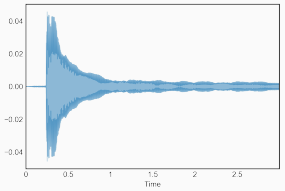
End to End ASR and Self Supervised Learning

Anirudh Gupta

Automatic Speech Recognition

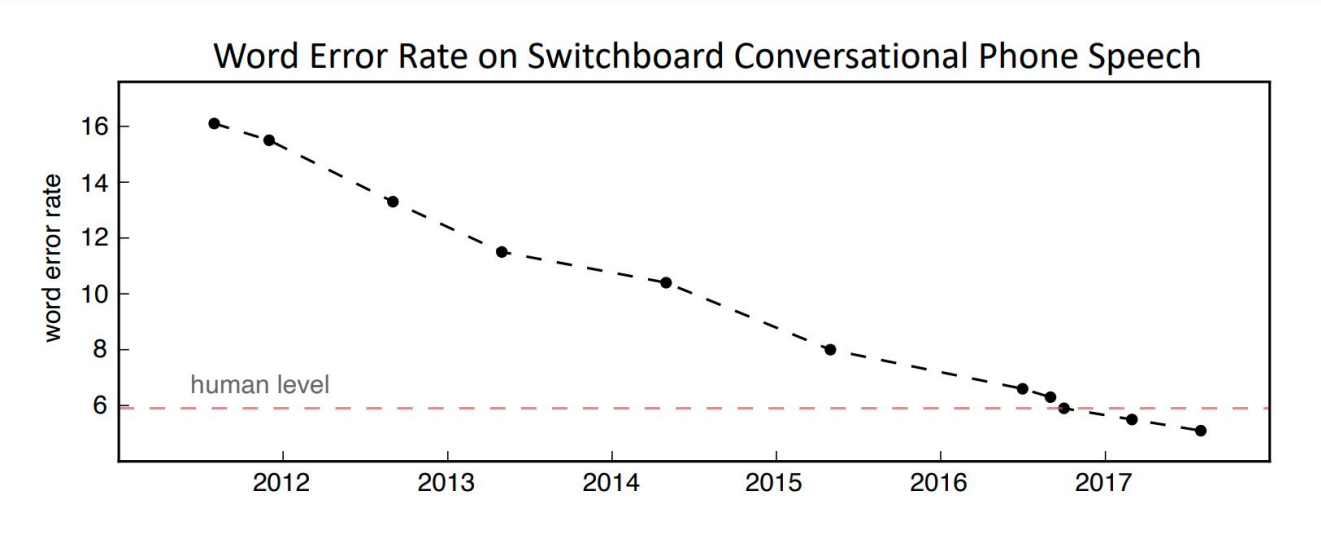
Goal: Input Speech --------> Output Transcription



