

Interpretable Knowledge Transfer for Machine Learning of Speech Tasks

Peter Plantinga

It takes a village...

Thanks to:

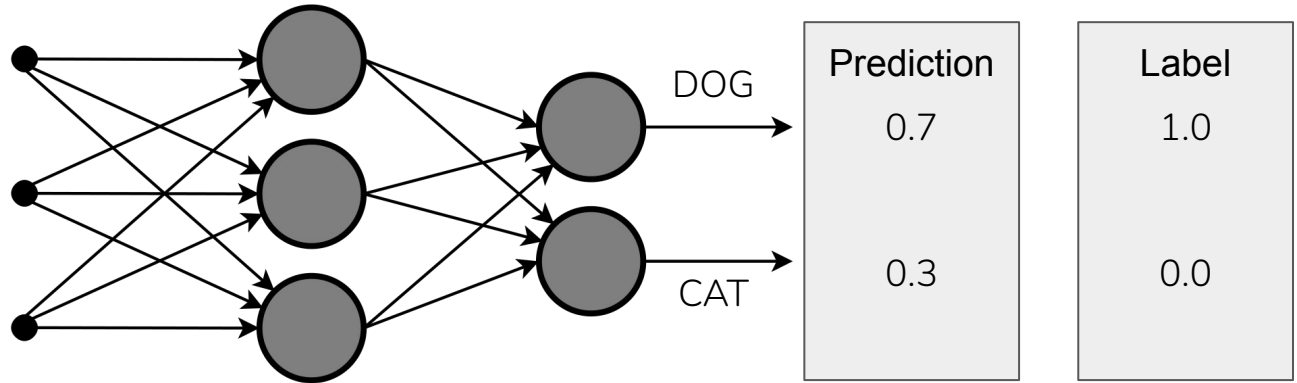
- You (seriously, thanks for coming!)
- Eric Fosler-Lussier
- SLaTe labmates
- SpeechBrain team
- OSU, NSF, Mila, Nvidia, OSC, etc.
- Family and friends

Introduction

Machine learning is often done with only one kind of feedback: direct comparison of prediction and label



x



Prediction
0.7
0.3

\hat{y}

Label
1.0
0.0

y

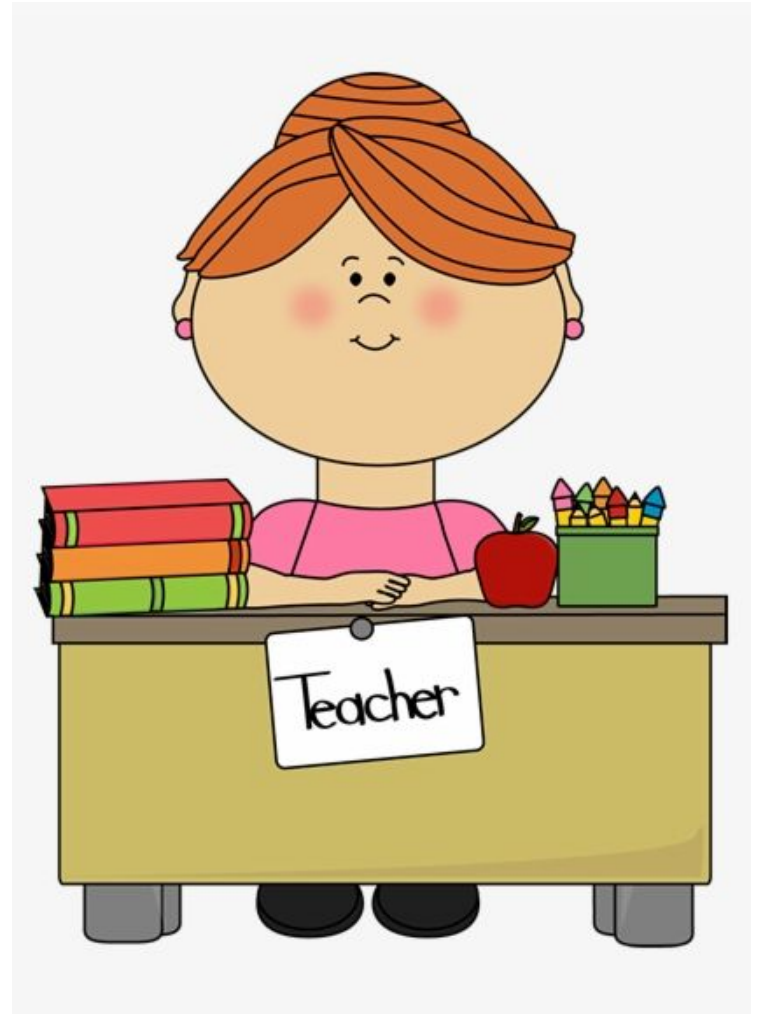
Introduction

By contrast, human educators provide more than just solutions.

Some examples:

- Explanations
- Comparisons
- Illustrations

(Vapnik and Vashist 2009)



Introduction

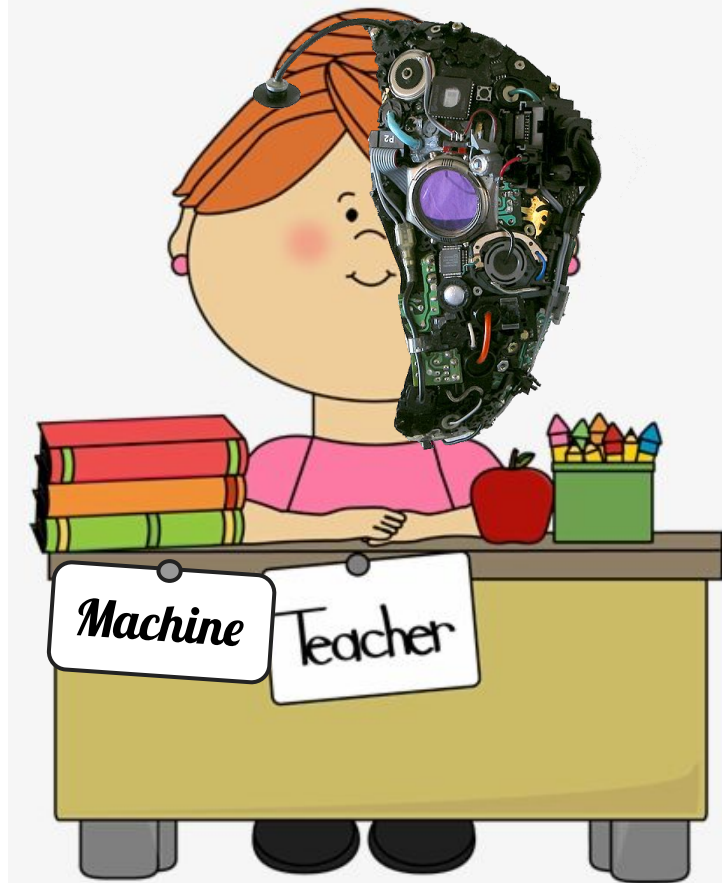
Turns out teacher models can help machines to learn too!

A teacher model adds “privileged information” to each sample (x^*):

$$(x_1, x_1^*, y_1), (x_2, x_2^*, y_2), \dots, (x_n, x_n^*, y_n)$$

Our work does this in new ways!

(Vapnik and Vashist 2009)



Outline

Approaches to knowledge transfer

New ways to use knowledge transfer:

1. For removing noise from speech recordings
2. For teaching kids how to read

Future work and conclusions

Outline

Approaches to knowledge transfer

New ways to use knowledge transfer:

1. For removing noise from speech recordings
2. For teaching kids how to read

Future work and conclusions

Knowledge Transfer

There are two main approaches to knowledge transfer

Provide better **labels**

Instead of “hard” binary labels, target “soft” labels with information such as confidence and relative likelihood between classes

E.g.



Cat? Dog? Bird? Horse?

“hard” binary label

[0.0, 1.0, 0.0, 0.0]

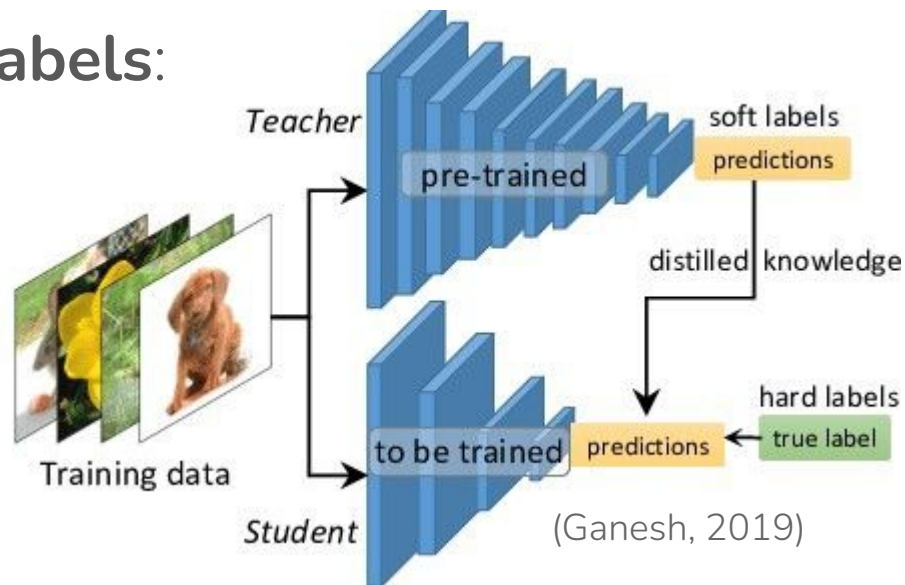
“soft” label provided by teacher

[0.1, 0.8, 0.03, 0.07]

Knowledge Transfer

Examples of providing better **labels**:

- Teacher-student learning
(Buciluă, Caruana, and Alexandru 2006)
- Knowledge distillation
(Hinton, Vinyals, and Dean 2015)



First used for **model compression**

Knowledge Transfer

There are two main approaches to knowledge transfer



Humans



Pixel-by-pixel mapping

Mapped by ConvNet



Provide better **objectives**

Map predictions and/or targets to alternate space for better comparison.

