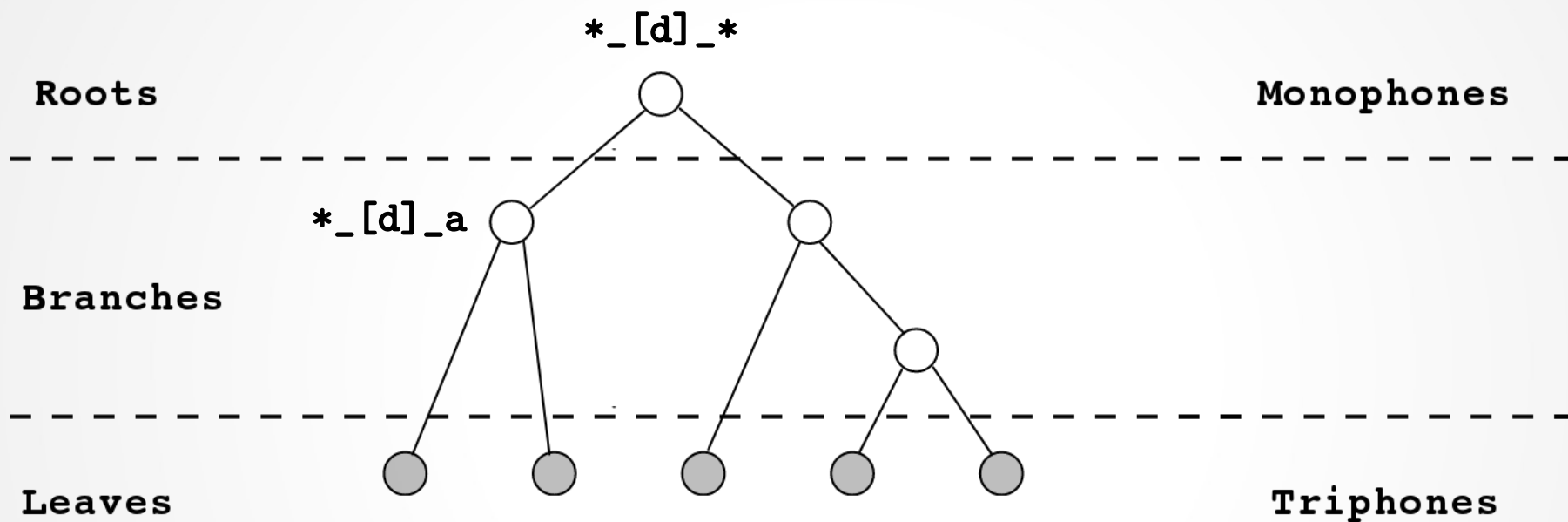
Phonetic Decision Tree



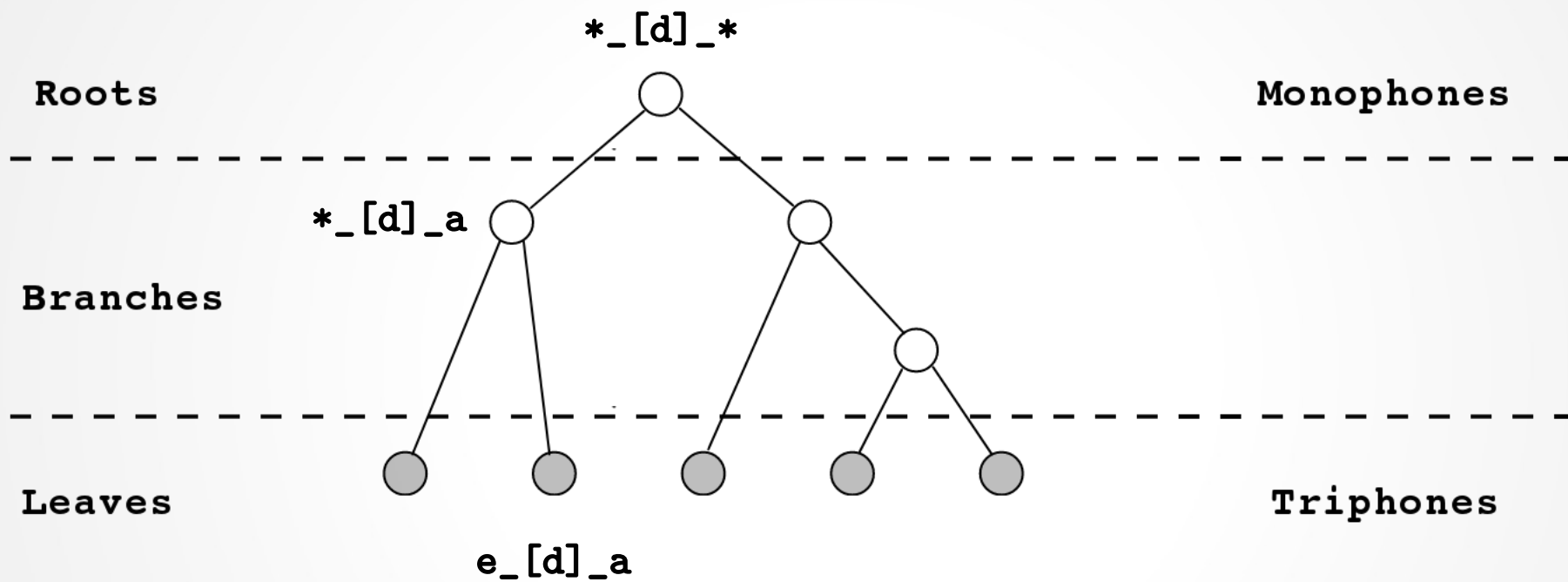
Phonetic Decision Tree



Phonetic Decision Tree

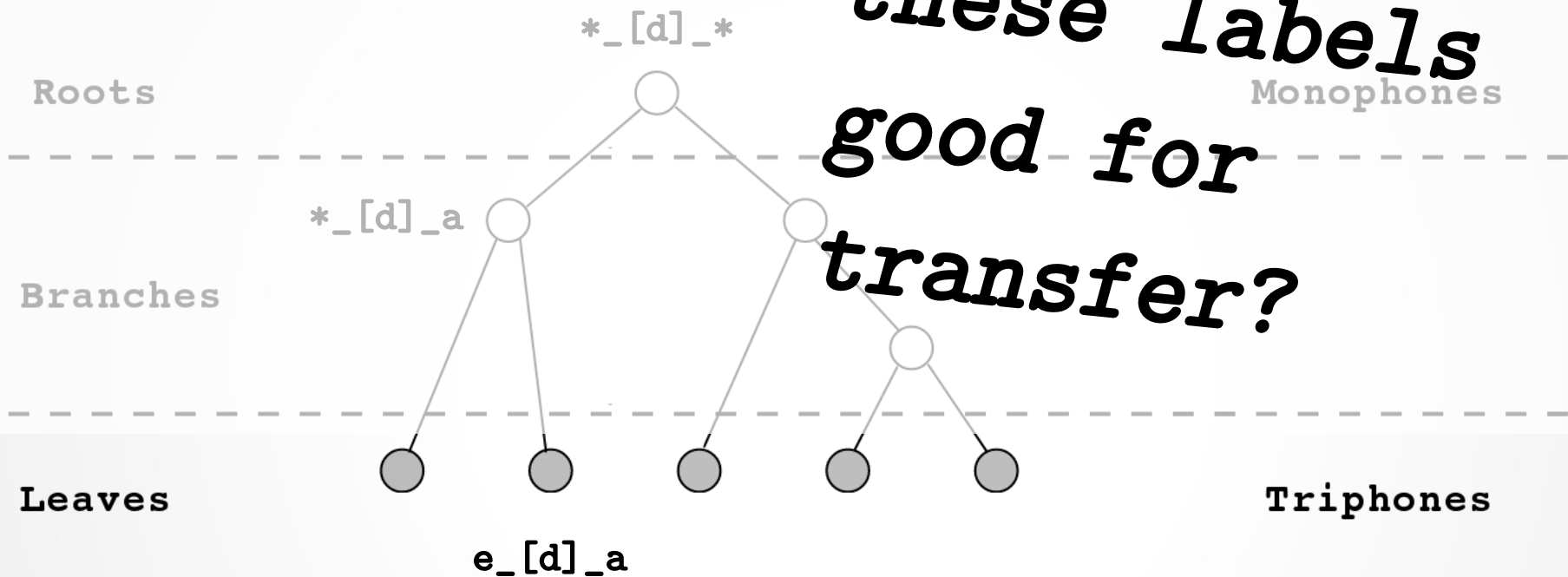


Phonetic Decision Tree



Phonetic Decision Tree

*But... are
these labels
good for
transfer?*



Transferring Bias

Multi-Task Learning

But first, what is a task?