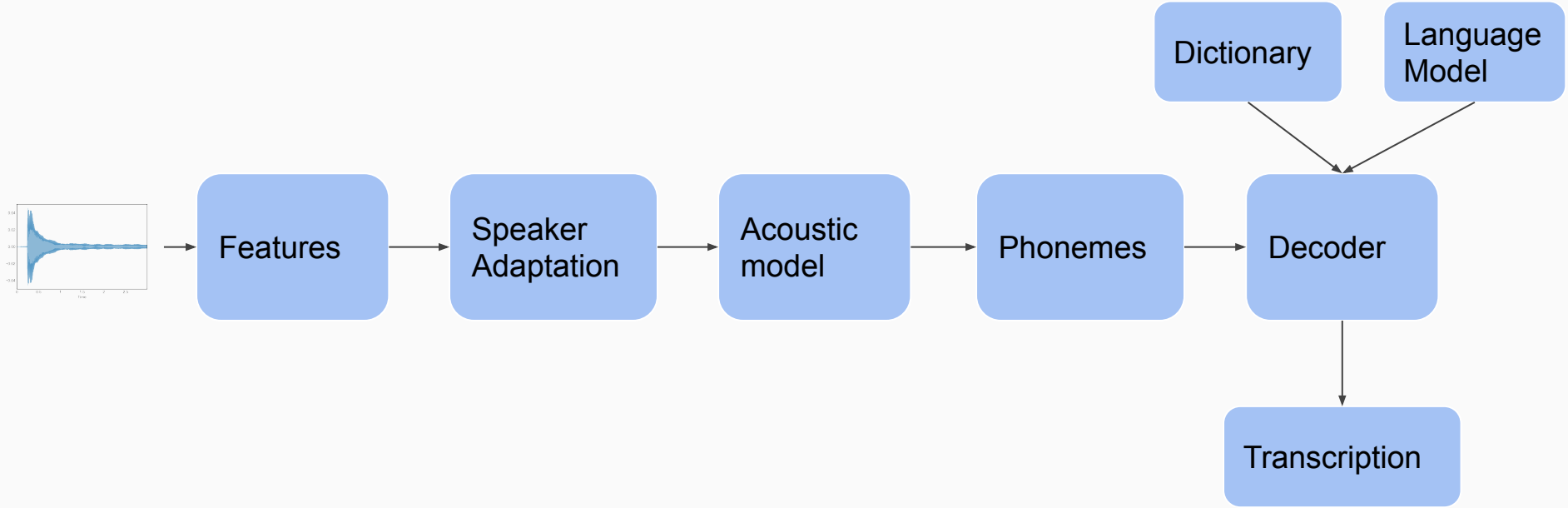
Automatic Speech Recognizer

The quick brown fox jumps over the lazy dog

Modern Speech Recognition



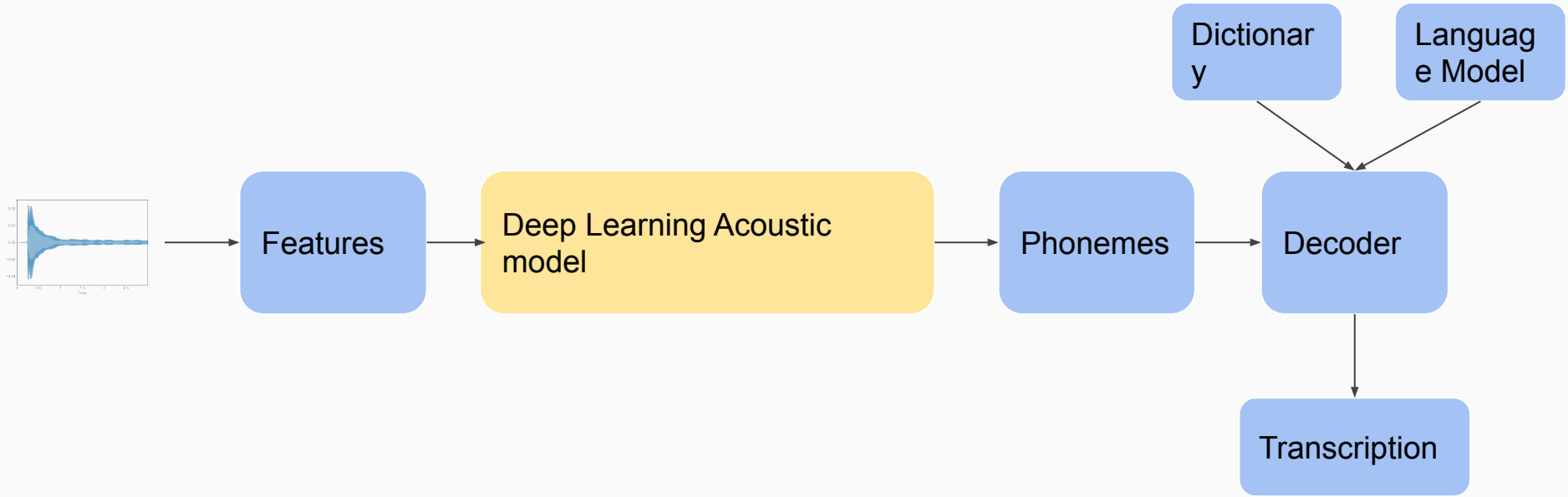
Traditional ASR Systems



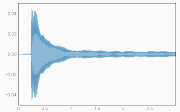
Traditional ASR Systems

- Traditional pipeline is highly tweak-able, but also hard to get working well.
- Historically, each part of the system has its own set of challenges.
  Eg - choosing feature representation
- Cascading errors due to different individual components.