Knowledge Transfer

Examples of providing better **objectives**:

- Perceptual Loss (Gatys et al. 2016)

(Thompson 2019)



- Generative adversarial networks (Goodfellow et al. 2014)

Outline

Approaches to knowledge transfer

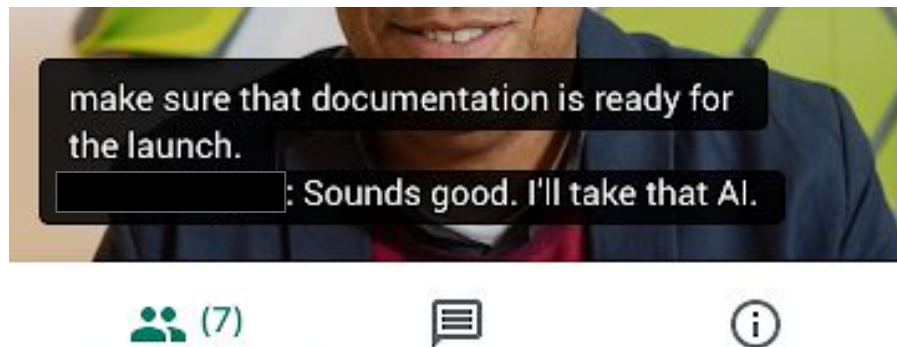
New ways to use knowledge transfer:

- 1. For removing noise from speech recordings**
2. For teaching kids how to read

Future work and conclusions

Enhancement for Intelligibility

Video conference calls (among other technologies) can benefit from speech enhancement for improving signal intelligibility.



- Usually improves human recognition (Healy et al. 2017)
- Sometimes improves machine recognition (Narayanan and Wang 2015)

Speech Recognizability

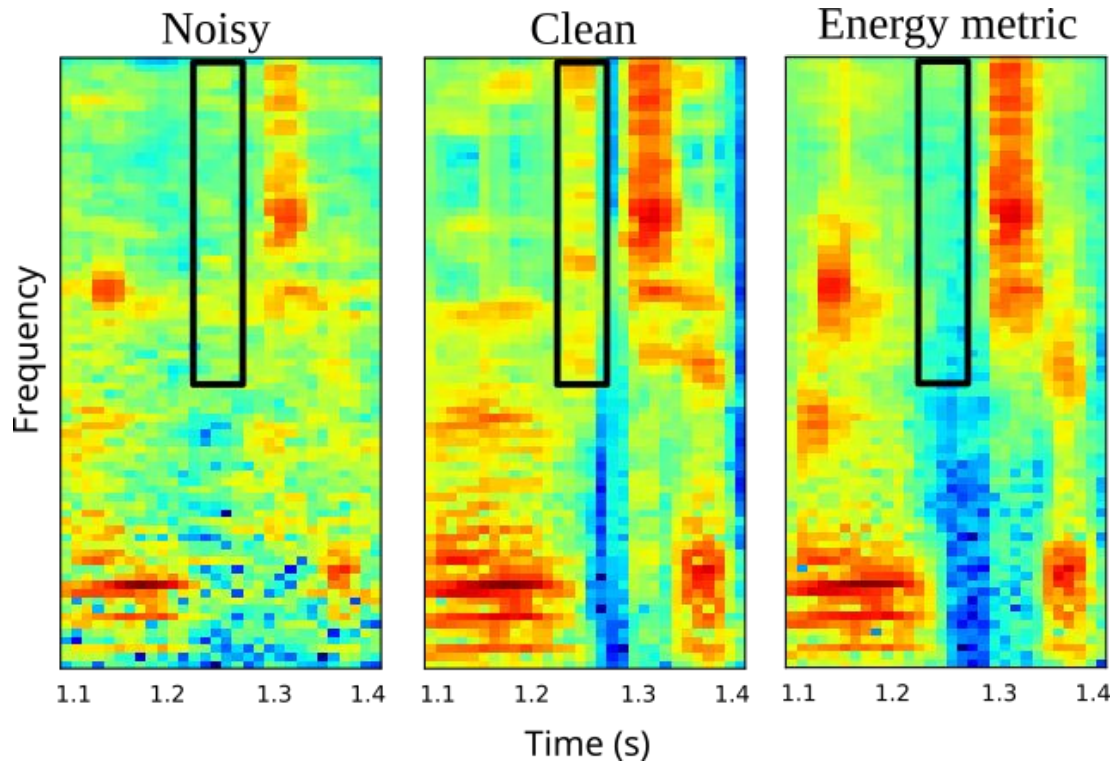
Enhancement and recognition objectives sometimes conflict

- Energy-based metrics don't use phonetic labels.
- ASR needs phonetic cues, but some phones are low-energy



Speech Recognizability

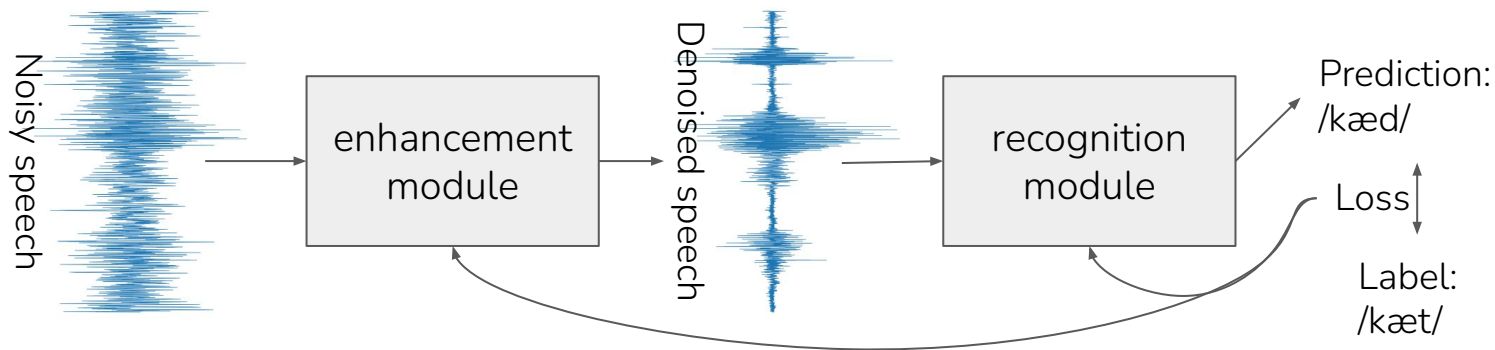
Low-energy phonemes (like /v/ or /θ/) are sometimes ignored by enhancement models trained with an energy-based metric. (Plantinga et al. 2020)



Speech Recognizability

Popular idea — train modules at the same time
i.e. “joint training” or “end-to-end training”

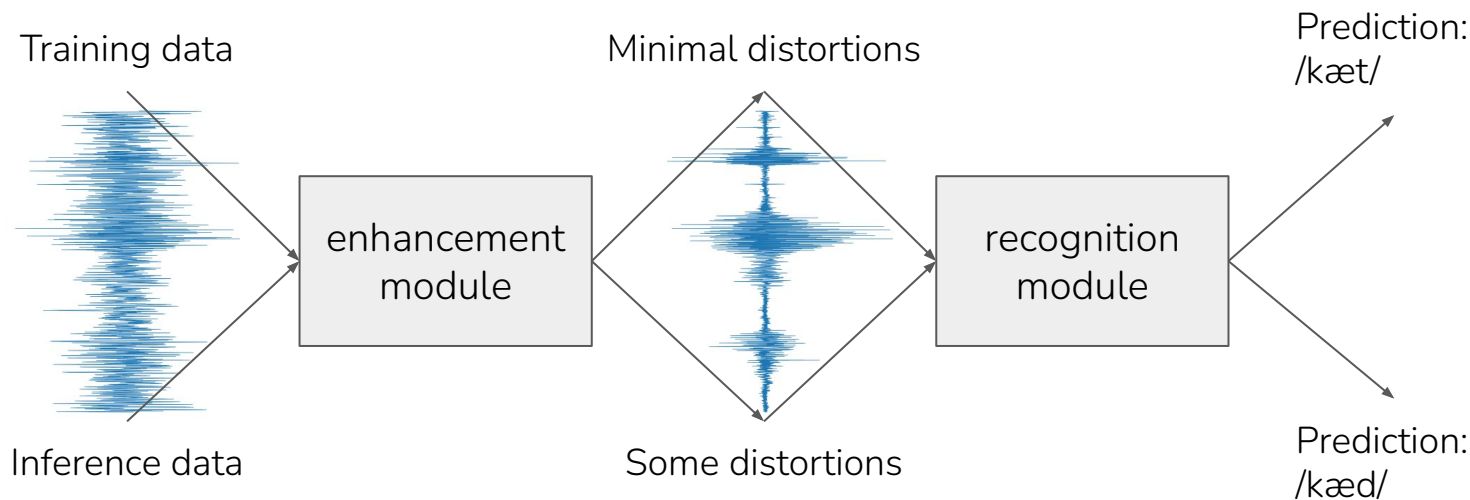
(Narayanan and Wang 2015)



Speech Recognizability

Joint training issue: “distortion problem”