Single-Task Learning



{rottweiler}



{collie}



{terrier}

Single-Task Learning



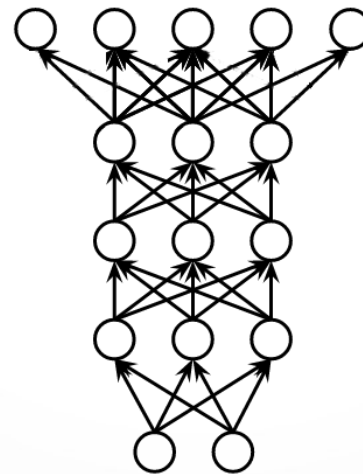
{rottweiler}



{collie}



{terrier}



Single-Task Learning



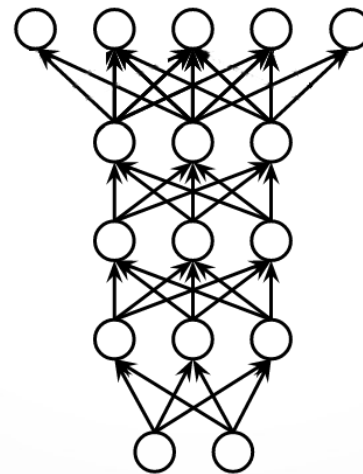
{rottweiler}



{collie}



{terrier}



Single-Task Learning



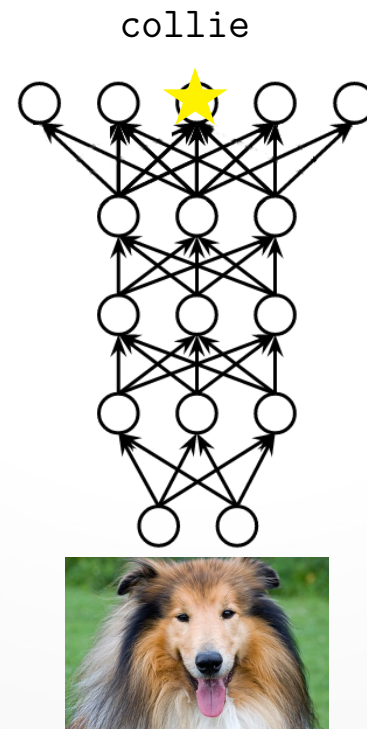
{rottweiler}



{collie}



{terrier}



Multi-Task Learning



{rottweiler, large}



{collie, large}



{terrier, small}

Multi-Task Learning



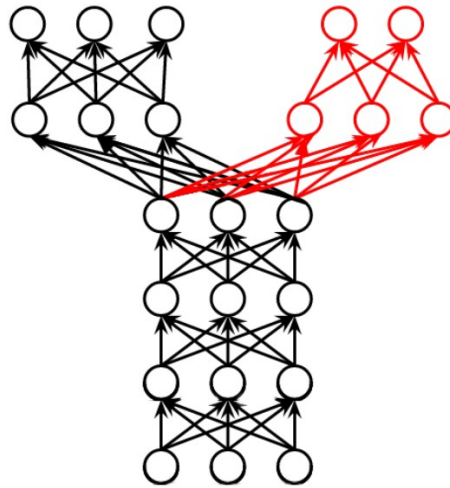
{rottweiler, large}



{collie, large}



{terrier, small}



Multi-Task Learning



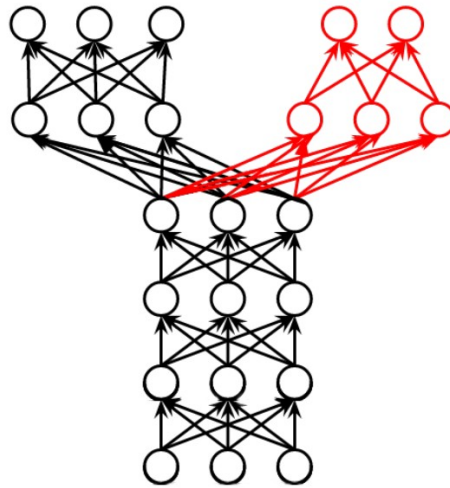
{rottweiler, large}



{collie, large}



{terrier, small}



Multi-Task Learning



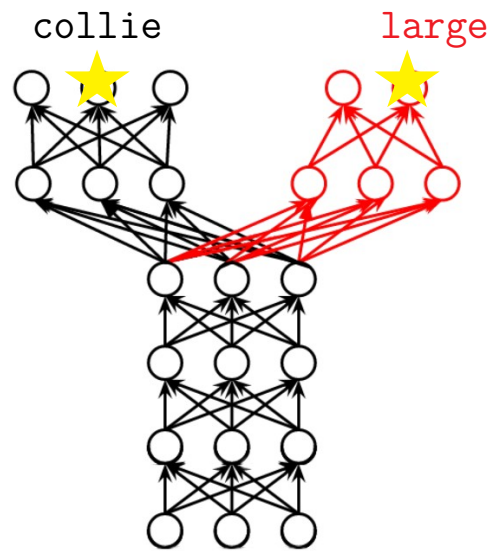
{rottweiler, large}



{collie, large}



{terrier, small}



Multi-Task Studies

Linguist-Crafted Tasks

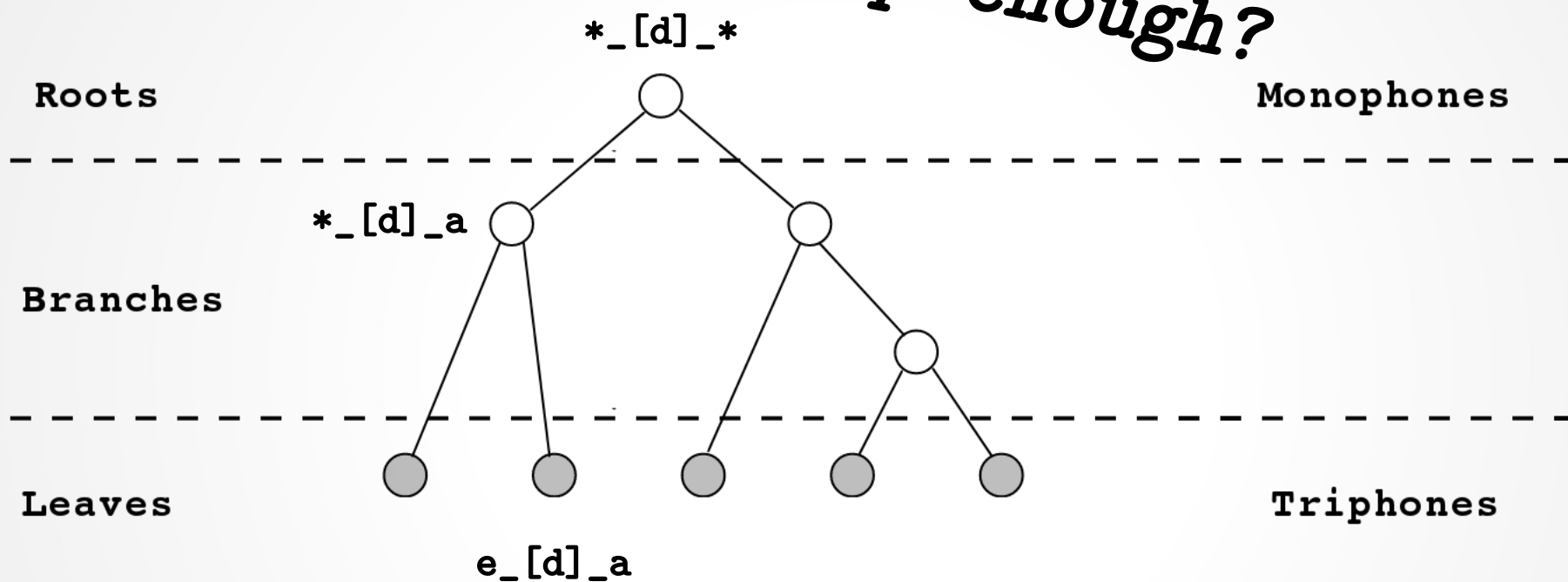
Linguist-Crafted Tasks

Can Linguistics help in a MTL Framework?

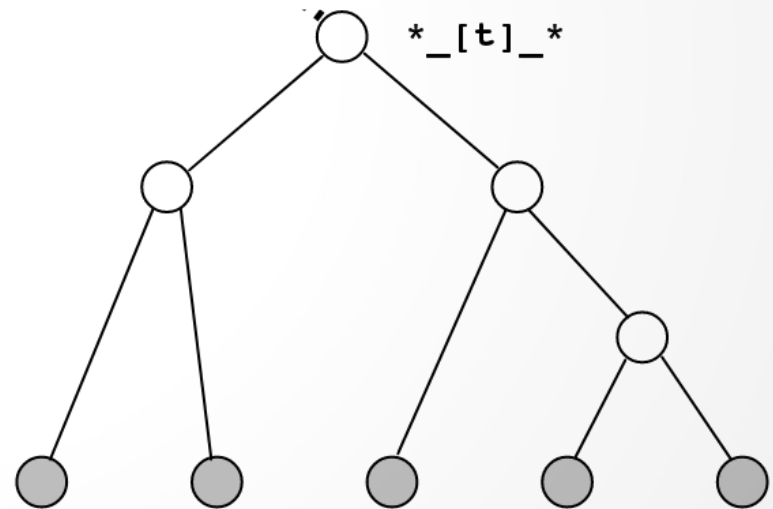
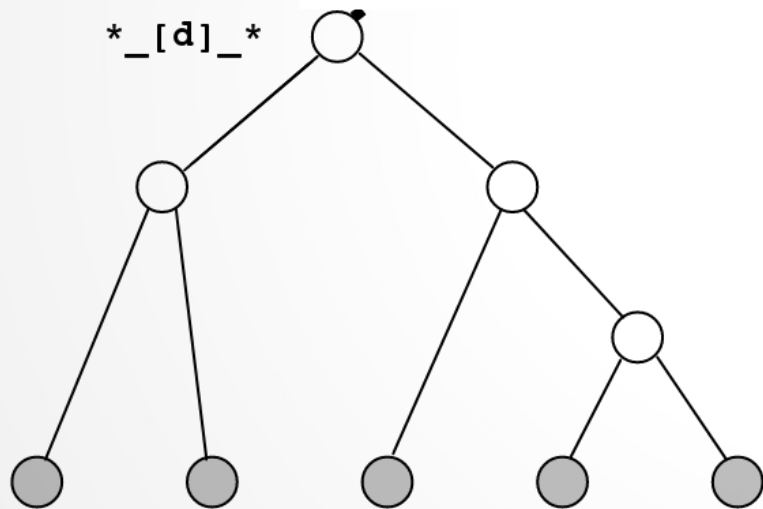
- Bells and Renals say “Yes!”
- XXX says “No!”

Linguist-Crafted Tasks

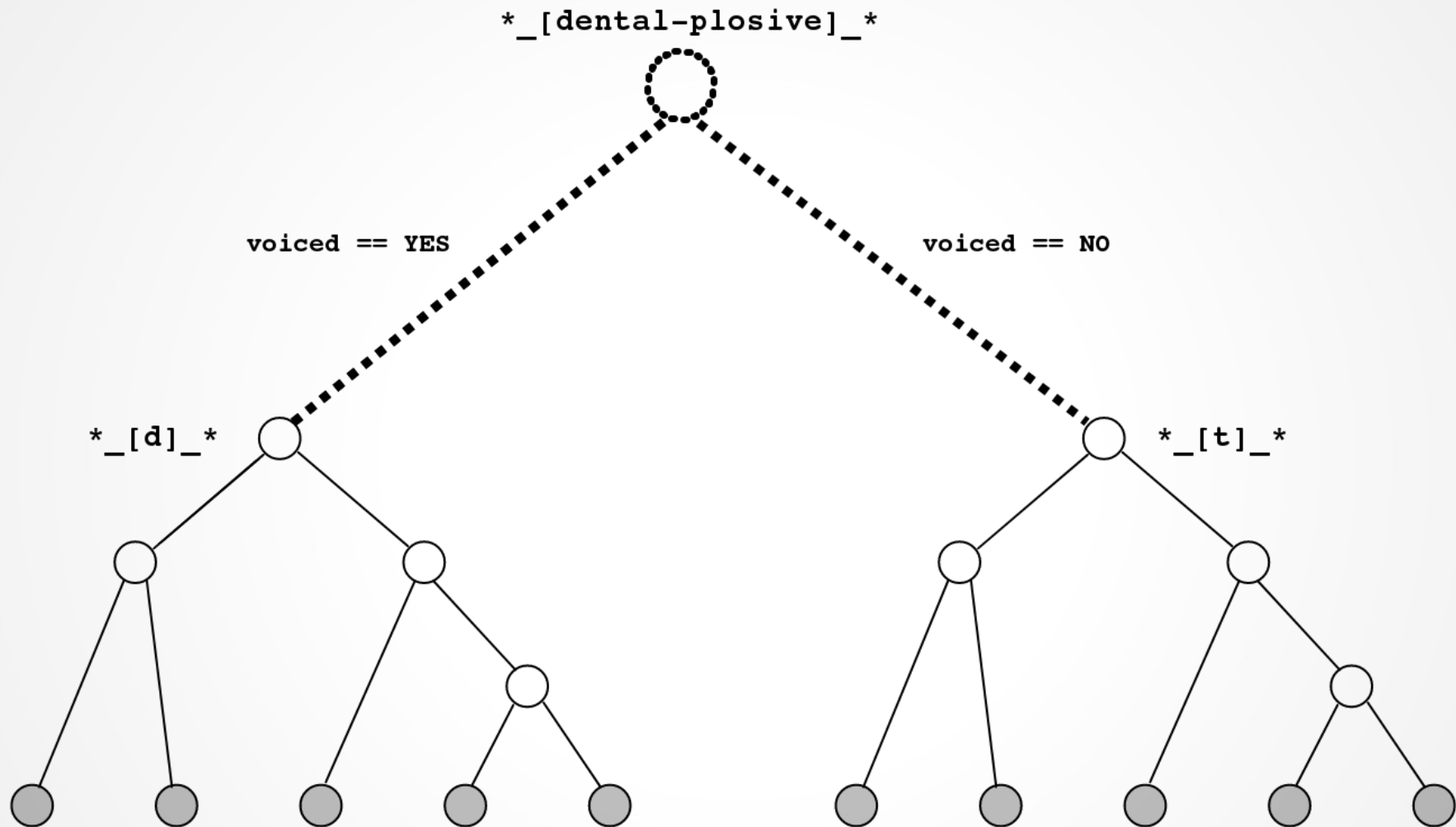
*Are the roots
deep enough?*



Linguist-Crafted Tasks



Linguist-Crafted Tasks



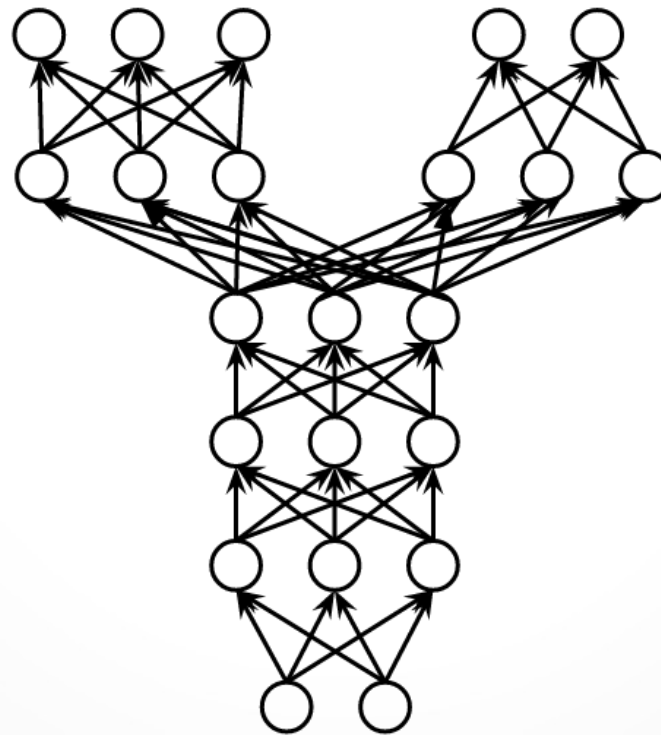
Linguistic Knowledge

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b		t d			ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ	n			ɳ	ɲ	ŋ	ɴ		
Trill	ʙ		r						ʀ		
Tap or Flap		ⱱ	ɾ			ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative			ɬ ɮ								
Approximant		ʋ	ɹ			ɻ	j	ɰ			
Lateral approximant			l			ɭ	ʎ	ʟ			

Linguistic Knowledge

Example: Collapsing on Voice

Baseline Triphones -Voicing Triphones



Training Data

Data

	CORPUS	
	Train	Test
Speaker	LibriSpeech-A	LibriSpeech-B
Language	LibriSpeech-A	Kyrgyz Audiobook

Data

CORPUS		
	Train	Test
Speaker	LibriSpeech-A	LibriSpeech-B
Language	LibriSpeech-A	Kyrgyz Audiobook

0.5 hours

4.86 hours

Data

CORPUS			
	Train	Test	<i>0.5 hours</i>
Speaker	LibriSpeech-A	LibriSpeech-B	
Language	LibriSpeech-A	Kyrgyz Audiobook	
	<i>4.86 hours</i>	<i>1.6 hours</i>	

Alignment Procedure

- GMM-HMM Alignment
- Monophones
 - 25 iterations of Baum-Welch
 - 1,000 Gaussian components
- Triphones
 - 25 iterations of Baum-Welch
 - 1,000 leaves
 - 2,000 Gaussian components

DNN Training Procedure

- 5 Layer, Time-Delay Neural Network
- 500 Nodes / Layer
- ReLU Activations
- Stochastic Gradient Descent for 2 epochs

Monolingual Experiments

CORPUS		
	Train	Test
Speaker	LibriSpeech-A	LibriSpeech-B
Language	LibriSpeech-A	Kyrgyz Audiobook

Monolingual Experiments

Auxiliary Tasks	WER%	
	Triphones	Monophones
STL Baseline		41.67
Voice	41.16	42.36
Place	42.66	40.61
Manner	42.03	41.70
Voice + Place	42.90	41.49
Voice + Manner	42.45	42.66
Place + Manner	42.66	41.82
Voice + Manner + Place	42.42	42.72

Monolingual Experiments

Not so great :(

Auxiliary Tasks	WER%	
	Triphones	Monophones
STL Baseline		41.67
Voice	41.16	42.36
Place	42.66	40.61
Manner	42.03	41.70
Voice + Place	42.90	41.49
Voice + Manner	42.45	42.66
Place + Manner	42.66	41.82
Voice + Manner + Place	42.42	42.72

Monolingual Experiments

The **main** task is **more** important...

Monolingual Experiments

The **main** task is **more** important...

Implement a relative **weighting**!