DL based acoustic models



End to end networks



Features

Deep Learning Acoustic model

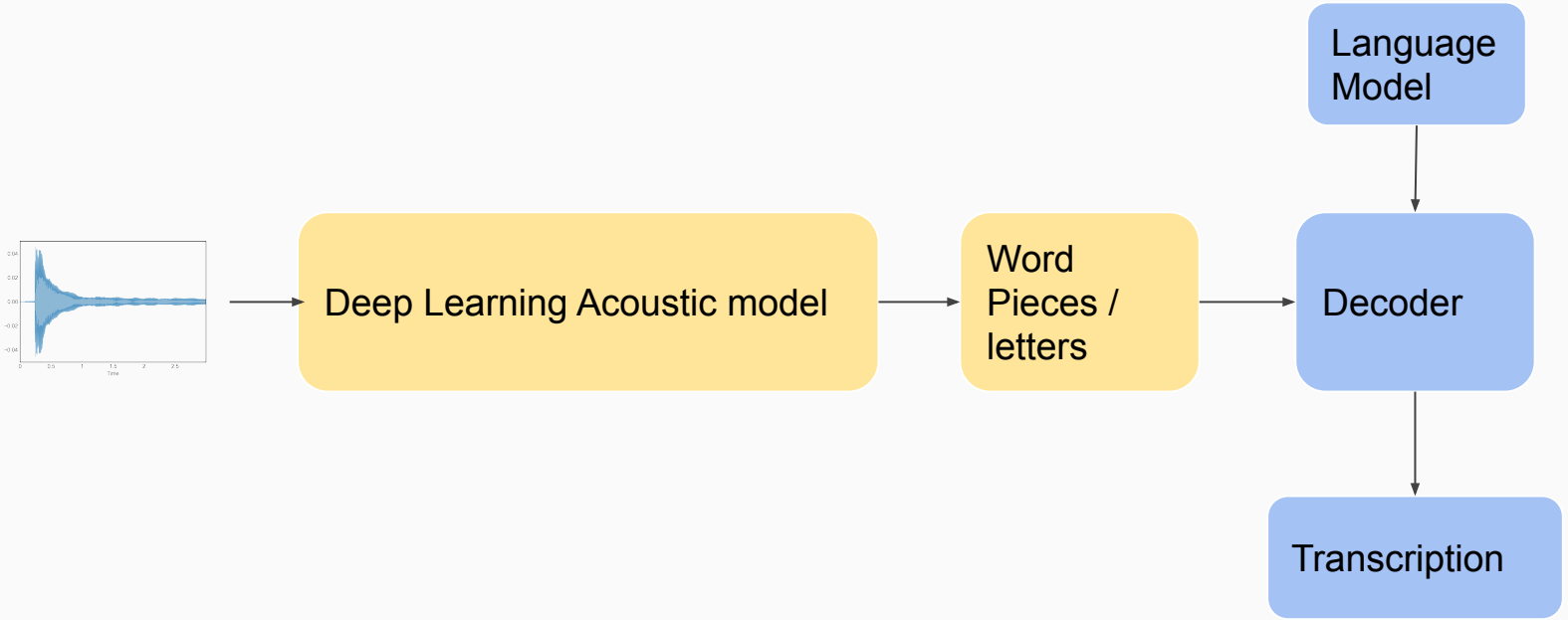
Word Pieces / letters

Language Model

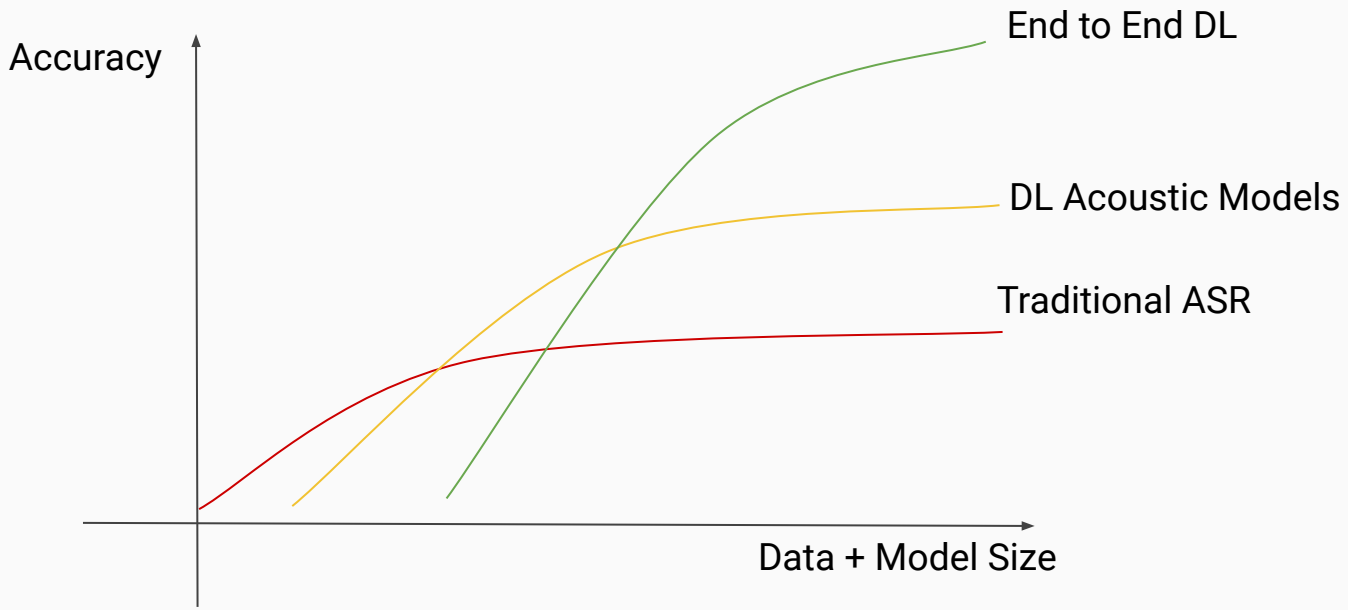
Decoder

Transcription

End to end networks

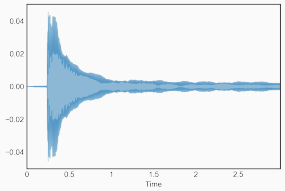


Moving towards end to end networks



Supervised E2E Networks

Towards end to end networks



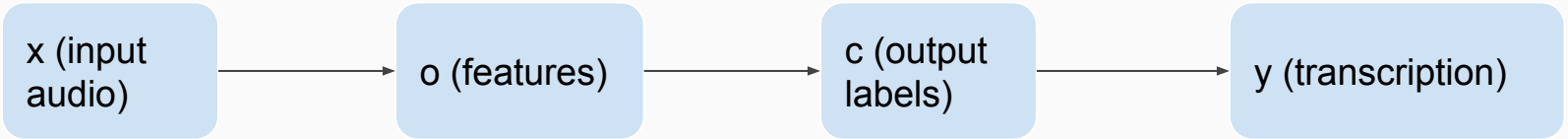
Deep learning network

The quick brown fox jumps over the lazy dog