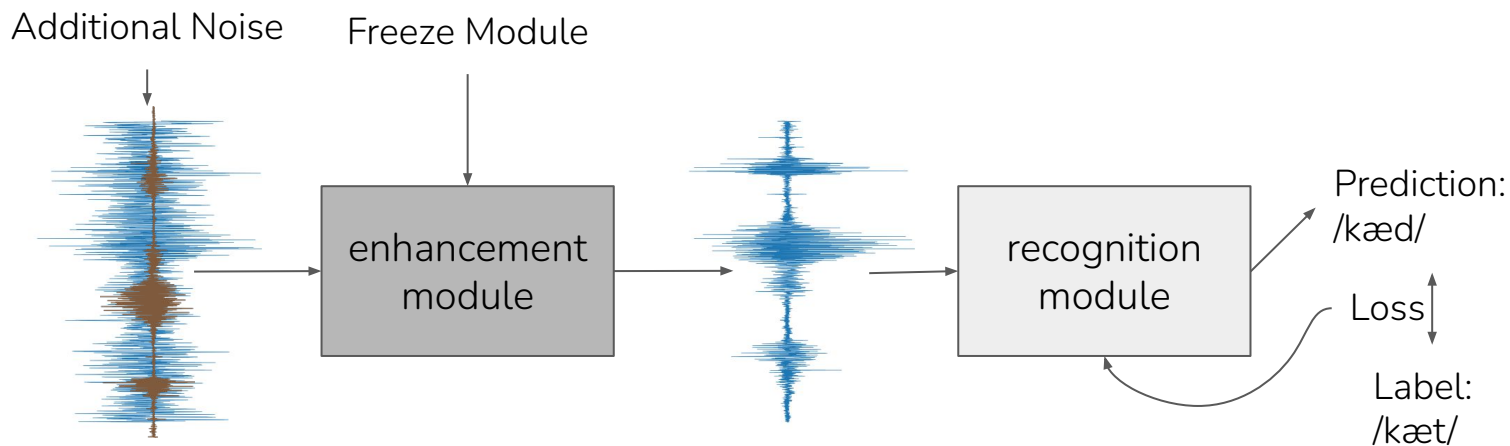
(Wang, Tan, and Wang 2019)



Speech Recognizability

A solution to the “distortion problem”

(Wang, Tan, and Wang 2019)

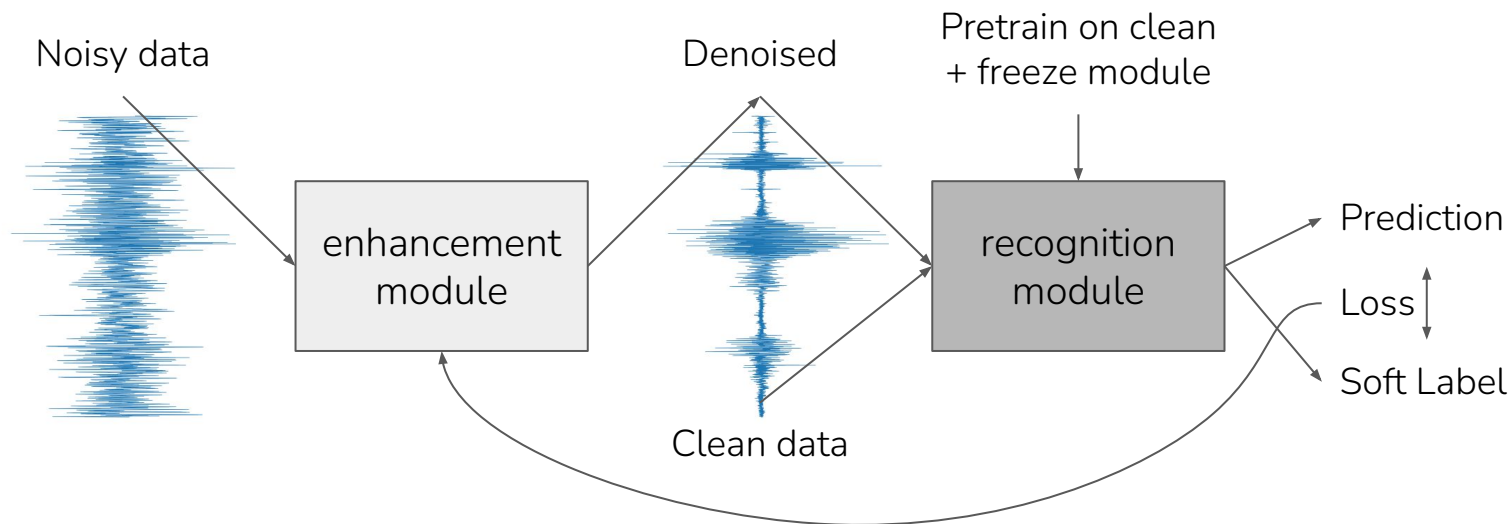


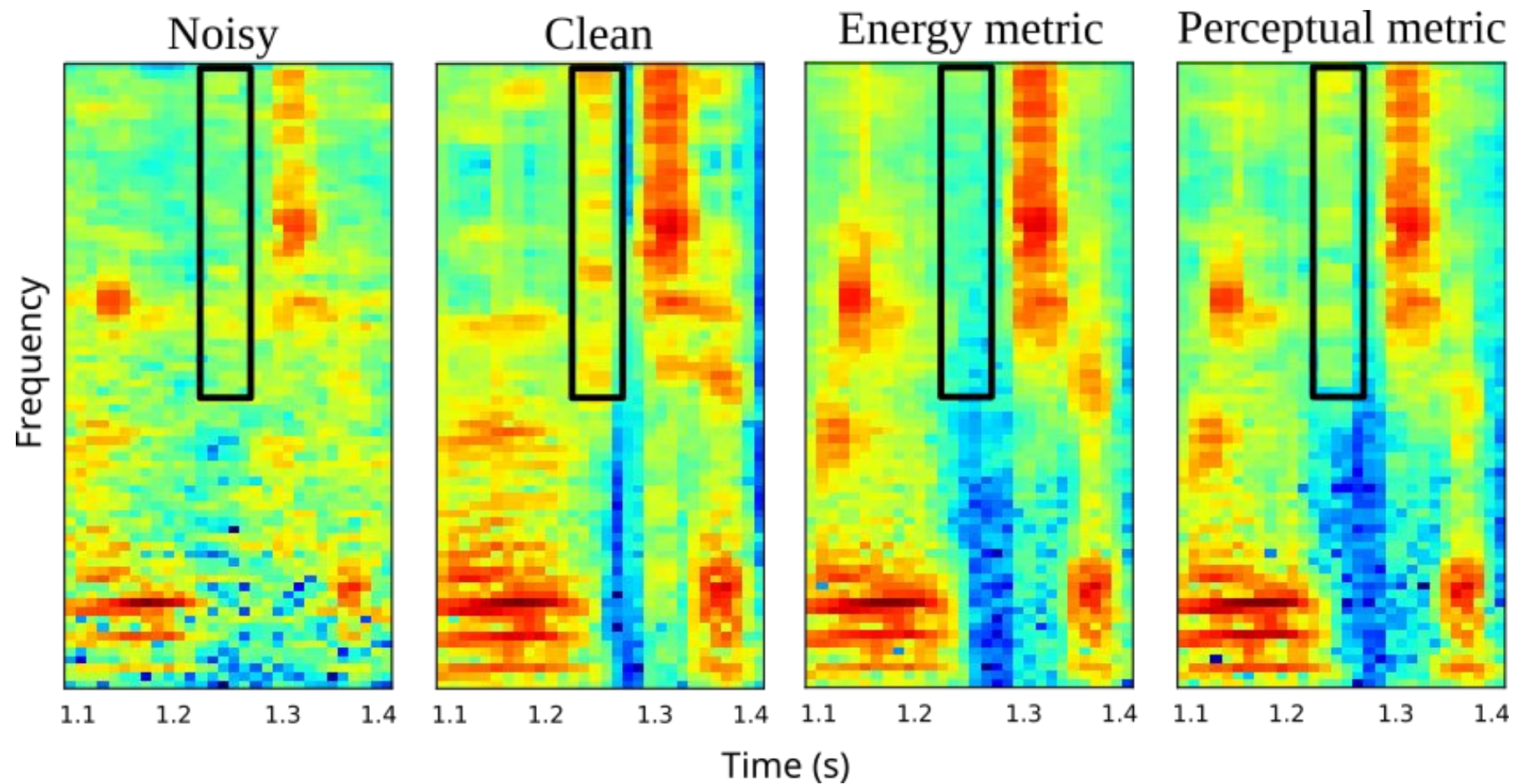
But we lose the benefits of joint training!

Speech Recognizability

Our proposal is to use a perceptual loss

(Bagchi, Plantinga, Stiff, and Fosler-Lussier 2018)





Samples w/ Perceptual Loss



Babble noise:

Noisy



Enhanced



Clean



Restaurant:

Noisy



Enhanced



Clean

Experiments

We run experiments on three datasets.

- CHiME-2 (ASR)
 - Recordings of reading WSJ articles with living room noise and reverb
 - Our enhancement and perceptual models were similar to Wide ResNet
 - We used an off-the-shelf Kaldi recipe to evaluate recognition rates



Experiments

We run experiments on three datasets.

- CHiME-2 (ASR)
- CHiME-4 (enhancement)
 - Includes both real and simulated noisy recordings
 - Enhancement model was state-of-the-art for time-domain (AECNN)
 - We tested in difficult scenario where no simulated data available



Experiments

We run experiments on three datasets.