Monolingual Experiments

Source:Target Weighting

1:1

1/3:1

Auxiliary Tasks	WER%		WER%	
	Triphones	Monophones	Triphones	Monophones
STL Baseline	41.67		41.67	
Voice	41.16	42.36	41.00	40.43
Place	42.66	40.61	41.37	41.46
Manner	42.03	41.70	40.43	41.34
Voice + Place	42.90	41.49	41.31	41.28
Voice + Manner	42.45	42.66	41.25	42.18
Place + Manner	42.66	41.82	42.03	42.48
Voice + Manner + Place	42.42	42.72	41.64	41.88

Monolingual Experiments

Now, that looks better :)

Source:Target Weighting

1:1

1/3:1

Auxiliary Tasks	WER%		WER%	
	Triphones	Monophones	Triphones	Monophones
STL Baseline	41.67		41.67	
Voice	41.16	42.36	41.00	40.43
Place	42.66	40.61	41.37	41.46
Manner	42.03	41.70	40.43	41.34
Voice + Place	42.90	41.49	41.31	41.28
Voice + Manner	42.45	42.66	41.25	42.18
Place + Manner	42.66	41.82	42.03	42.48
Voice + Manner + Place	42.42	42.72	41.64	41.88

Multilingual Experiments

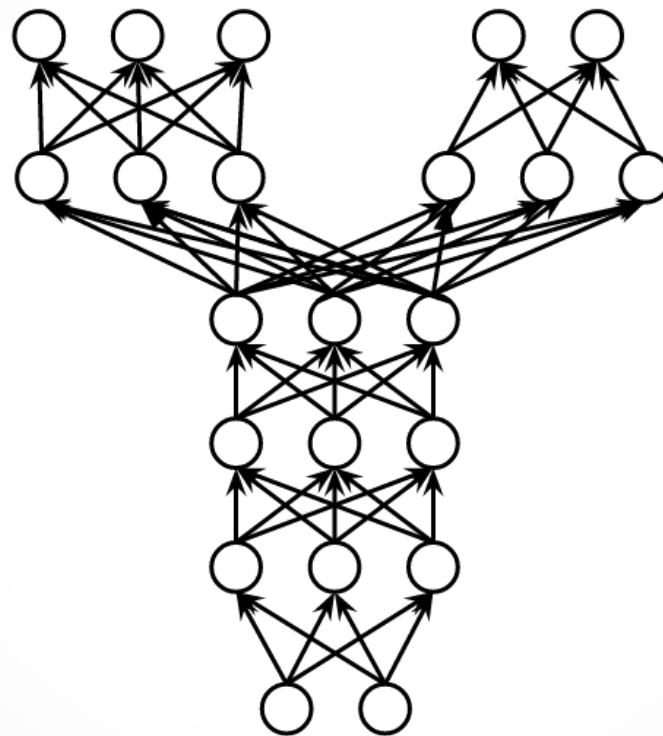
Multilingual Experiments

CORPUS		
	Train	Test
Speaker	LibriSpeech-A	LibriSpeech-B
Language	LibriSpeech-A	Kyrgyz Audiobook

Multilingual Experiments

Standard
Kyrgyz

Linguistic
English



Multilingual Experiments

Auxiliary Tasks	WER%	
	Triphones	Monophones
STL Baseline		53.07
Phonemes	53.95	52.78
Voice	54.05	53.85
Place	55.22	53.95
Manner	53.37	53.27
Voice + Place	55.22	53.46
Voice + Manner	55.12	53.46
Place + Manner	55.51	53.66
Voice + Manner + Place	54.15	54.44

Multilingual Experiments

Not so great :(

Auxiliary Tasks	WER%	
	Triphones	Monophones
STL Baseline		53.07
Phonemes	53.95	52.78
Voice	54.05	53.85
Place	55.22	53.95
Manner	53.37	53.27
Voice + Place	55.22	53.46
Voice + Manner	55.12	53.46
Place + Manner	55.51	53.66
Voice + Manner + Place	54.15	54.44

Multilingual Experiments

Source:Target Weighting

1:1

1/3:1

Auxiliary Tasks	WER%		WER%	
	Triphones	Monophones	Triphones	Monophones
STL Baseline		53.07		53.07
Phonemes	53.95	52.78	51.80	51.61
Voice	54.05	53.85	52.39	53.46
Place	55.22	53.95	51.90	52.29
Manner	53.37	53.27	52.00	51.80
Voice + Place	55.22	53.46	52.68	52.78
Voice + Manner	55.12	53.46	51.22	51.32
Place + Manner	55.51	53.66	50.83	53.66
Voice + Manner + Place	54.15	54.44	52.78	52.39

Multilingual Experiments

Now, that looks better :)

Source:Target Weighting

1:1

1/3:1

Auxiliary Tasks	WER%		WER%	
	Triphones	Monophones	Triphones	Monophones
STL Baseline	53.07		53.07	
Phonemes	53.95	52.78	51.80	51.61
Voice	54.05	53.85	52.39	53.46
Place	55.22	53.95	51.90	52.29
Manner	53.37	53.27	52.00	51.80
Voice + Place	55.22	53.46	52.68	52.78
Voice + Manner	55.12	53.46	51.22	51.32
Place + Manner	55.51	53.66	50.83	53.66
Voice + Manner + Place	54.15	54.44	52.78	52.39

Summary: Linguistic Experiments

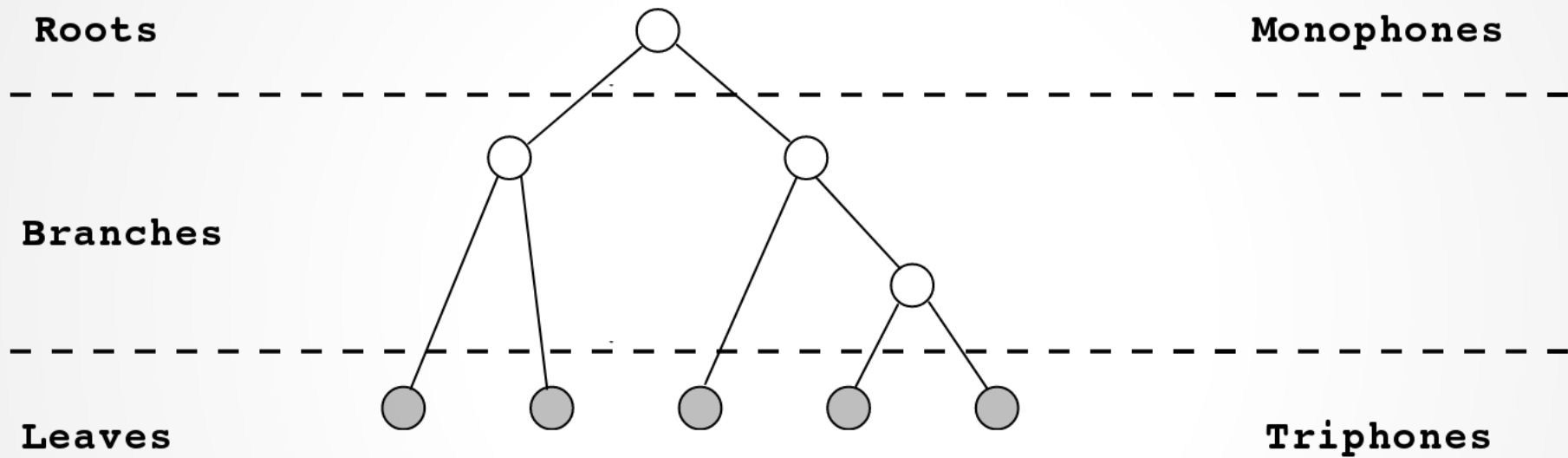
- Linguistics **can** help
 - But we must keep in mind weighting
- Multilingual transfer more affected
 - Triphone leaves of the tree
not best for transfer

Engineered (Multilingual) Tasks

Engineered Tasks

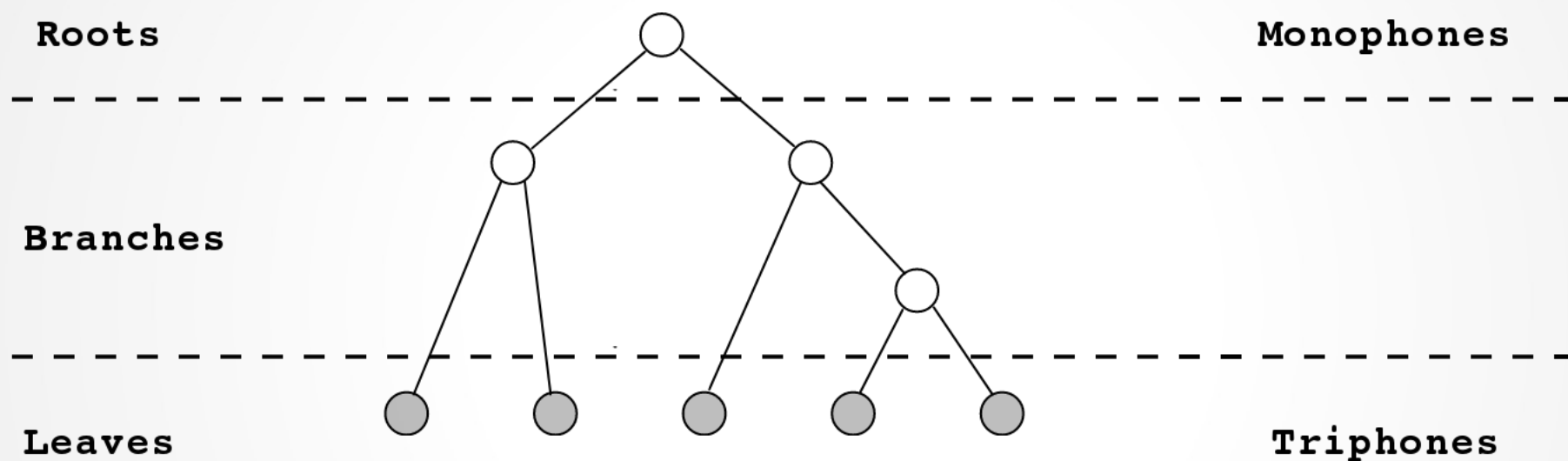
Can we find useful linguistic bias
without a linguist?

Linguist-Crafted Tasks



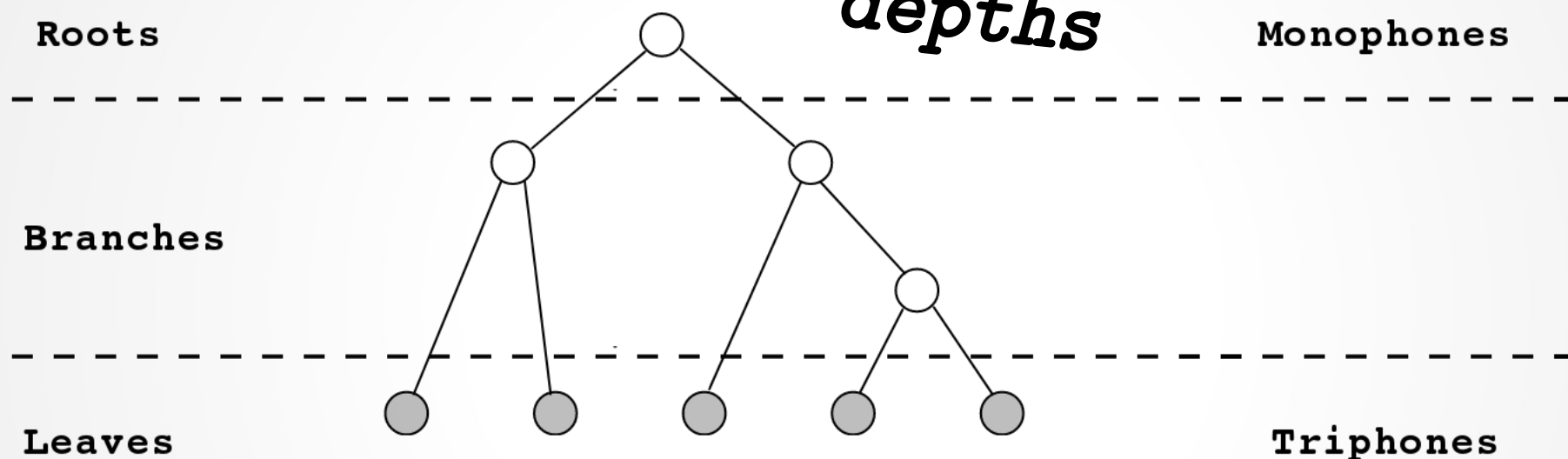
Linguist-Crafted Tasks

Lots of unused structure...