- Main issue: length(x) != length(y)
  - Difficult to understand how y maps to frames of audio. (eg. same word over different time scales)
- Initial fix: Train a network that would predict the sound at every frame. (alignment)

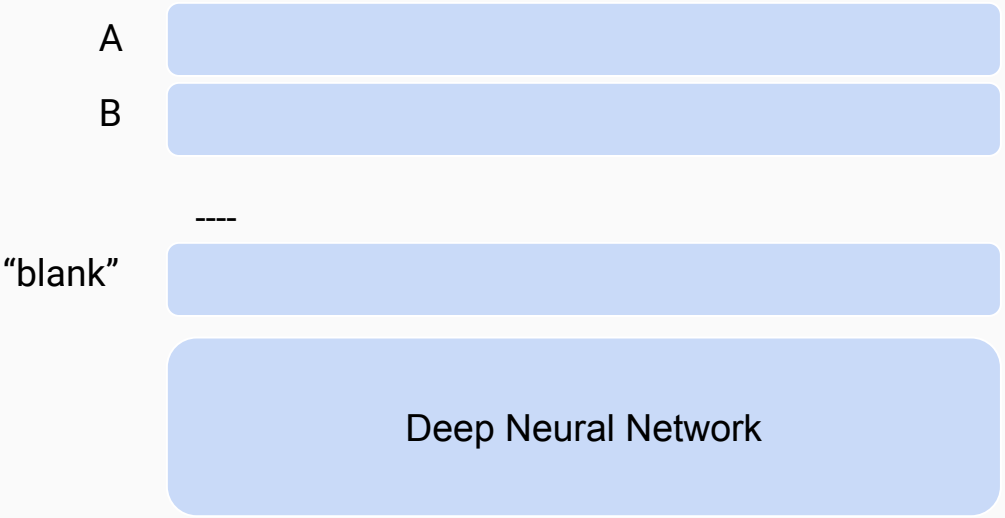
Connectionist Temporal Classification (CTC)

- A network outputs c, a distribution over symbols. (length(c) = length(x))
  $c \in$ {A, B, C, …. , Z, blank}
- Define a mapping α : c ---> y
- Maximize likelihood of y* under this model.