- CHiME-2 (ASR)
- CHiME-4 (enhancement)
- Voicebank + DEMAND (both ASR and enhancement)
 - Diverse set of voices with wide variety of environmental noises
 - Direct comparison against joint training (possible with SpeechBrain)
 - Recipe made public with this public toolkit and public dataset



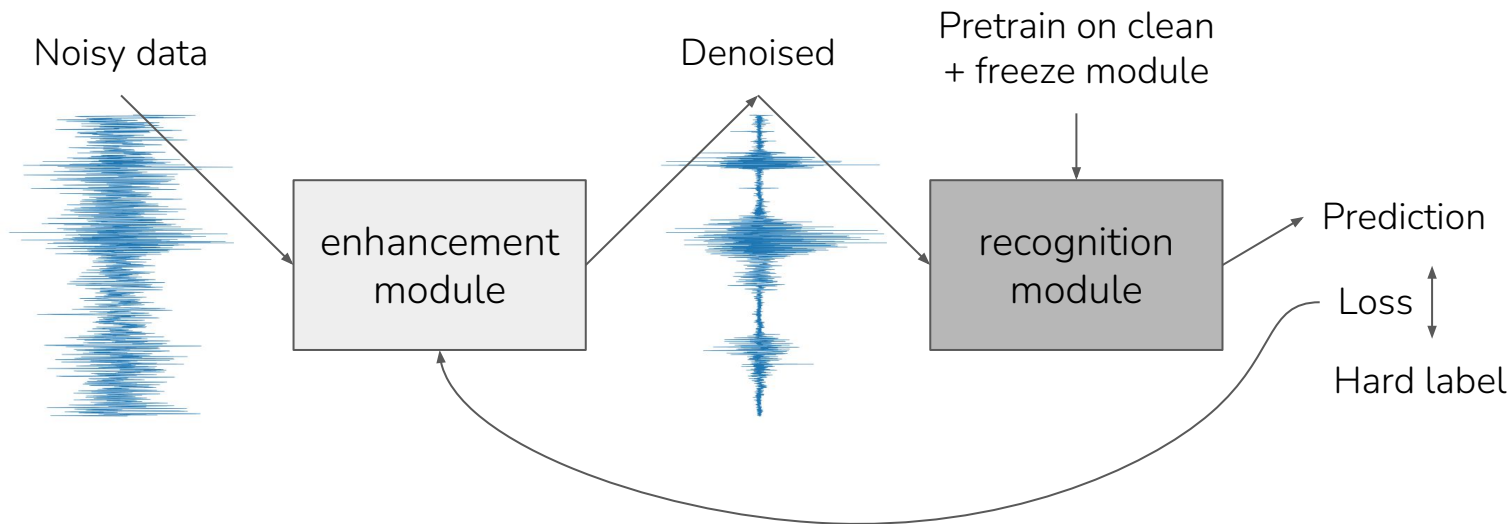
CHiME-2 (Ours vs SotA)

Enhancement model	Joint Training	Extra Features	WER error rate
Noisy (no enhancement)	-	-	17.4
CNN (Chen et al. 2015)	Yes	-	16.0
DNN (Narayanan and Wang 2015)	Yes	Yes	15.4
LSTM (Weninger et al. 2015)	-	Yes	13.8
Joint Training (Wang and Wang 2016)	Yes	Yes	10.6
Wide ResNet (Plantinga et al. 2018)	-	-	10.8
Wide ResNet + Perceptual Loss	-	-	8.7

No Parallel Clean & Noisy Data!

Perceptual loss works without access to parallel speech data

(Plantinga, Bagchi, and Fosler-Lussier 2020)



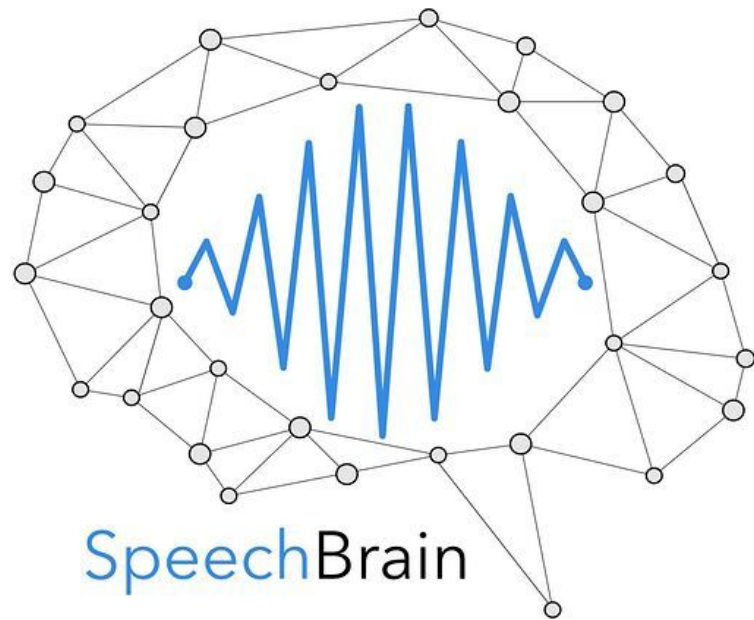
CHiME-4 (Perceptual vs Joint)

Model	Parallel?	Perceptual model	Joint?	SI-SDR speech quality	eSTOI intelligibility
Noisy	No	-	-	7.5	68.3
AECNN	No	Wide ResNet	No	1.6	72.6
AECNN	No	Wide ResNet	Yes	0.6	47.0
AECNN	Yes	-	-	11.7	78.9
AECNN	Yes	Wide ResNet	No	11.9	79.8
AECNN	Yes	Wide ResNet	Yes	11.7	79.5

Aside: SpeechBrain

State-of-the-art recipes for:

- End-to-end ASR
- Speaker embeddings
- Speaker diarization
- Speech separation
- Spoken language understanding
- etc.

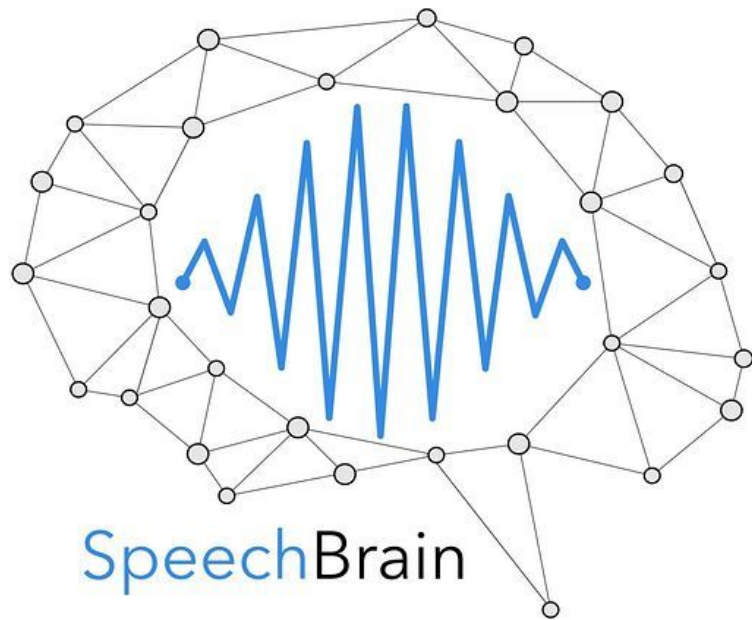


Aside: SpeechBrain

Easy to combine recipes, e.g.

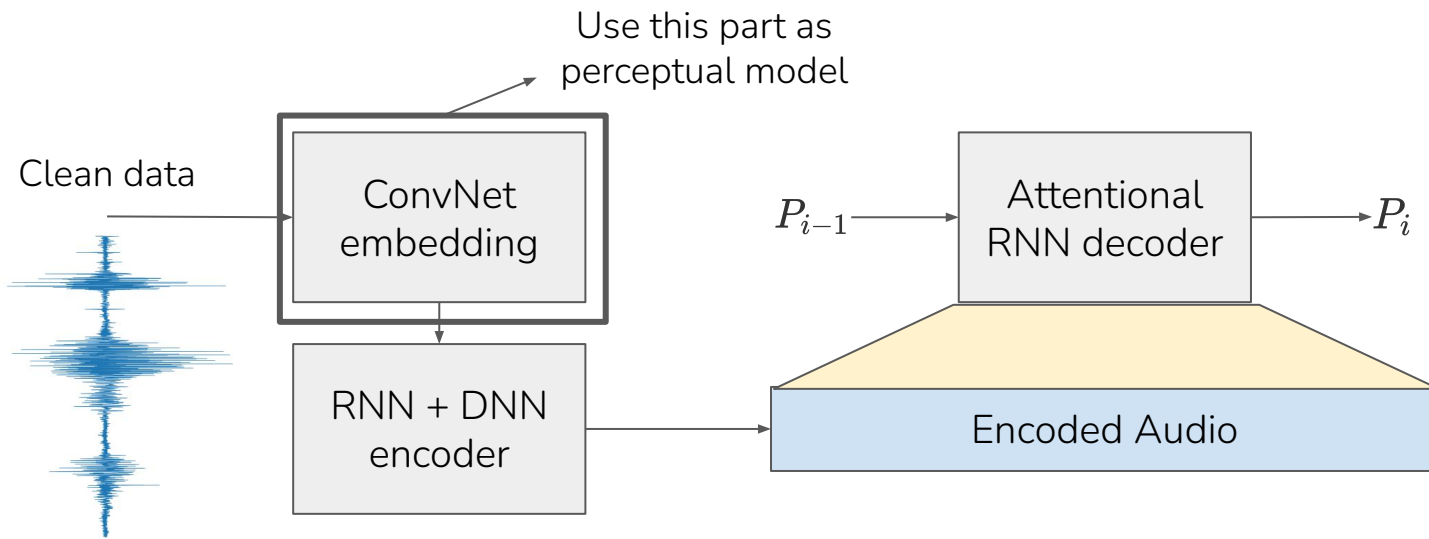
- Attentional model, trained on LibriSpeech, gets 3.0 WER
- Recipe for enhancement on Voicebank + DEMAND

We use these recipes so we can compare against joint training



Sequential Perceptual Model

Use part of seq2seq model for perceptual model



Voicebank (Perceptual vs Joint)

Enhancement model	Perceptual model	Joint?	PESQ speech quality	eSTOI intelligibility	Dev WER	Test WER
Clean	-	-	4.50	100.	1.44	2.29
Noisy	-	-	1.97	78.7	4.33	3.60
Wide ResNet	-	-	2.94	86.5	2.95	3.24
Wide ResNet	ConvNet	-	3.05	86.8	2.58	3.06
Wide ResNet	ConvNet	YES	3.08	86.6	2.91	3.08

Voicebank (Ours vs SotA)

System	PESQ speech quality	CSIG signal distortion	CBAK background distortion	COVL overall distortion
Noisy	1.97	3.35	2.44	2.63
MetricGAN (Fu et al. 2019)	2.86	3.99	3.18	3.42
PHASEN (Yin et al. 2020)	2.99	4.21	3.55	3.62
DEMUCS (Defossez et al. 2020)	3.07	4.31	3.40	3.63
Wide ResNet + Perceptual loss	3.05	4.36	3.51	3.73

Outline

Approaches to knowledge transfer

New ways to use knowledge transfer:

1. For removing noise from speech recordings
- 2. For teaching kids how to read**

Future work and conclusions

Reading Verification

Learning to read is an important task!
Can we design a model that helps?