Linguist-Crafted Tasks

We can slice the tree at varying depths



Alignment Procedure

- GMM-HMM Alignment
- Monophones
 - 25 iterations of Baum-Welch
 - 1,000 Gaussian components
- “Half-phones”
 - 25 iterations of Baum-Welch
 - 792 leaves
 - 5,000 Gaussian components
- Triphones
 - 25 iterations of Baum-Welch
 - 1584 leaves
 - 5,000 Gaussian components

Alignment Procedure

- GMM-HMM Alignment
- Monophones
 - 25 iterations of Baum-Welch
 - 1,000 Gaussian components
- “Half-phones” (**Half-way down the tree**)
 - 25 iterations of Baum-Welch
 - 792 leaves
 - 5,000 Gaussian components
- Triphones
 - 25 iterations of Baum-Welch
 - 1584 leaves
 - 5,000 Gaussian components

DNN Training Procedure

- 5 Layer, Time-Delay Neural Network
- 500 Nodes / Layer
- ReLU Activations
- Stochastic Gradient Descent for 10 epochs

DNN Training Procedure

Source:Target Ratio	Target Weighting
2:1	1.53x
1:1	3.06x
1:2	6.12x

DNN Training Procedure

Smarter weighting

Source:Target Ratio	Target Weighting
2:1	1.53x
1:1	3.06x
1:2	6.12x

Multilingual Engineered Tasks

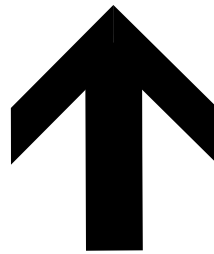
Auxiliary (Source Lang) Tasks	Source:Target Weighting		
	<i>1-to-2</i>	<i>1-to-1</i>	<i>2-to-1</i>
STL Baseline		50.54	
Monophones	48.20	47.32	47.41
Halfphones	48.68	46.73	48.68
Triphones	49.37	47.12	46.73
Monophones + Halfphones	48.20	48.49	48.10
Halfphones + Triphones	50.05	48.00	47.90
Monophones + Halfphones + Halfphones	48.88	48.20	48.59

End-to-End Transfer Studies

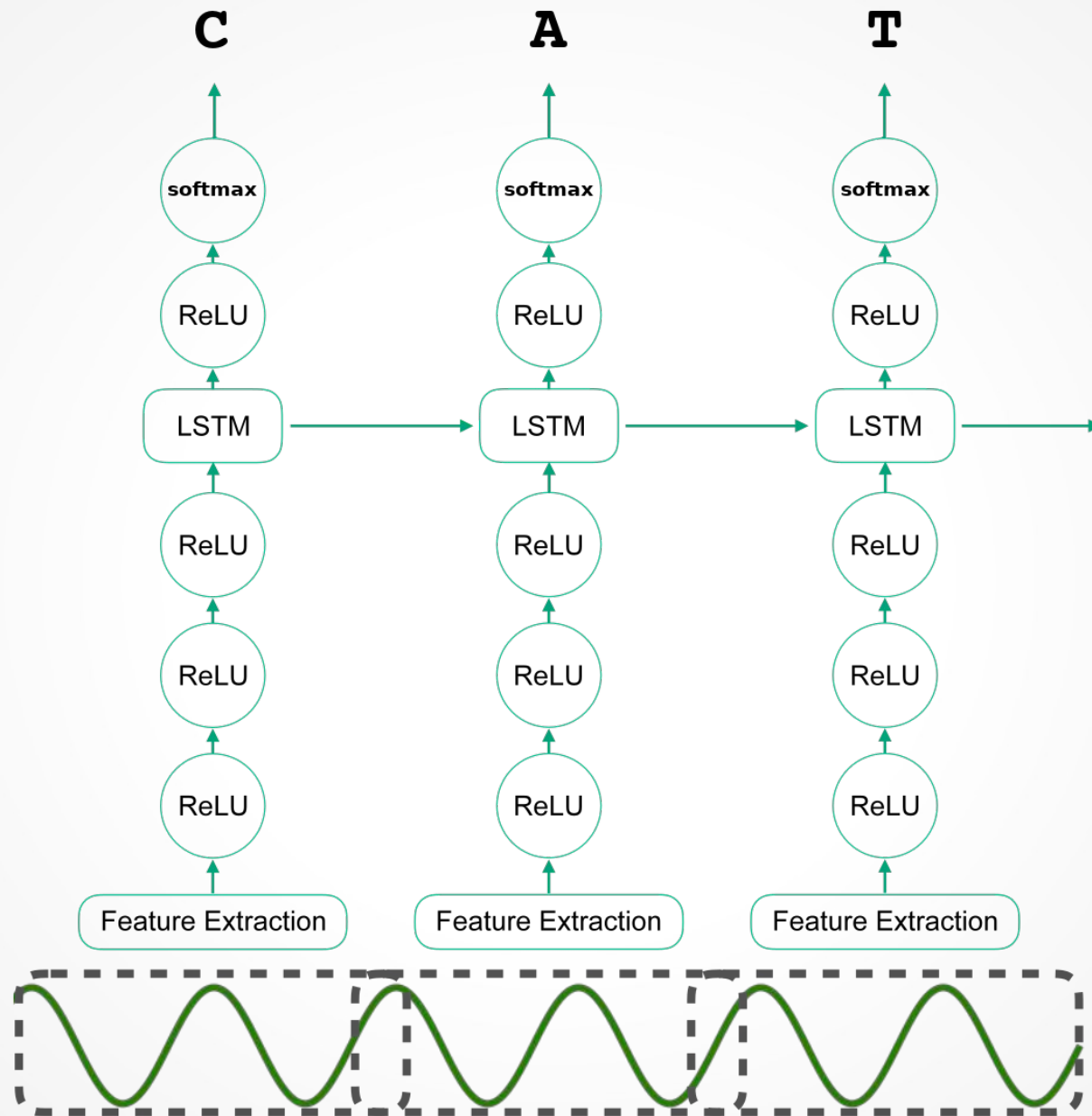
End-to-end?

End-to-end

"THE DOG"