[Graves et. al, 2006]

Connectionist Temporal Classification (CTC)



Connectionist Temporal Classification (CTC)

Assuming independence a distribution can be defined over the whole character sequences.

$$P(c|x) = \Pi_{i=1}^{N} P(c_i|x)$$

Eg. P(c = HHH_E_LL_LO_ _ _ | x) = P(c_1 = H|x) P(c_2 = H|x) .... P(c_{15} = blank|x)

Connectionist Temporal Classification (CTC)

Define a mapping α : c ---> y

Given a specific character sequence c, squeeze out duplicates and blanks to yield transcription.

y = α(c) = α(HHH_E_ _LL_LO_ _ _) = “HELLO”

$$P(y|x) = \sum_{c:\alpha(c)=y} P(c|x)$$

Connectionist Temporal Classification (CTC)

$$\theta^* = argmax_\theta \sum_i log P(y^{*(i)}|x^{(i)})$$

$$= argmax_\theta \sum_i log \sum_{c:\alpha(c)=y^{*(i)}} P(c|x^{(i)})$$

[Graves et. al 2006] provides an efficient dynamic programming algorithm to compute the summation and its gradient.

DeepSpeech

- Trained on almost ~10,000 hours of audio data.
- Does not need hand designed components to handle noise, reverberation or speaker variation.
- Does not need a phoneme dictionary.
- Achieved 16% WER on Switchboard Hub5'00

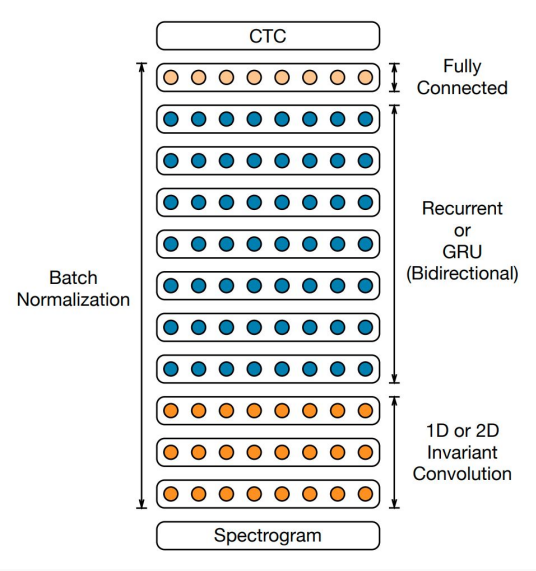


Fig: Network Architecture

Source

Types of Speech Data

- Find data that matches our goals.

Styles of Speech: Read, conversational, Spontaneous

Issues: Disfluency, noise, mic quality, reverb, echo

Applications: Dictation, meeting transcriptions, call centres etc.

Speech Data

- Transcribing speech data isn’t cheap.
- Typical speech benchmarks offer 100s to few 1000s hours of data.
  - LibriSpeech
  - LDC corpora
  - LDCIL

Why Self Supervised Learning?

- AI has made tremendous progress in developing systems that can learn from massive amounts of carefully labelled data.
- However, this is a bottleneck for building more generalist models that can perform multiple tasks without massive amounts of data.
- It is practically impossible to label everything in the world!