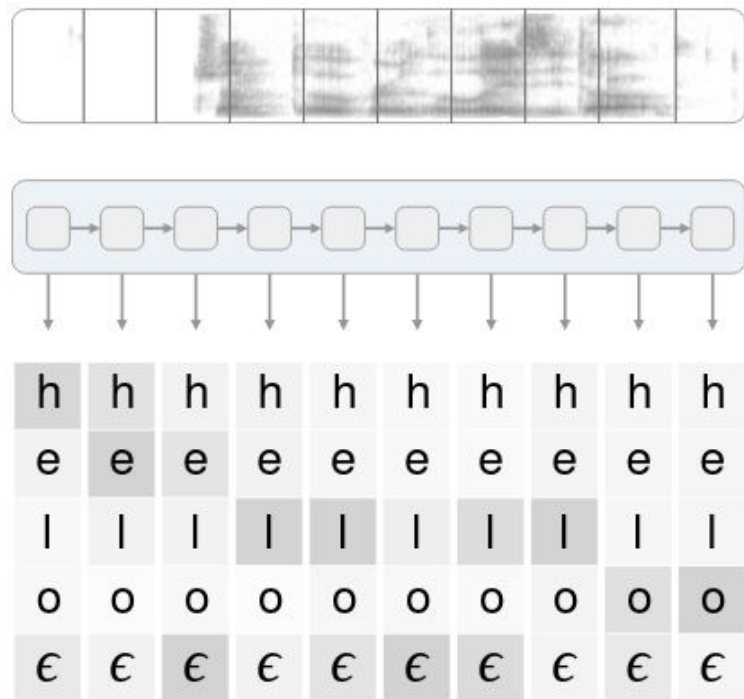
Considerations:

1. Evaluation (correct WPM)
2. Prompting (real-time)
3. Problem words

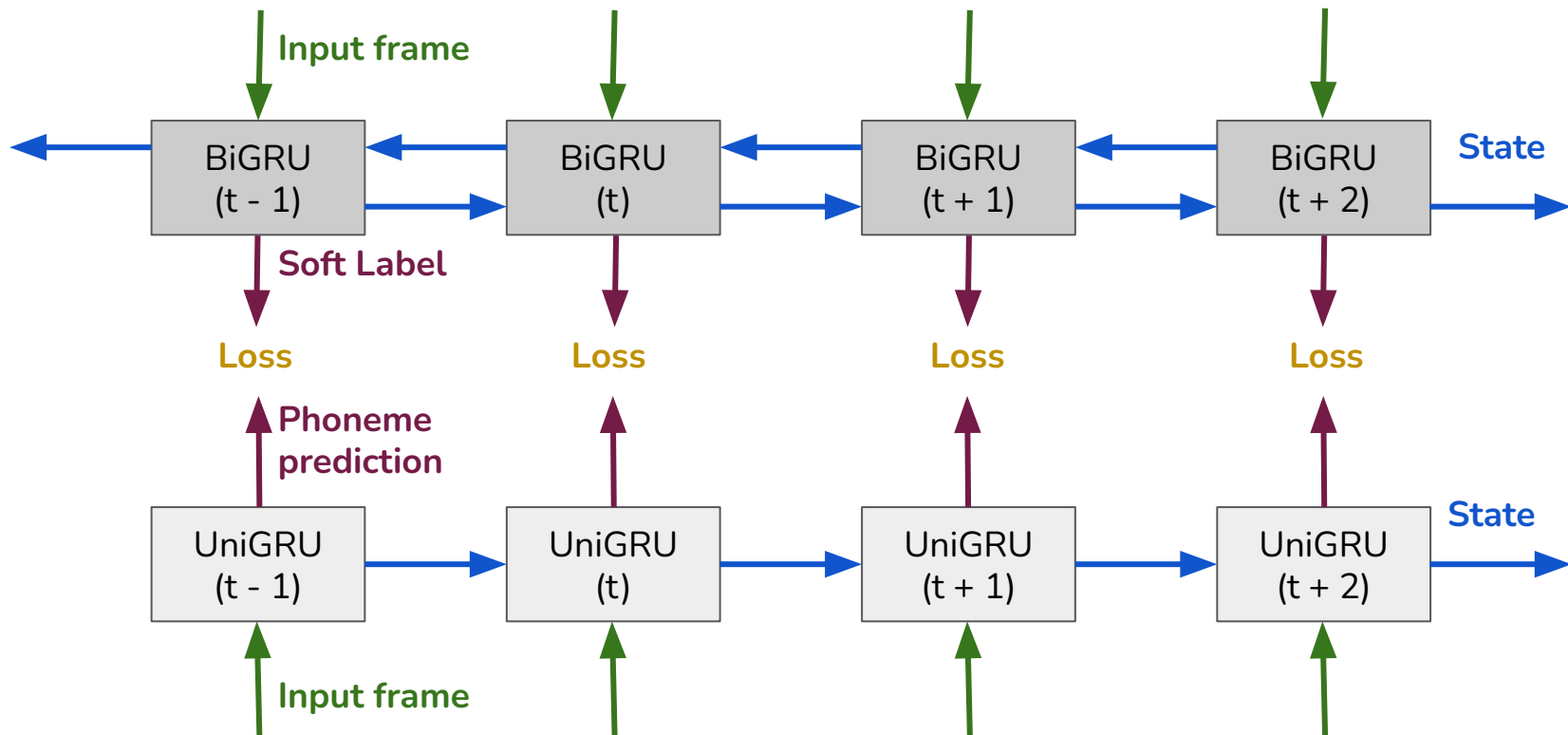


CTC Loss

We focus on real-time aspect.
For this, we chose CTC loss:
CTC loss adds blank label ϵ
and combines identical terms,
summing over equivalent paths.
Can be decoded in real time!



Teacher-Student Loss



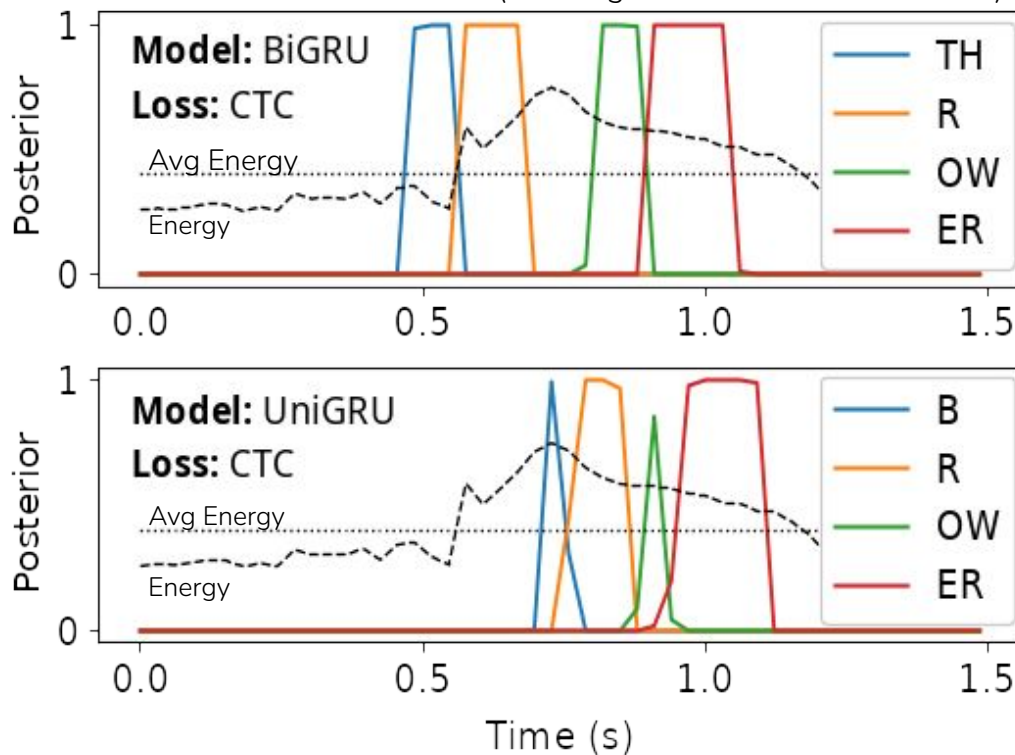
Alignment Problem

With CTC, outputs are not aligned with the evidence

BiGRU outputs can occur before evidence

UniGRU outputs occur after evidence

(Plantinga and Fosler-Lussier 2019)