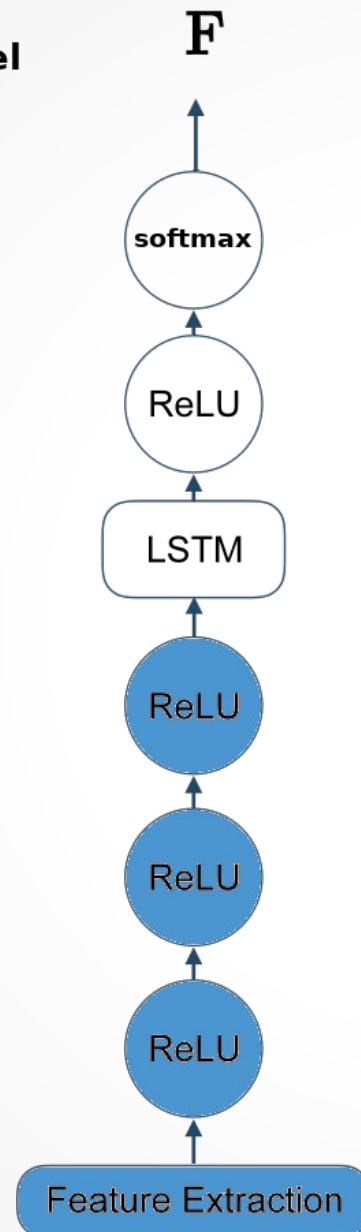
DeepSpeech



Transfer Experiments

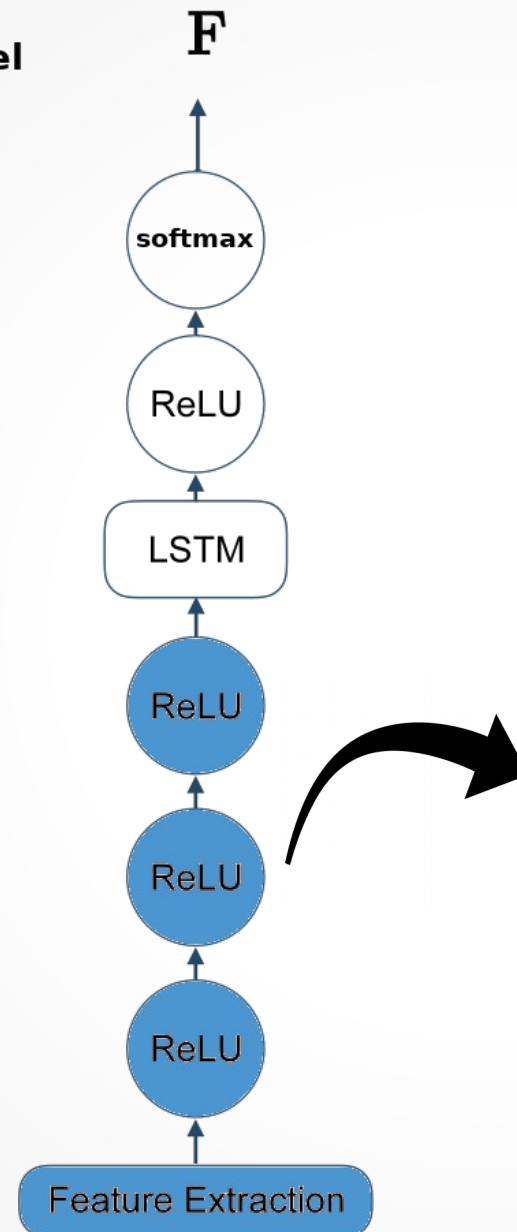
CTC Transfer Experiments

English
Source Model



CTC Transfer Experiments

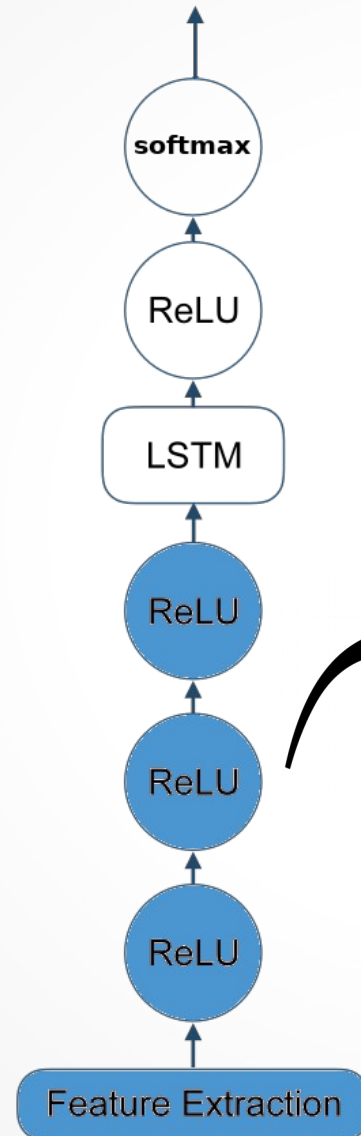
English
Source Model



CTC Transfer Experiments

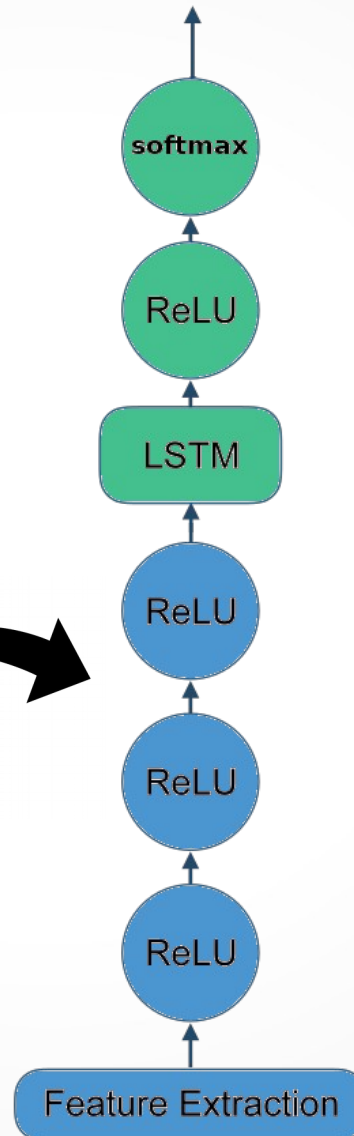
English
Source Model

F



X

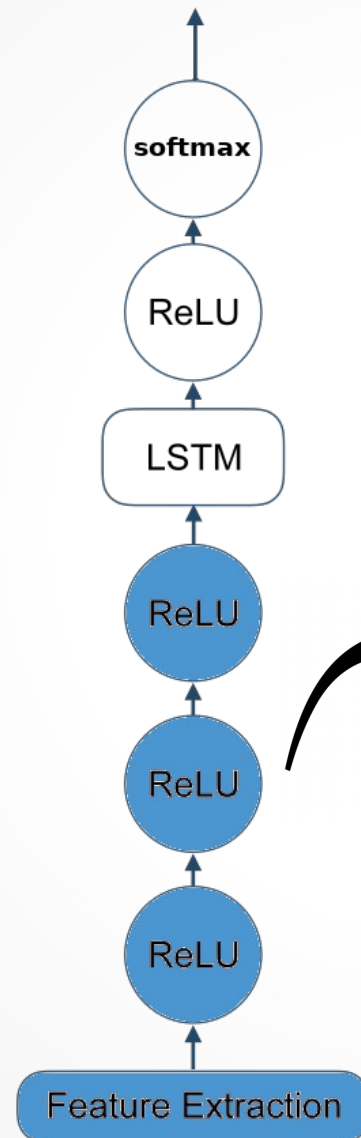
Target Language
Model



CTC Transfer Experiments

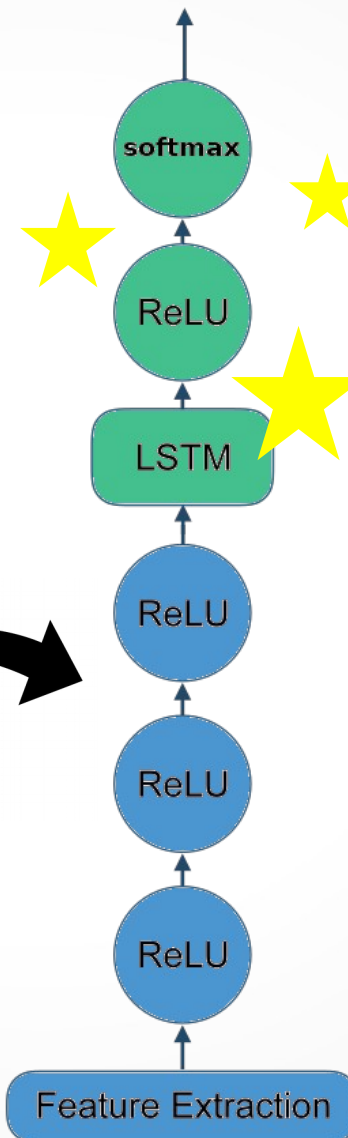
English
Source Model

F



Target Language
Model

X



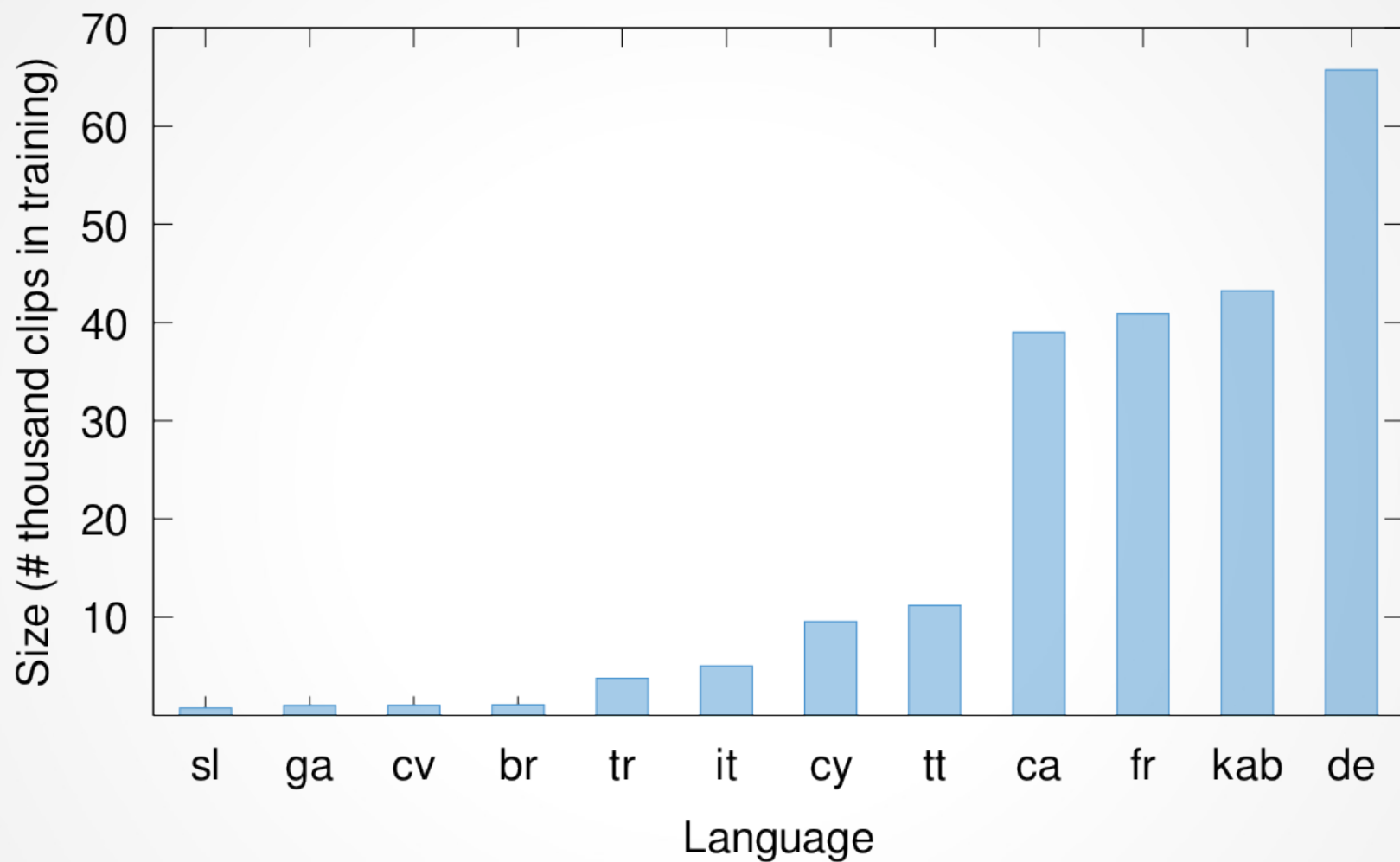
Experimental Design

- 5 depths for slicing source model
- 2 update scenarios (frozen vs. fine-tuned)
- 12 target languages
- 120 experiments, in total

Hyperparameters

- Single GPU training
- 24 train batch, 48 dev batch
- 20% dropout rate
- 0.0001 learning rate with ADAM
- Early stopping based on last 5 steps

Data (Spoken Corpora)



Frozen Transfer Results

Lang.	Character Error Rate					
	Number of Layers Copied from English					
	None	1	2	3	4	5
sl	23.35	23.93	25.30	18.87	17.53	26.24
ga	31.83	29.08	36.14	27.22	29.07	32.27
cv	48.10	46.13	47.83	38.00	35.23	42.88
br	21.47	19.17	20.76	18.33	17.72	21.03
tr	34.66	32.98	35.47	33.00	33.66	36.71
it	40.91	39.20	41.55	38.16	39.40	43.21
cy	34.15	32.46	33.93	31.57	35.26	36.56
tt	32.61	29.20	30.52	27.37	28.28	31.28
ca	38.01	36.44	38.70	36.51	42.26	47.96
fr	43.33	43.30	43.47	43.37	43.75	43.79
kab	25.76	25.57	25.97	25.45	27.77	29.28
de	43.76	44.48	44.08	43.70	43.77	43.69

Table 2. Frozen Transfer Learning Character-error rates (CER)

Frozen Transfer Results

Lang.	Character Error Rate					
	Number of Layers Copied from English					
	None	1	2	3	4	5
sl	23.35	23.93	25.30	18.87	17.53	26.24
ga	31.83	29.08	36.14	27.22	29.07	32.27
cv	48.10	46.13	47.83	38.00	35.23	42.88
br	21.47	19.17	20.76	18.33	17.72	21.03
tr	34.66	32.98	35.47	33.00	33.66	36.71
it	40.91	39.20	41.55	38.16	39.40	43.21
cy	34.15	32.46	33.93	31.57	35.26	36.56
tt	32.61	29.20	30.52	27.37	28.28	31.28
ca	38.01	36.44	38.70	36.51	42.26	47.96
fr	43.33	43.30	43.47	43.37	43.75	43.79
kab	25.76	25.57	25.97	25.45	27.77	29.28
de	43.76	44.48	44.08	43.70	43.77	43.69

Table 2. Frozen Transfer Learning Character-error rates (CER)

Frozen Transfer Results

Character Error Rate						
Number of Layers Copied from English						
Lang.	None	1	2	3	4	5
sl	23.35	23.93	25.30	18.87	17.53	26.24
ga	31.83	29.08	36.14	27.22	29.07	32.27
cv	48.10	46.13	47.83	38.00	35.23	42.88
br	21.47	19.17	20.76	18.33	17.72	21.03
tr	34.66	32.98	35.47	33.00	33.66	36.71
it	40.91	39.20	41.55	38.16	39.40	43.21
cy	34.15	32.46	33.93	31.57	35.26	36.56
tt	32.61	29.20	30.52	27.37	28.28	31.28
ca	38.01	36.44	38.70	36.51	42.26	47.96
fr	43.33	43.30	43.47	43.37	43.75	43.79
kab	25.76	25.57	25.97	25.45	27.77	29.28
de	43.76	44.48	44.08	43.70	43.77	43.69

Table 2. Frozen Transfer Learning Character-error rates (CER)

Fine-Tuning Transfer Results

Lang.	Character Error Rate					
	Number of Layers Copied from English					
	None	1	2	3	4	5
sl	23.35	21.65	26.44	19.09	15.35	17.96
ga	31.83	31.01	32.2	27.5	25.42	24.98
cv	48.1	47.1	44.58	42.75	27.21	31.94
br	21.47	19.16	20.01	18.06	15.99	18.42
tr	34.66	34.12	34.83	31.79	27.55	29.74
it	40.91	42.65	42.82	36.89	33.63	35.10
cy	34.15	31.91	33.63	30.13	28.75	30.38
tt	32.61	31.43	30.80	27.79	26.42	28.63
ca	38.01	35.21	39.02	35.26	33.83	36.41
fr	43.33	43.26	43.51	43.24	43.20	43.19
kab	25.76	25.5	26.83	25.25	24.92	25.28
de	43.76	43.69	43.62	43.60	43.76	43.69

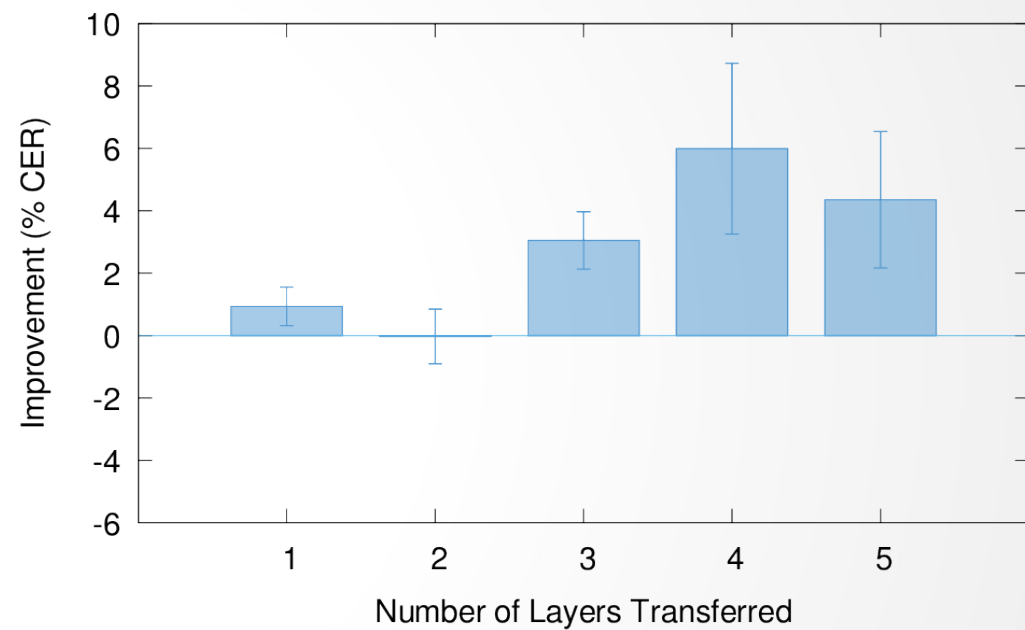
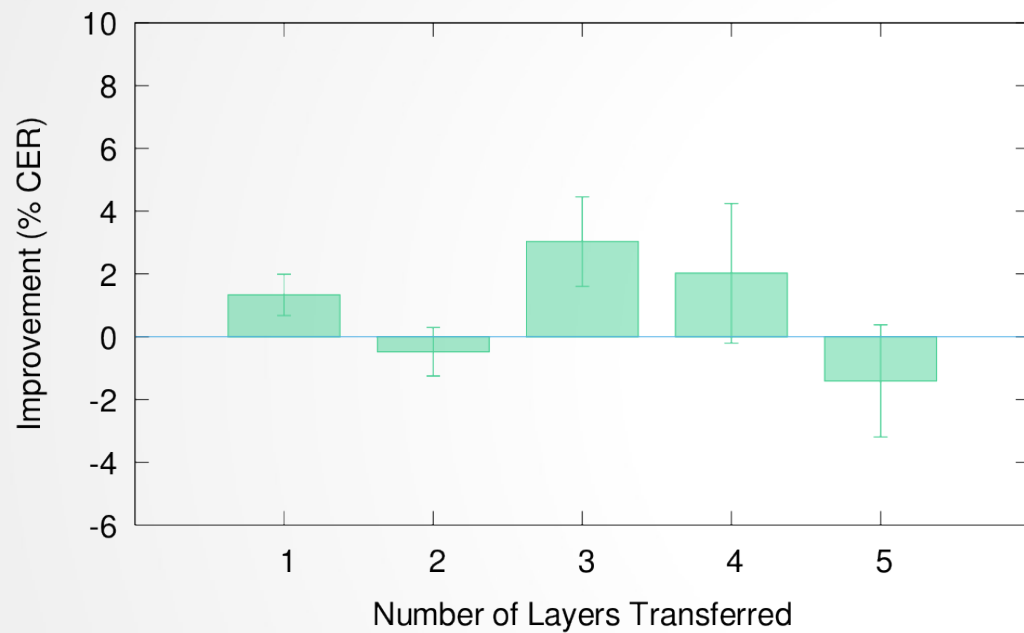
Table 3. Fine-Tuned Transfer Learning Character-error rates (CER)

Fine-Tuning Transfer Results

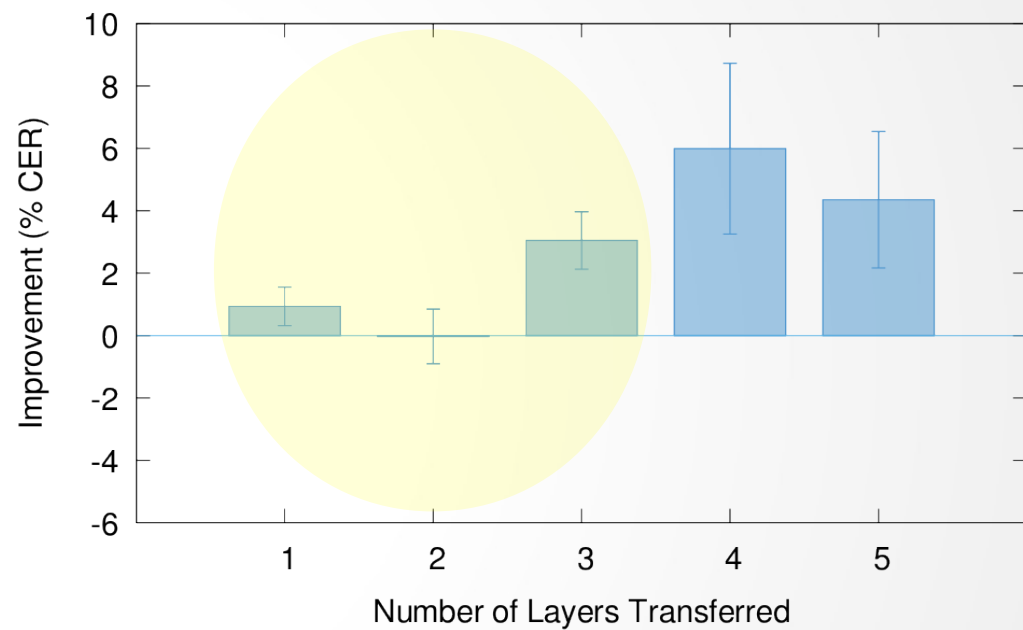
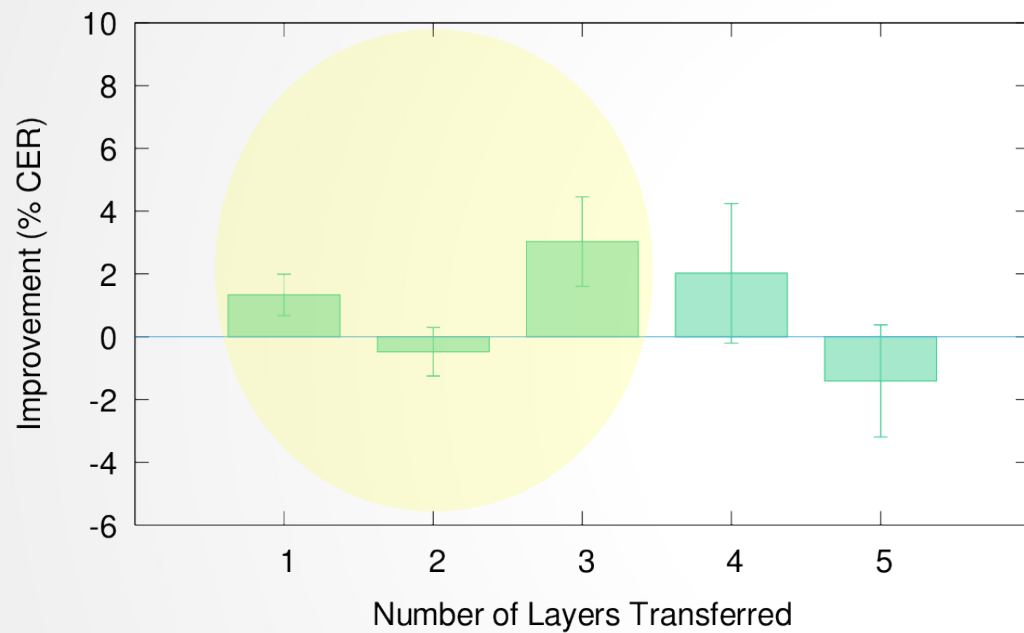
Lang.	Character Error Rate					
	Number of Layers Copied from English					
	None	1	2	3	4	5
sl	23.35	21.65	26.44	19.09	15.35	17.96
ga	31.83	31.01	32.2	27.5	25.42	24.98
cv	48.1	47.1	44.58	42.75	27.21	31.94
br	21.47	19.16	20.01	18.06	15.99	18.42
tr	34.66	34.12	34.83	31.79	27.55	29.74
it	40.91	42.65	42.82	36.89	33.63	35.10
cy	34.15	31.91	33.63	30.13	28.75	30.38
tt	32.61	31.43	30.80	27.79	26.42	28.63
ca	38.01	35.21	39.02	35.26	33.83	36.41
fr	43.33	43.26	43.51	43.24	43.20	43.19
kab	25.76	25.5	26.83	25.25	24.92	25.28
de	43.76	43.69	43.62	43.60	43.76	43.69

Table 3. Fine-Tuned Transfer Learning Character-error rates (CER)

Frozen vs. Fine-Tuned

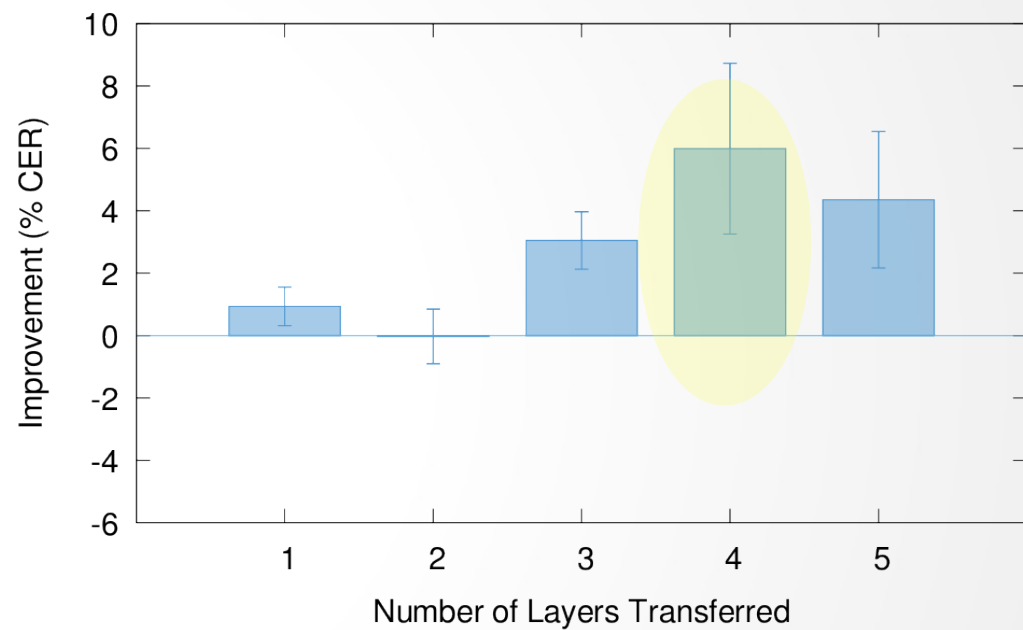
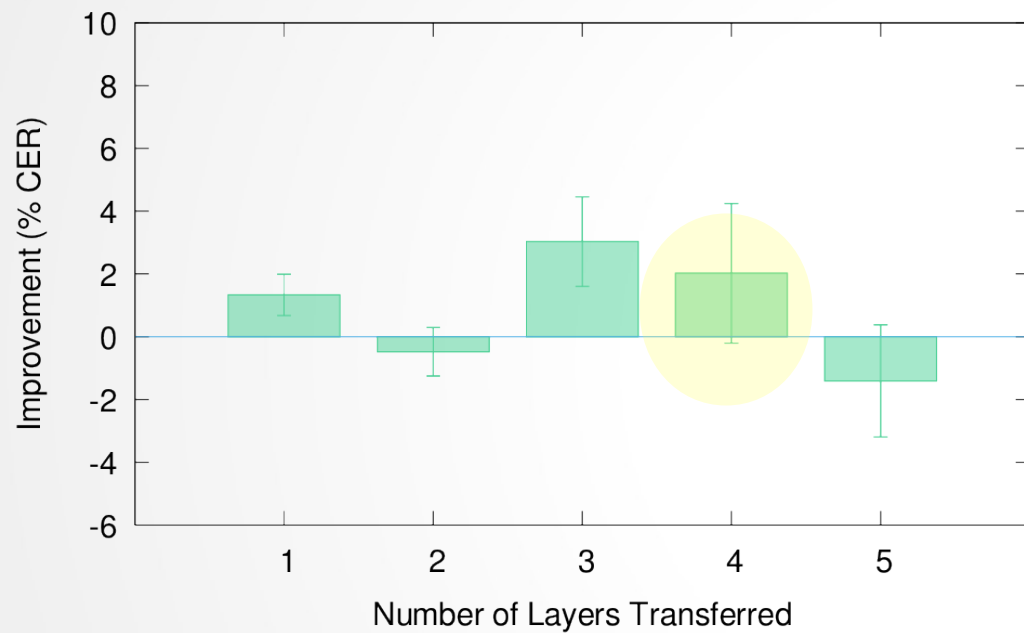


Frozen vs. Fine-Tuned

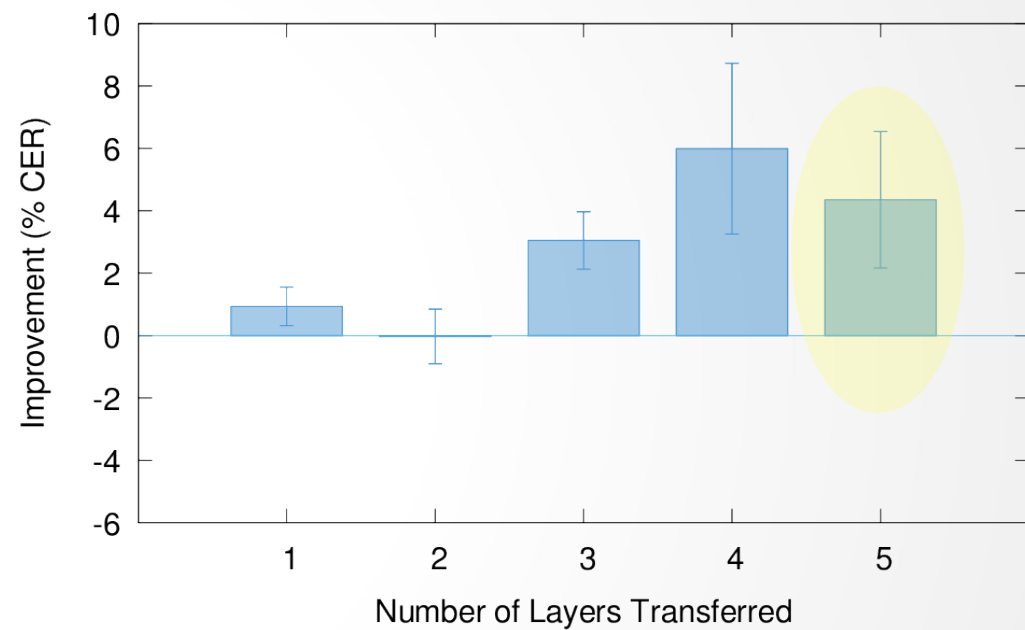
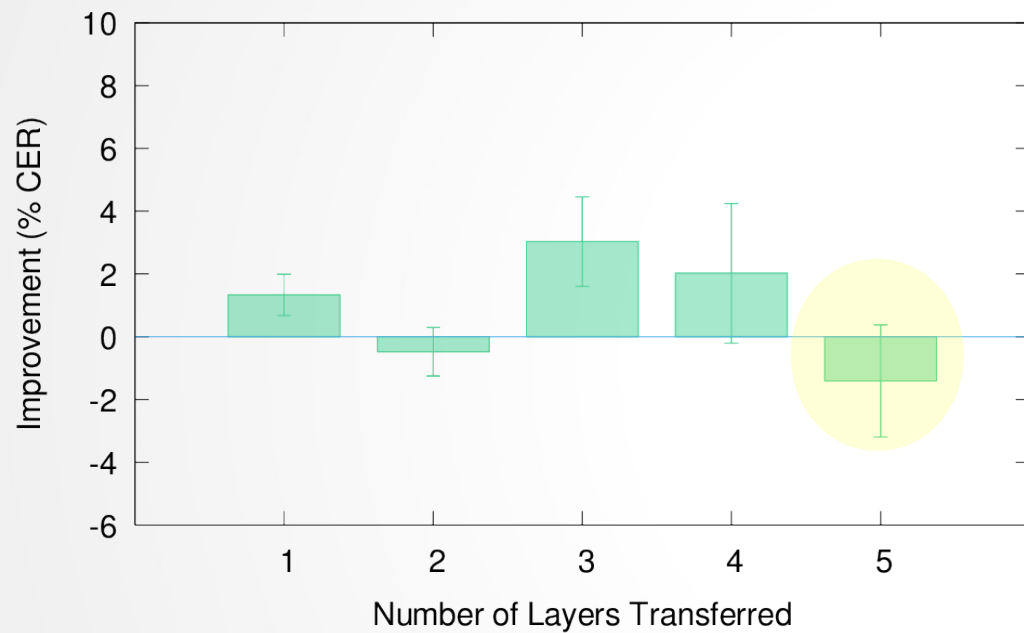


Frozen vs. Fine-Tuned

LSTM!