Alignment Loss

Solution #1: Alignment loss

$$L_{align}(t) = -\log \left(\sum_{n=1}^N I(E_t) \hat{p}_n \right)$$

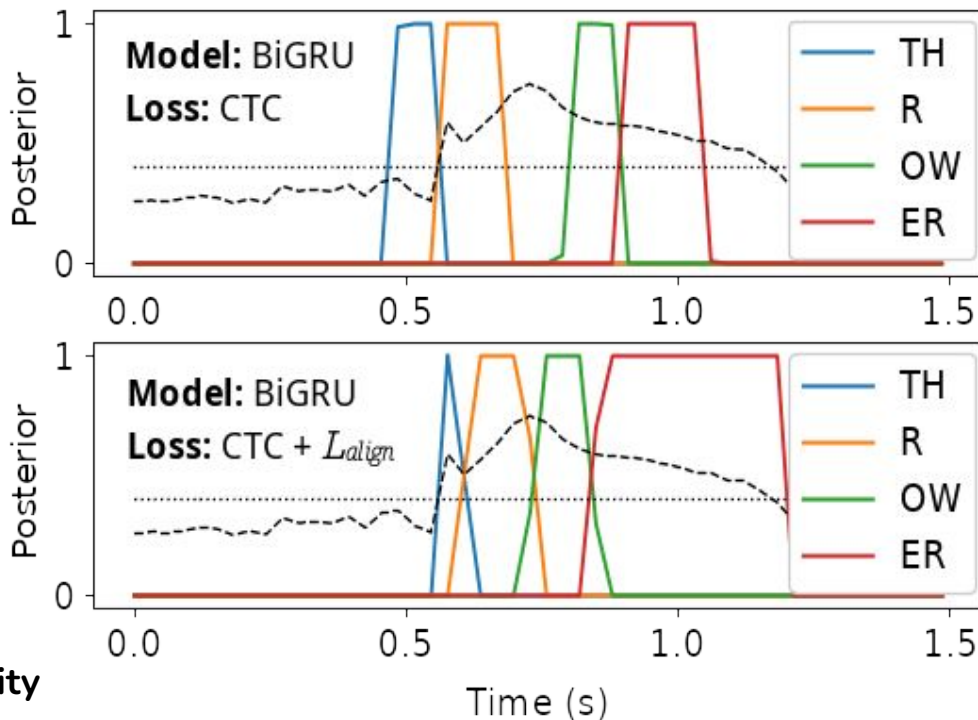
loss at **frame t**

sum over the phoneme
vocabulary (size: N)

indicator function selects
blank or non-blank symbols
depending on relative energy

n-th phoneme's
predicted **probability**

(Plantinga and Fosler-Lussier 2019)

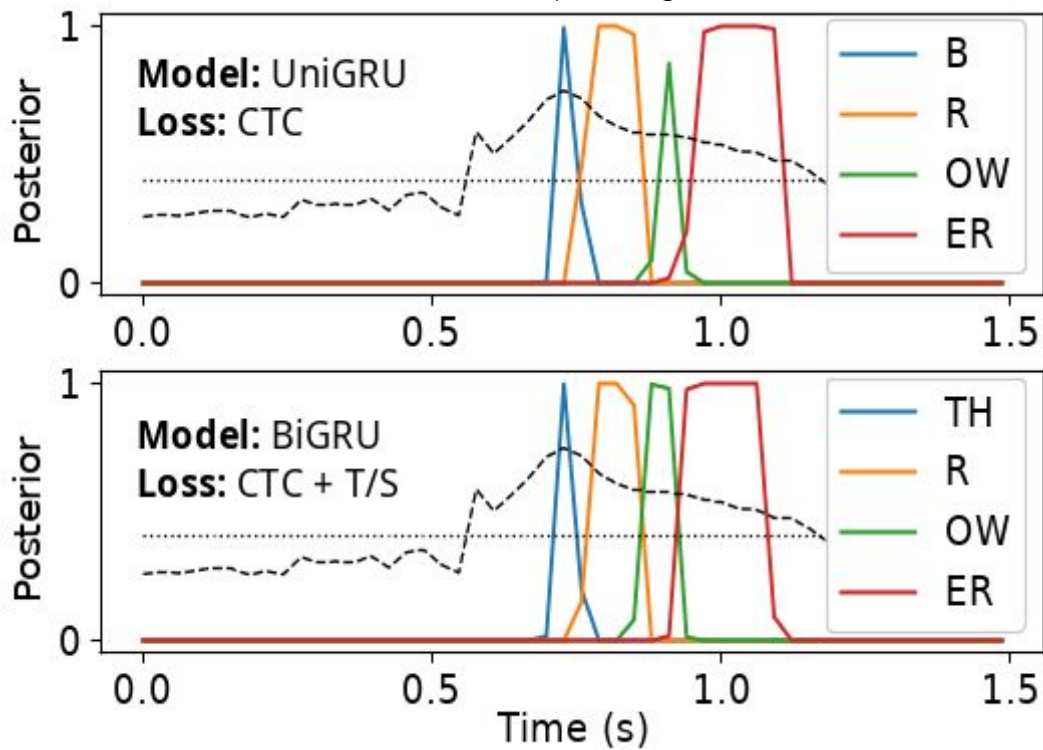


Reversed T/S learning

(Plantinga and Fosler-Lussier 2019)

Solution #2:
UniGRU as teacher

BiGRU learns to put
outputs in the most
helpful place



OGI kids' speech dataset

Recordings of kids in grades K-10

~10 mins from ~100 kids in each grade

Fluency labels (binary) are based on recording quality (missing word, etc.)

Simple metric for error detection:
more than one prediction error



Recognition Results

Model	Student Loss	Teacher Loss	PER error rate
BiGRU	CTC	-	12.6
UniGRU	CTC	-	19.5
UniGRU	CTC + T/S	CTC	21.3
UniGRU	CTC + T/S	CTC + T/S	18.4
UniGRU	CTC + T/S	CTC + T/S + Align	19.0

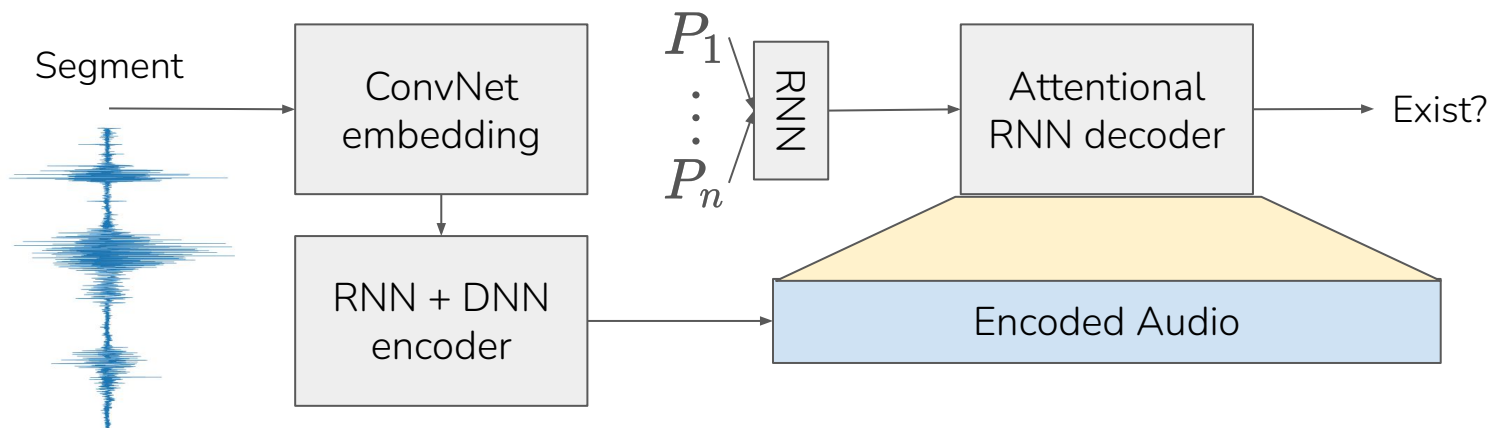
Verification Results

Model	Student Loss	Teacher Loss	F1	Delay
BiGRU	CTC + Align	-	60.4	0 ms
UniGRU	CTC	-	54.7	159 ms
UniGRU	CTC + T/S	CTC + Align	49.1	24 ms
UniGRU	CTC + T/S	CTC + T/S	56.7	237 ms
UniGRU	CTC + T/S	CTC + T/S + Align	56.2	162 ms

Word Detection

Follow-up: can we detect if a word has been said?

Cut up audio and label whether each word is inside segment



Word Detection Results

Encoder Model	Pretrained?	Frozen?	Detection Model	F1
None	-	-	Attentional RNN	90.9
CRDNN	Yes	Yes	Attentional RNN	92.4
CRDNN	No	No	Attentional RNN	95.1
CRDNN	Yes	No	Attentional RNN	95.5

Outline

Approaches to knowledge transfer

New ways to use knowledge transfer:

1. For removing noise from speech recordings
2. For teaching kids how to read

Future work and conclusions

Future work


- GAN loss + perceptual loss for non-parallel enhancement
- Use Reading RACES data
- Add disfluency detection: stutters, prompts, etc.
- Add enhancement due to noisy classroom environment

Conclusions

- Knowledge transfer is useful for speech tasks
- Perceptual loss can improve both enhancement and noise-robust ASR at the same time
- Alignment loss and reversed T/S trained teacher model can trade-off accuracy and latency
- SpeechBrain is a great toolkit for knowledge transfer for speech tasks!

slido

Audience Q&A Session

 Start presenting to display the audience questions on this slide.

Graduate Publications

Ravanelli, Mirco; Parcollet, Titouan; Rouhe, Aku; **Plantinga, Peter**; Rastorgueva, Elena; Lugosch, Loren; Dawalatabad, Nauman; Ju-Chieh, Chou; Heba, Abdel et al. “SpeechBrain.” GitHub Repository, 2021.

Plantinga, Peter; Bagchi, Deblin; Fosler-Lussier, Eric. “Phonetic feedback for speech enhancement with and without parallel speech data.” ICASSP 2020.

Plantinga, Peter; Fosler-Lussier, Eric. “Towards real-time mispronunciation detection in kids’ speech.” ASRU 2019.

Plantinga, Peter; Bagchi, Deblin; Fosler-Lussier, Eric. “An Exploration of Mimic Architectures for Residual Network Based Spectral Mapping.” SLT 2018.

Bagchi, Deblin; **Plantinga, Peter**; Stiff, Adam; Fosler-Lussier, Eric. “Spectral Feature Mapping with Mimic Loss for Robust Speech Recognition.” ICASSP 2018.

References

Buciluă, Cristian; Caruana, Rich; Alexandru, Niculescu-Mizil. "Model compression." SIGKDD, 2006.

Ganesh, Prakar. "Knowledge distillation : simplified." Towards Data Science, 2019.

Gatys, Leon; Ecker, Alexander; Bethge, Matthias. "Image style transfer using convolutional neural networks." CVPR, 2016.

Goodfellow, Ian; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron; Bengio, Yoshua. "Generative adversarial networks." NIPS, 2014.

Hannun, Awni; "Sequence modelling with CTC." Distill, 2017.

Healy, Eric; Delfarah, Masood; Vasko, Jordan; Carter, Brittney; Wang, Deliang. "An algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker." The Journal of the Acoustical Society of America, 2017.

References

Hinton, Geoffrey; Vinyals, Oriol; Dean, Jeff. "Distilling the knowledge in a neural network." Stat, 2015.

Narayanan, Arun; Wang, Deliang. "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training." Transactions, 2015.

Thomson, James. "Neural style transfer with swift for TensorFlow." Medium, 2019.

Vapnik, Vladimir; Vashist, Akshay. "A new learning paradigm: learning using privileged information." Neural Networks, 2009.

Wang, Peidong; Tan, Ke; Wang, Deliang. "Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling"

Zhang, Richard; Isola, Phillip; Efros, Alexei; Shechtman, Eli; Wang, Oliver. "The unreasonable effectiveness of deep features as a perceptual metric." CVPR 2018.

CHiME-2 Results

Enhancement ↓	Perceptual→	None	6-layer DNN	Wide * ResNet	WRBN
Noisy (no enhancement)		17.4	-	-	-
2-layer DNN (Bagchi et al. 2018)		16.0	14.4	14.2	14.0
Wide ResNet (Plantinga et al. 2018)		10.8	10.5	8.7	9.3

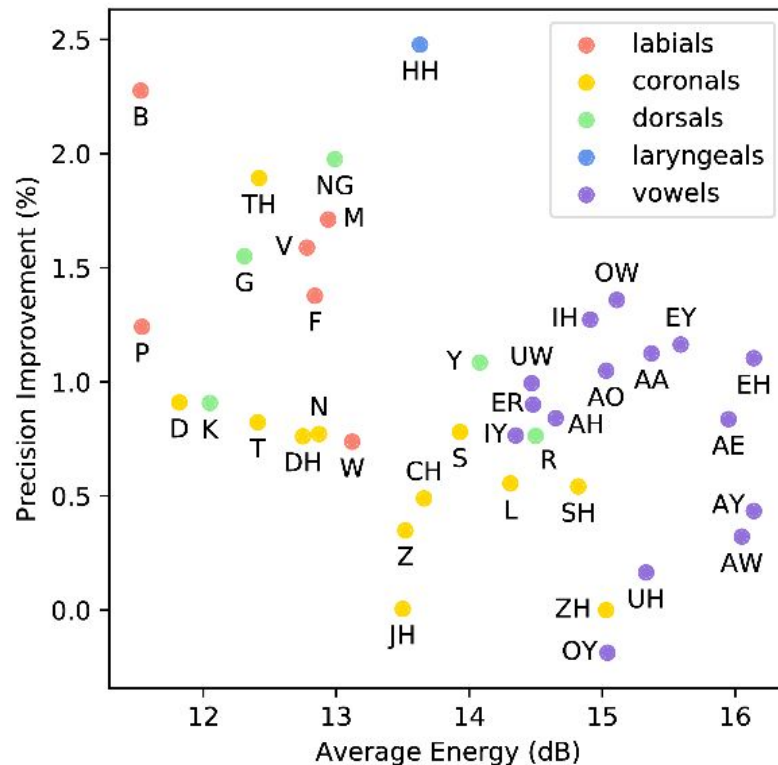
* Wide ResNet perceptual model unpublished

CHiME-2 Layer-wise

Layer of Perceptual Model Used	WER
Noisy	16.0
Layer 1	15.0
Layer 3	14.7
Layer 6	14.4
Layers 4+5+6	14.3

CHiME-2 Analysis

- Our model improves more on lower energy phonemes
- Less improvement on vowels, since they are long and have higher energy
- Correlation coefficient of the consonants is around -0.5



Voicebank w/ Transformer

Enh. model	Mask training	Joint?	PESQ	eSTOI	dWER	tWER
Noisy	-	-	1.97	78.7	4.33	3.60
Transformer	MSE loss	Yes	2.45	83.3	3.40	3.12
Transformer	MSE + Mimic loss	Yes	2.58	83.5	3.50	3.32
Transformer	MSE loss	No	2.72	84.8	3.48	3.12
Transformer	MSE + Mimic loss	No	2.92	85.3	3.20	2.96
Wide Resnet	MSE + Mimic loss	No	3.05	86.8	2.58	3.06

Transformer system is mixed with noisy, 0.7 mask + 0.3 noisy

Transformer Example

Dataset: Voicebank + DEMAND

Noisy: 

Enhanced: 

Clean: 



Reading Verification

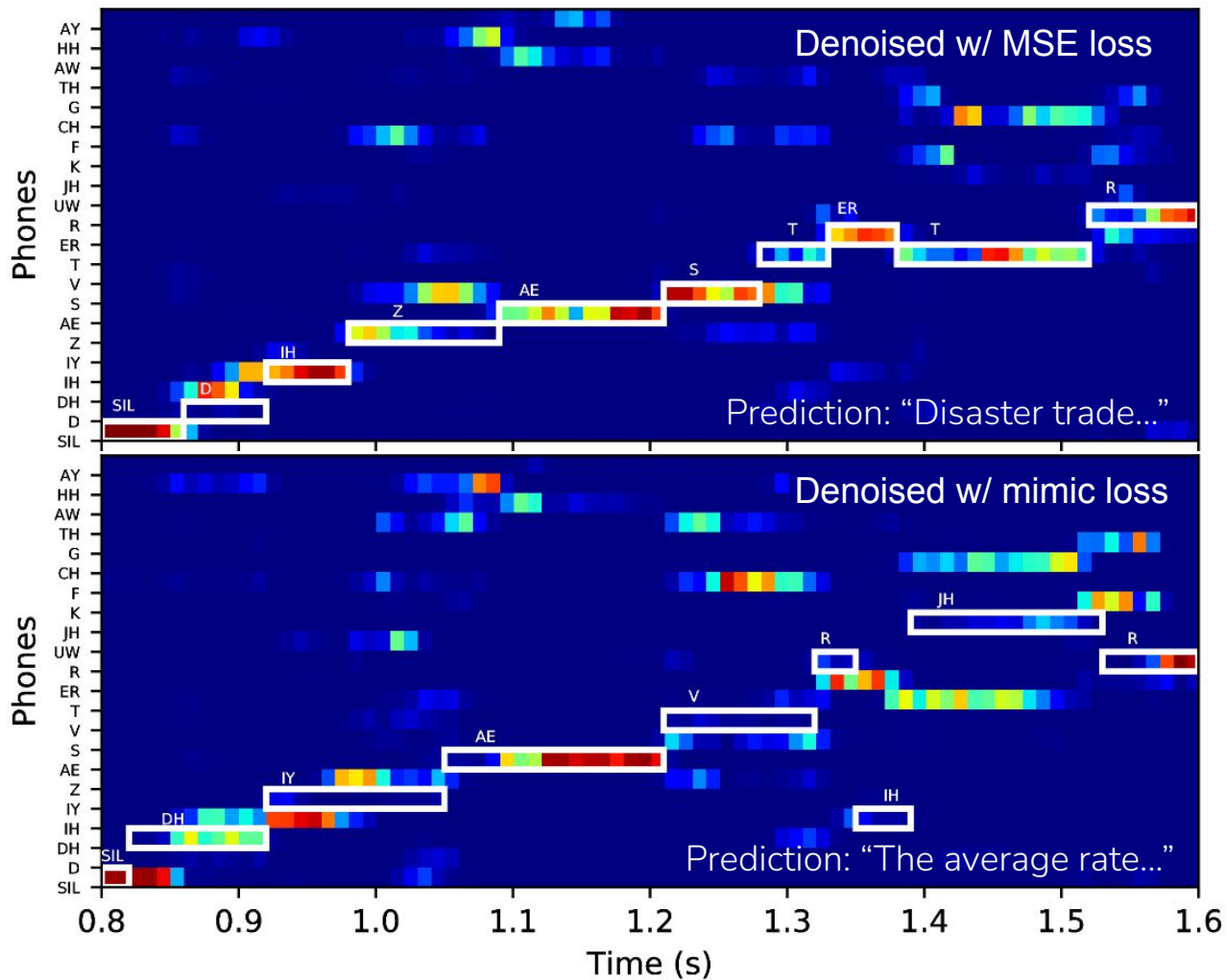
Model	Student Loss	Teacher Loss	PER
BiGRU	CTC	-	12.6
BiGRU	CTC + T/S	CTC	12.5
BiGRU	CTC + T/S + Align	CTC	13.1
BiGRU	CTC + Align	-	13.3

Model	Student Loss	Teacher Loss	F1	Delay
BiGRU	CTC	-	61.2	153 ms
BiGRU	CTC + Align	-	60.4	0 ms