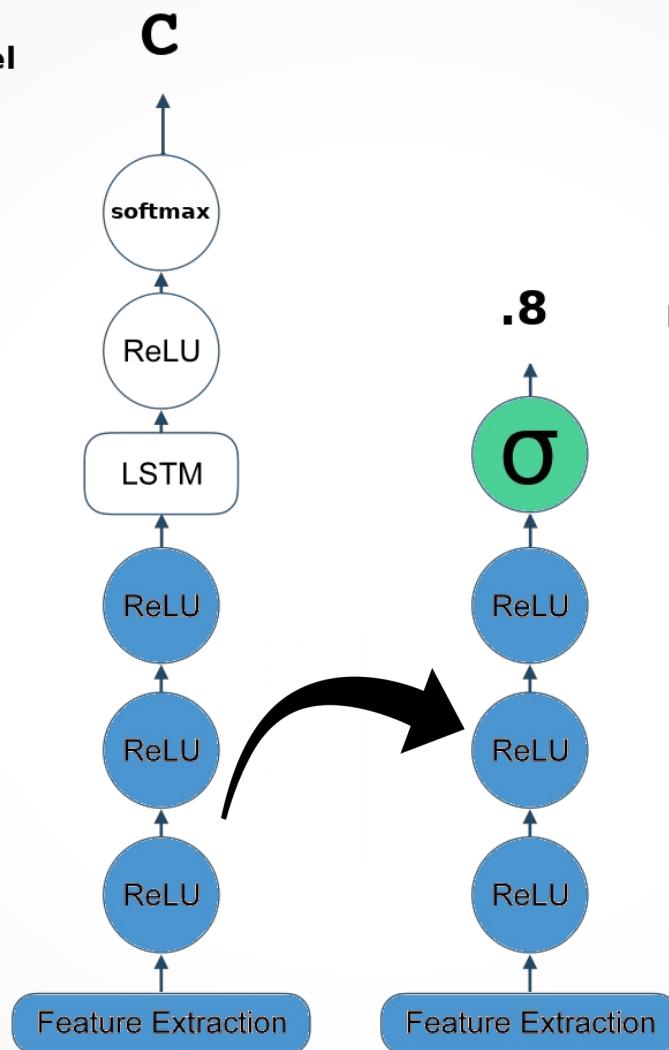
Frozen vs. Fine-Tuned



Interpretability Experiments

Regression on Embeddings

CTC ASR
Source Model



Logistic Regression
Target Task

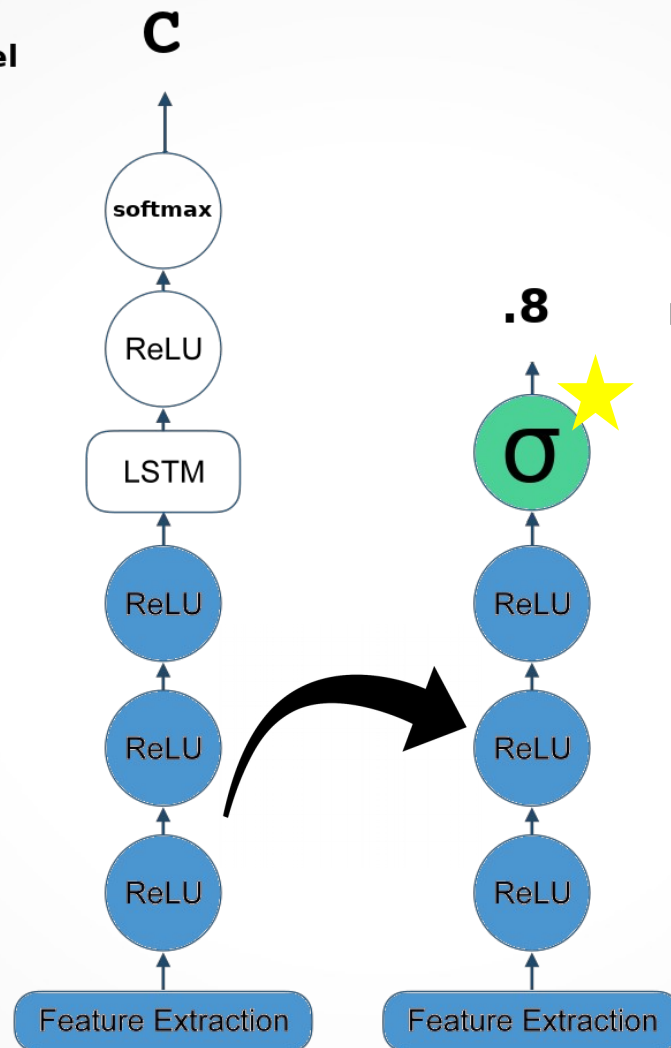
.8
 σ



Feature Extraction

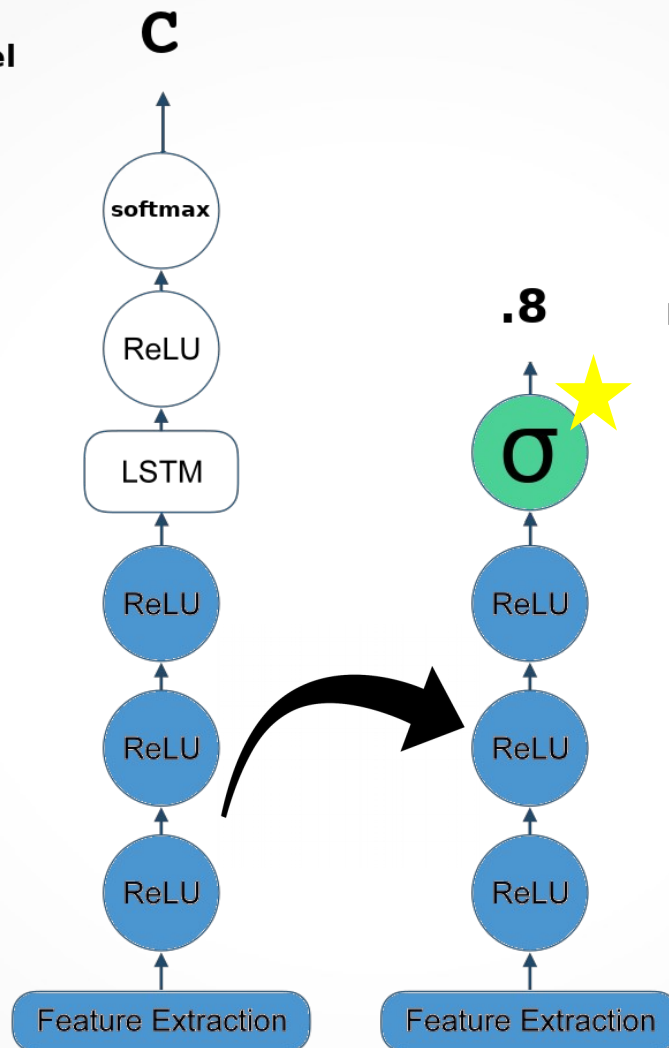
Regression on Embeddings

CTC ASR
Source Model



Regression on Embeddings

**CTC ASR
Source Model**



**Logistic Regression
Target Task**

Trained for 3 epochs
w/ Cross Entropy Loss

Regression Results

Speech vs. Noise

- Copied layers, added final FC layer with single output and logistic activation
- 13 languages vs. UrbanSound8k
- 5,005 train clips, 442 test clips per class

Regression Results

Classification Accuracy					
Number of Layers Copied from English					
1	2	3	4	5	6
51.01	93.68	92.82	95.30	94.55	93.53

Table 4. Speech vs. Non-Speech Audio Classification Accuracy

- Copied layers, added final FC layer with single output and logistic activation
- 13 languages vs. UrbanSound8k
- 5,005 train clips, 442 test clips per class

Regression Results

Classification Accuracy					
Number of Layers Copied from English					
1	2	3	4	5	6
51.01	93.68	92.82	95.30	94.55	93.53

Table 4. Speech vs. Non-Speech Audio Classification Accuracy

- Copied layers, added final FC layer with single output and logistic activation
- 13 languages vs. UrbanSound8k
- 5,005 train clips, 442 test clips per class

Regression Results

English vs. German

Regression Results

English vs. German

- Copied layers, added final FC layer with single output and logistic activation
- English vs. German
- 5,000 train clips, 500 test clips per class

Regression Results

English vs. German

Classification Accuracy					
Number of Layers Copied from English					
1	2	3	4	5	6
66.51	66.38	52.77	86.21	74.97	85.00

Table 5. English vs. German Audio Classification Accuracy (%)

Regression Results

Classification Accuracy					
Number of Layers Copied from English					
1	2	3	4	5	6
51.01	93.68	92.82	95.30	94.55	93.53

Table 4. Speech vs. Non-Speech Audio Classification Accuracy

Classification Accuracy					
Number of Layers Copied from English					
1	2	3	4	5	6
66.51	66.38	52.77	86.21	74.97	85.00

Table 5. English vs. German Audio Classification Accuracy (%)

Discussion

Discussion

1) Transfer in ASR

- Fine-tuning always helps
- LSTM transfer is best, but only with fine-tuning

2) Interpretability Studies

- At the third layer, the model has learned general speech, but language-agnostic representations

Thank you for your attention!

APPENDIX A: Multi-Task

Linguistic Knowledge

Example: Collapsing on Voice

B P	--> P	bilabial plosives
CH JH	--> CH	alveo-palatal affricates
D T	--> T	alveolar plosives
DH TH	--> TH	interdental fricatives
F V	--> F	labio-dental fricatives
G K	--> G	velar plosives
S Z	--> S	alveolar fricatives
SH ZH	--> SH	alveo-palatal fricatives

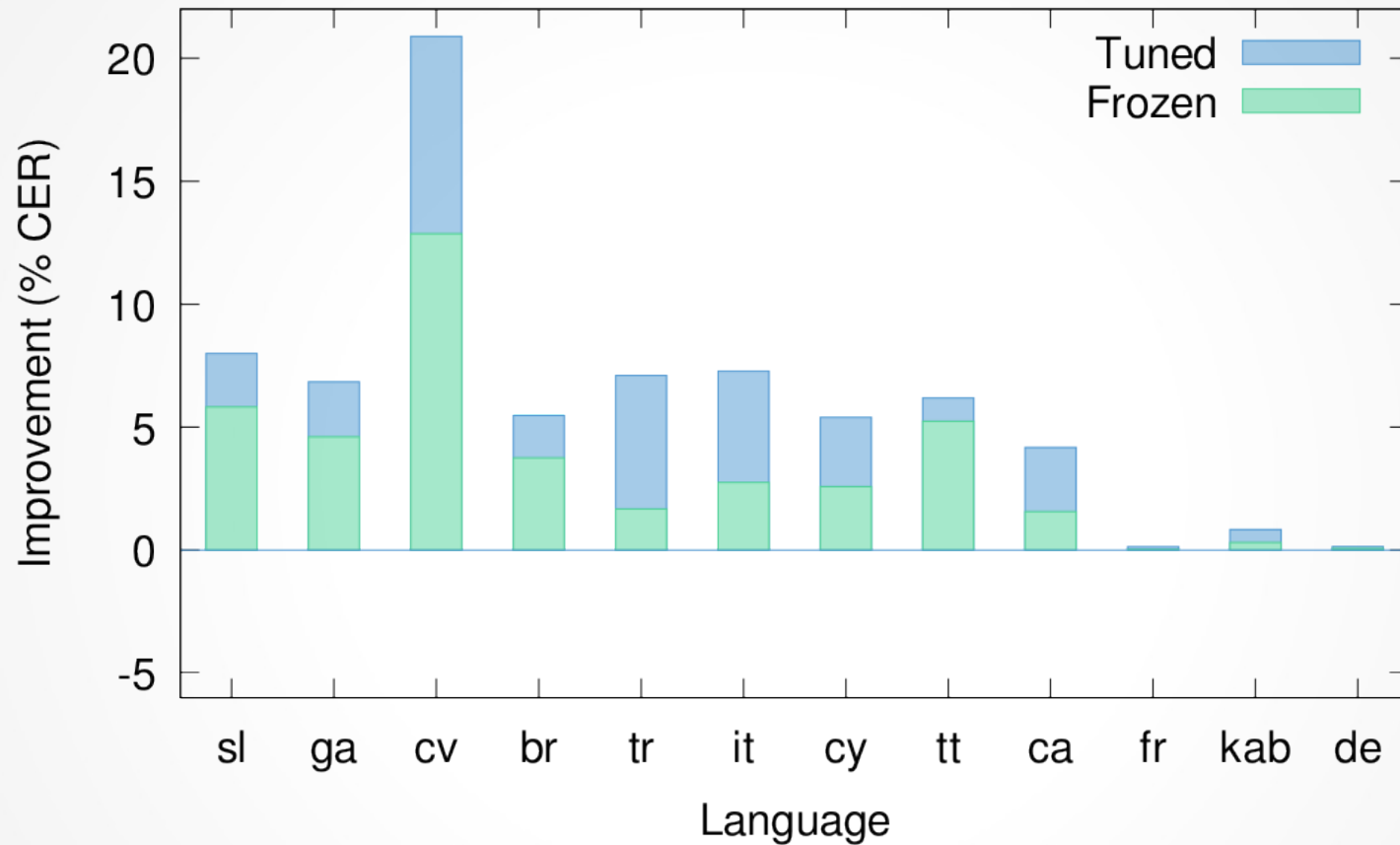
APPENDIX B: DeepSpeech

Data Details

Language	Code	Dataset Size					
		Audio Clips			Unique Speakers		
		Dev	Test	Train	Dev	Test	Train
Slovenian	sl	110	213	728	1	12	3
Irish	ga	181	138	1001	4	12	6
Chuvash	cv	96	77	1023	4	12	5
Breton	br	163	170	1079	3	15	7
Turkish	tr	407	374	3771	32	89	32
Italian	it	627	734	5019	29	136	37
Welsh	cy	1235	1201	9547	51	153	75
Tatar	tt	1811	1164	11187	9	64	3
Catalan	ca	5460	5037	38995	286	777	313
French	fr	5083	4835	40907	237	837	249
Kabyle	kab	5452	4643	43223	31	169	63
German	de	7982	7897	65745	247	1029	318