Posterior Graph

Outlined are predicted phones

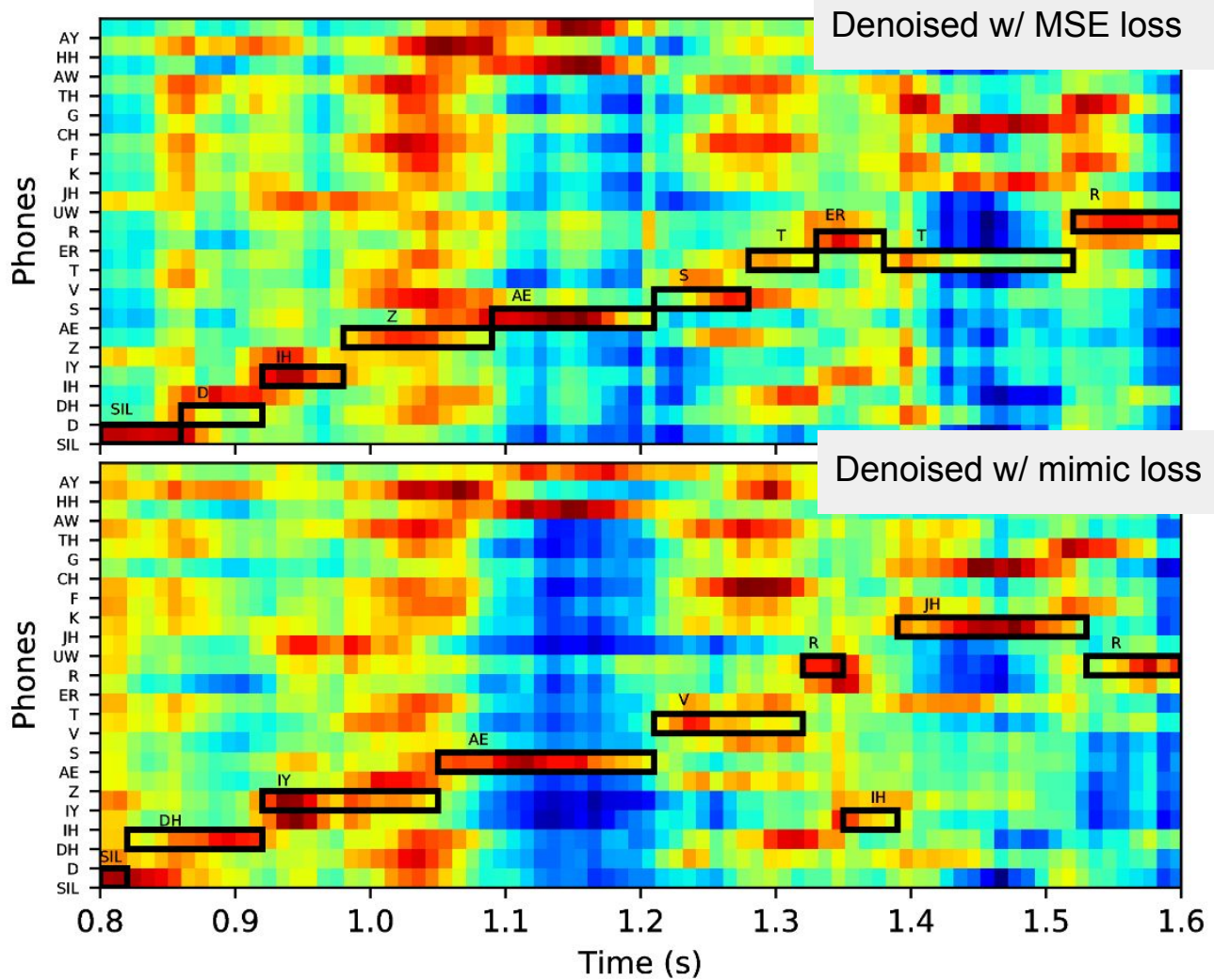
Proposed model makes the correct predictions



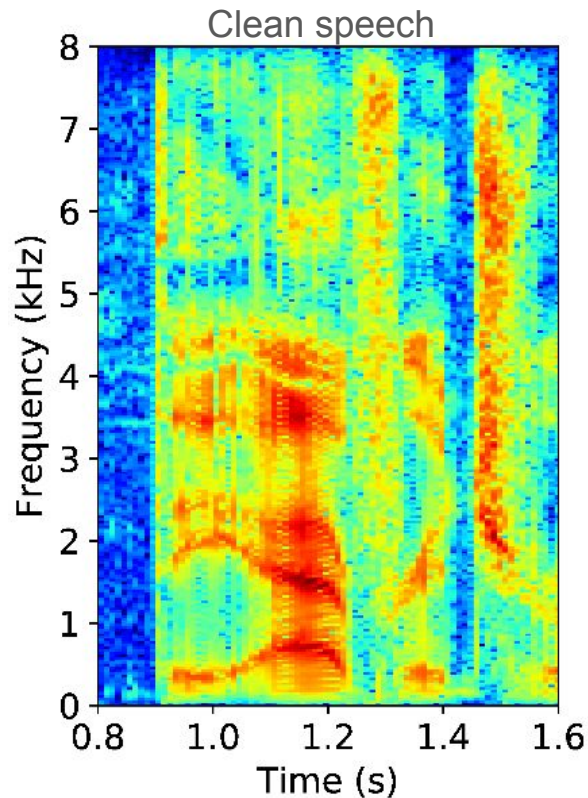
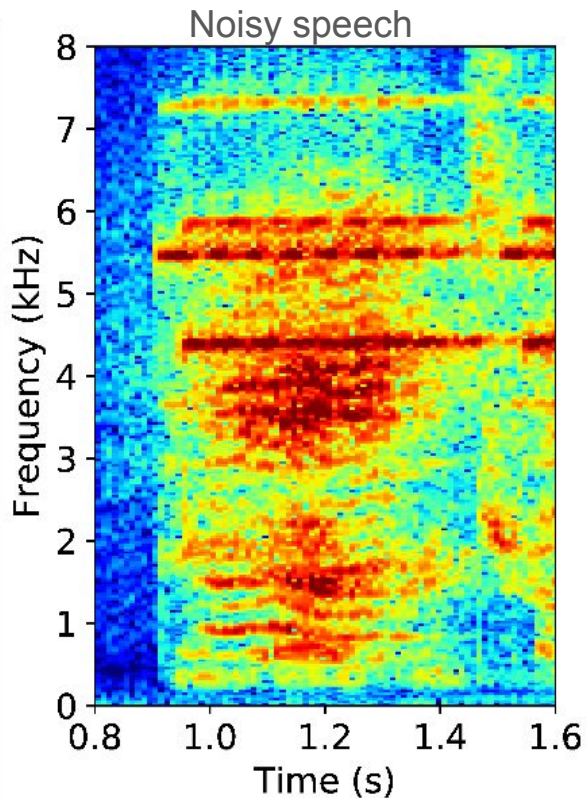
Likelihood Graph

Outlined are predicted phones

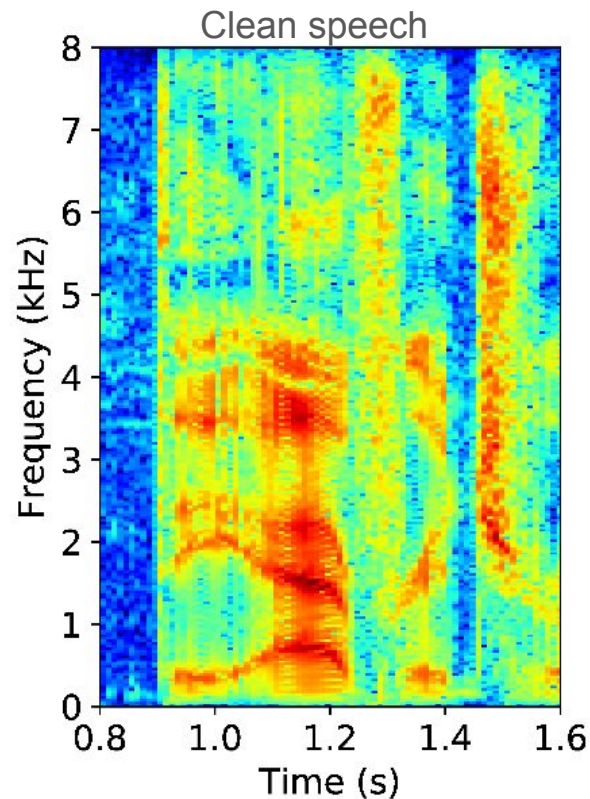
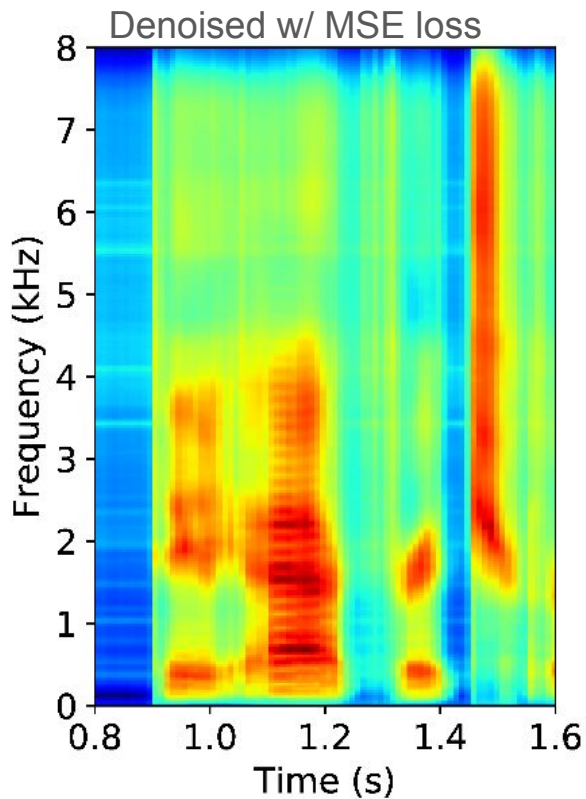
Proposed model makes the correct predictions



Denoising Example



Denoising Example



Denoising Example