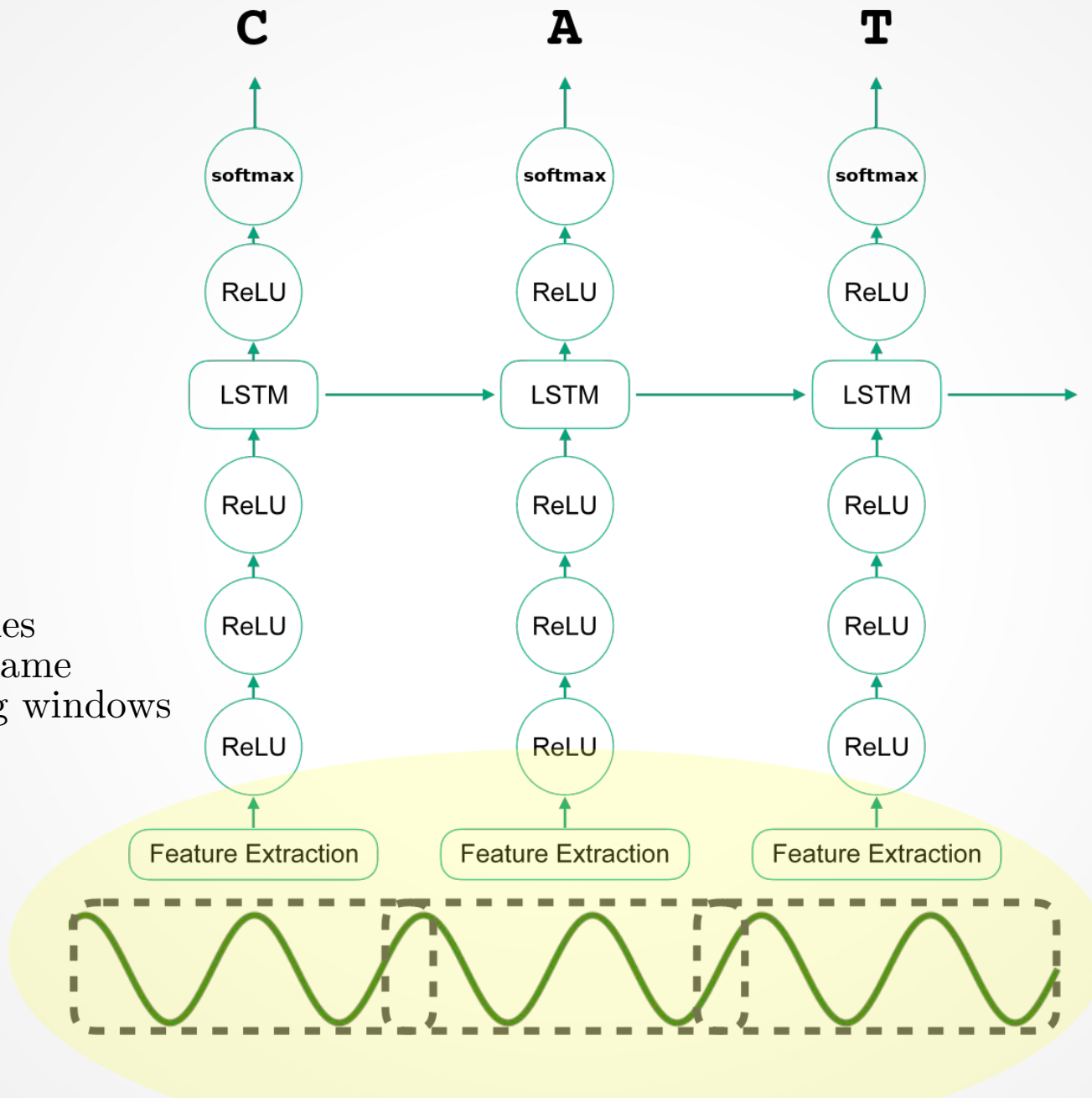
Table 1. Number of audio clips and unique speakers per language per dataset split.

Effect of Data Size



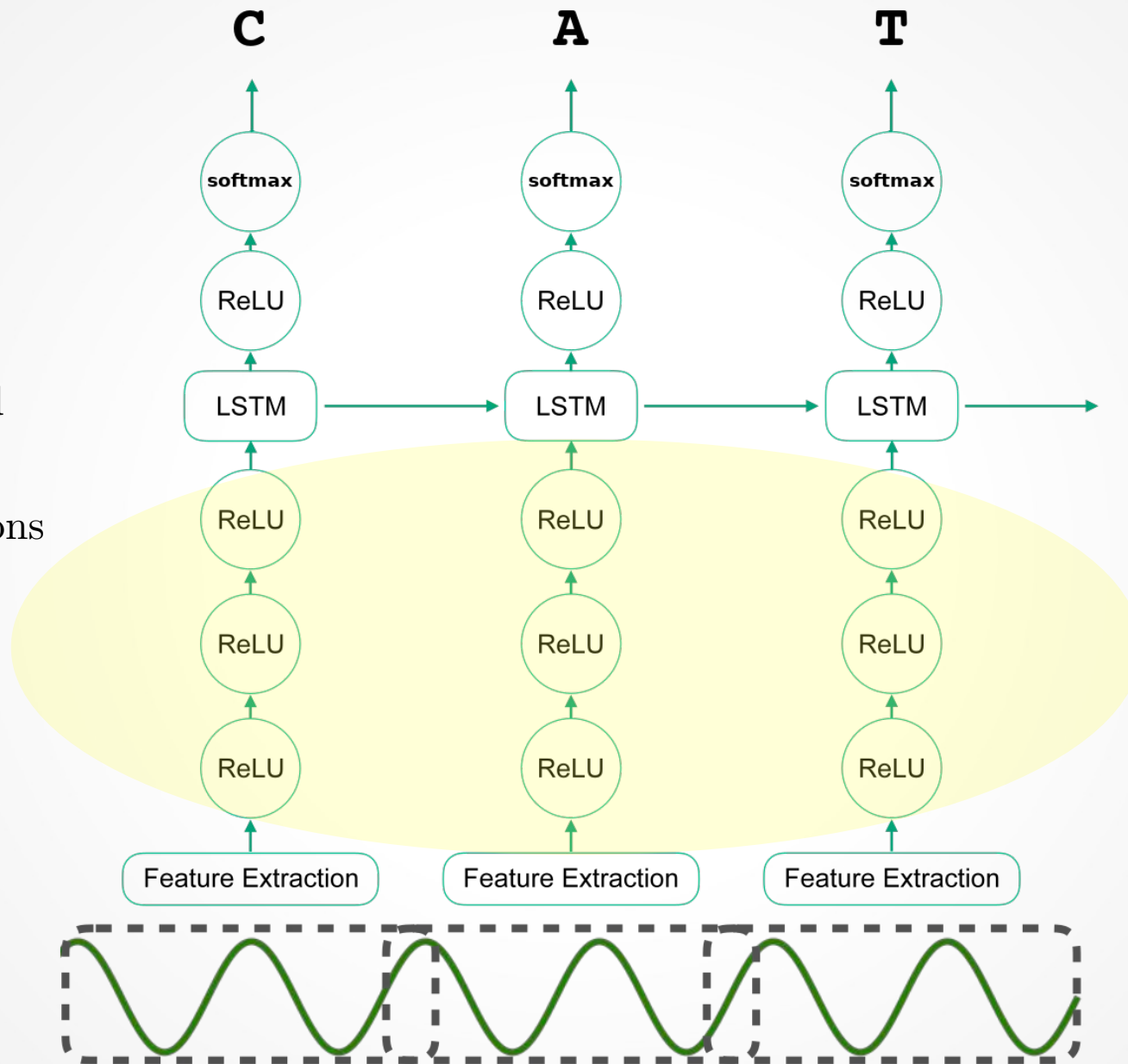
Model Architecture

19 spliced frames
26 MFCCs / frame
32ms Hamming windows
20ms timestep



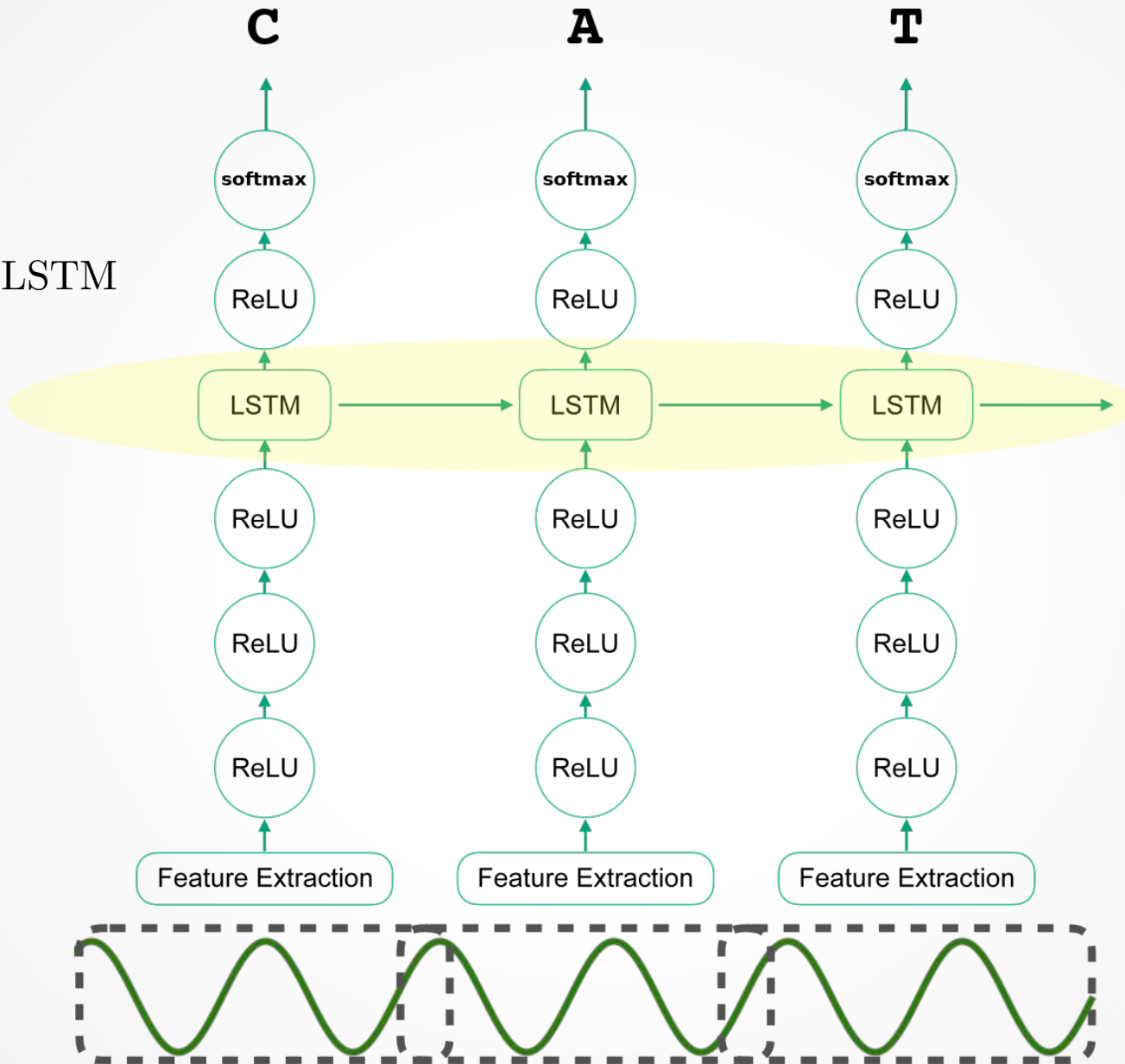
Model Architecture

Fully connected
Feed-Forward
2048 dims
ReLU activations

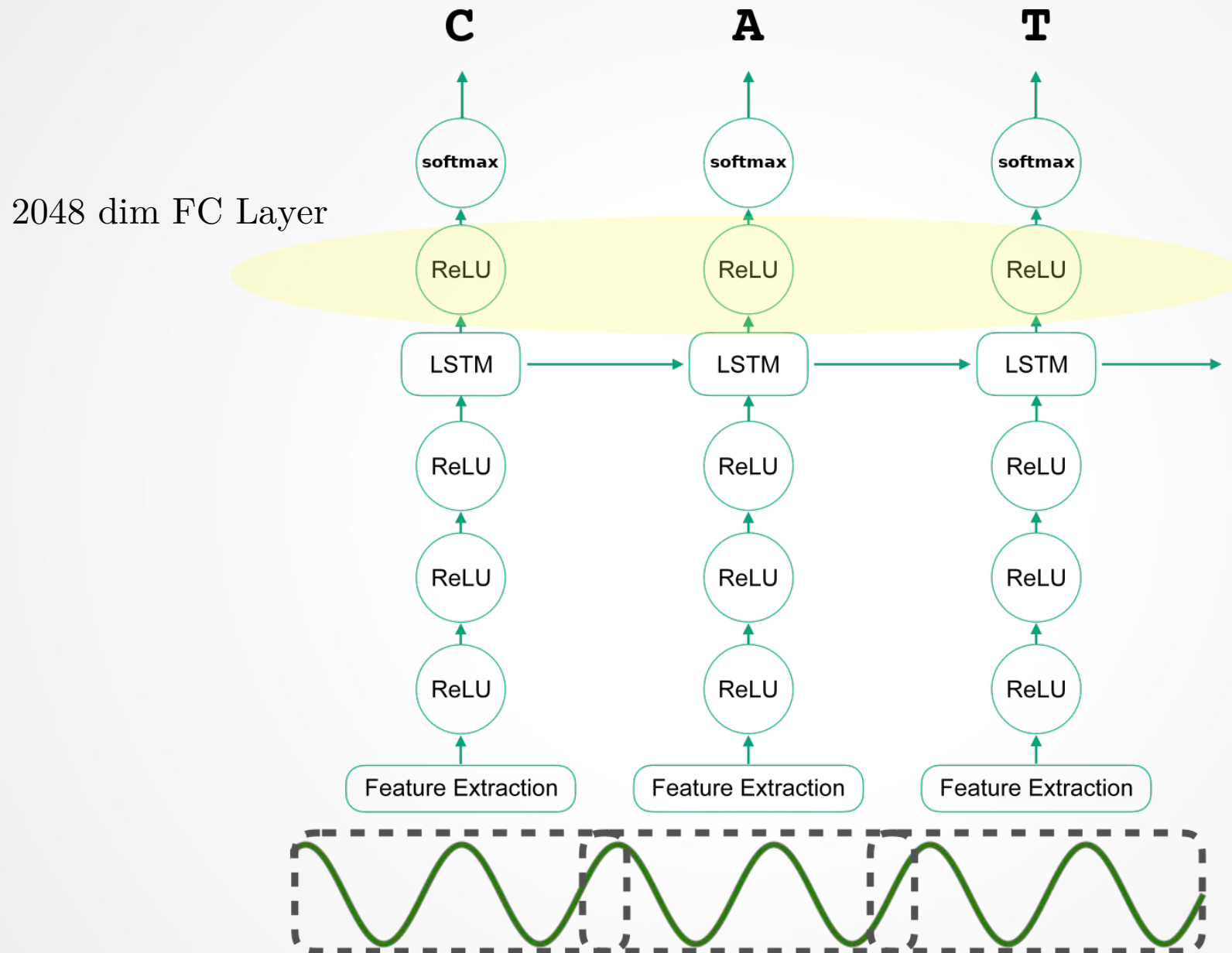


Model Architecture

Unidirectional LSTM
2048 dims



Model Architecture



Model Architecture