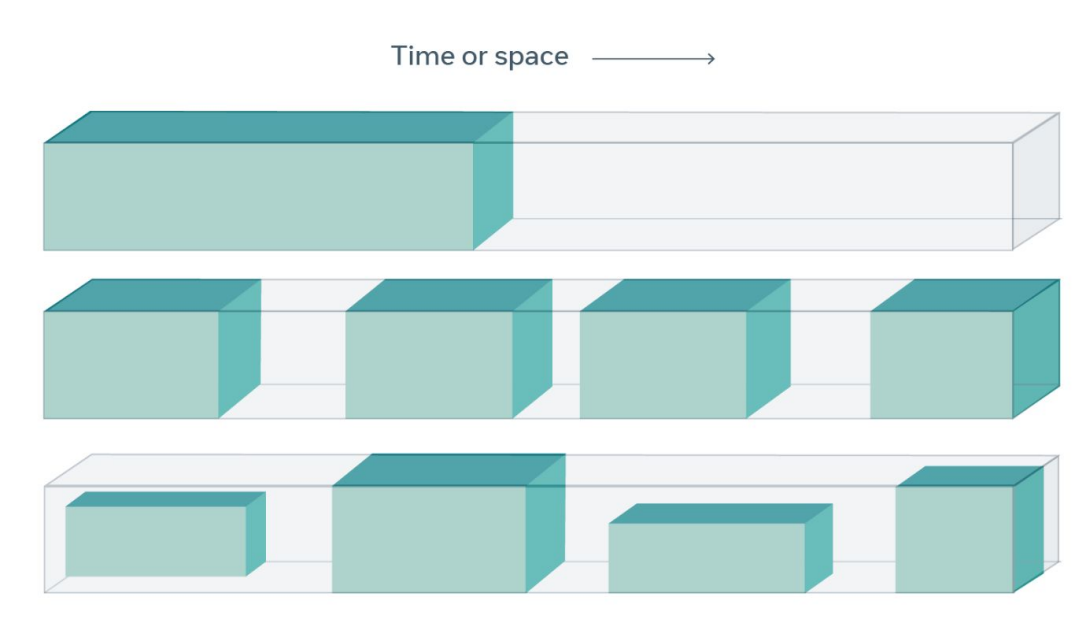
“Alternate” Supervision

- Getting labelled data is extremely difficult and expensive. (Switchboard)
- Labelling does not scale very well.

Semi automatic process: Obtain data using forced alignment and other techniques.

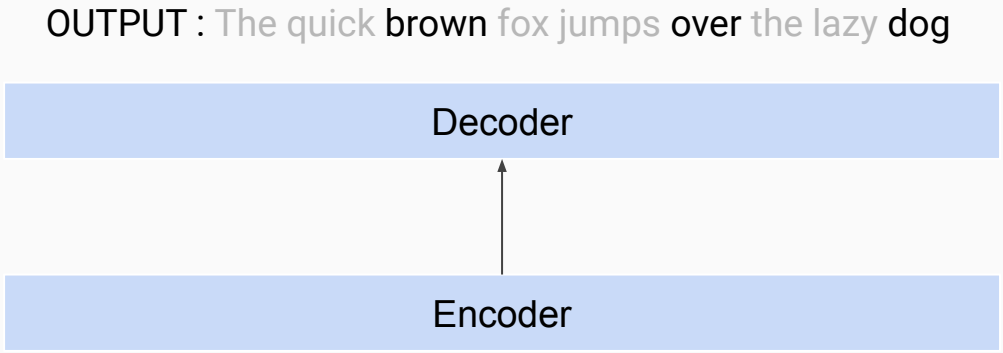
Using the data itself: **“self”- supervised**

Self Supervision as predictive learning



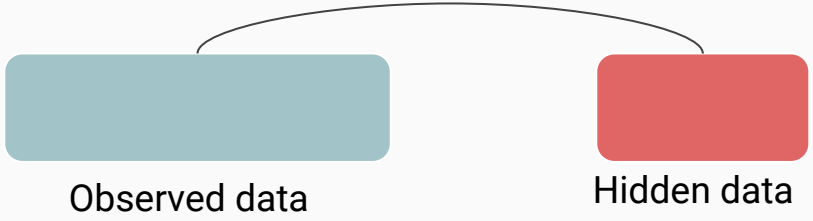
Source

Self Supervised Learning in NLP



Contrastive Learning

- Technique to learn general features of a dataset without labels by teaching the model which data points are similar or different.
- Given a set of input labels $\{x_i\}$, each corresponding to a label $y_i \in \{1, ..., L\}$ among L classes.
- We would like to learn a function $f_\theta(.) : X \text{ ------> } R^d$ that encodes x_i into an embedding vector space such that examples from same classes have similar embeddings and sampled from different classes have very different ones.



Wav2vec 2.0

- Learn audio representations in a pretraining task solely from audio data as a contrastive task.
- Fine tune on ***some*** amount of audio data for speech recognition task using CTC loss function.

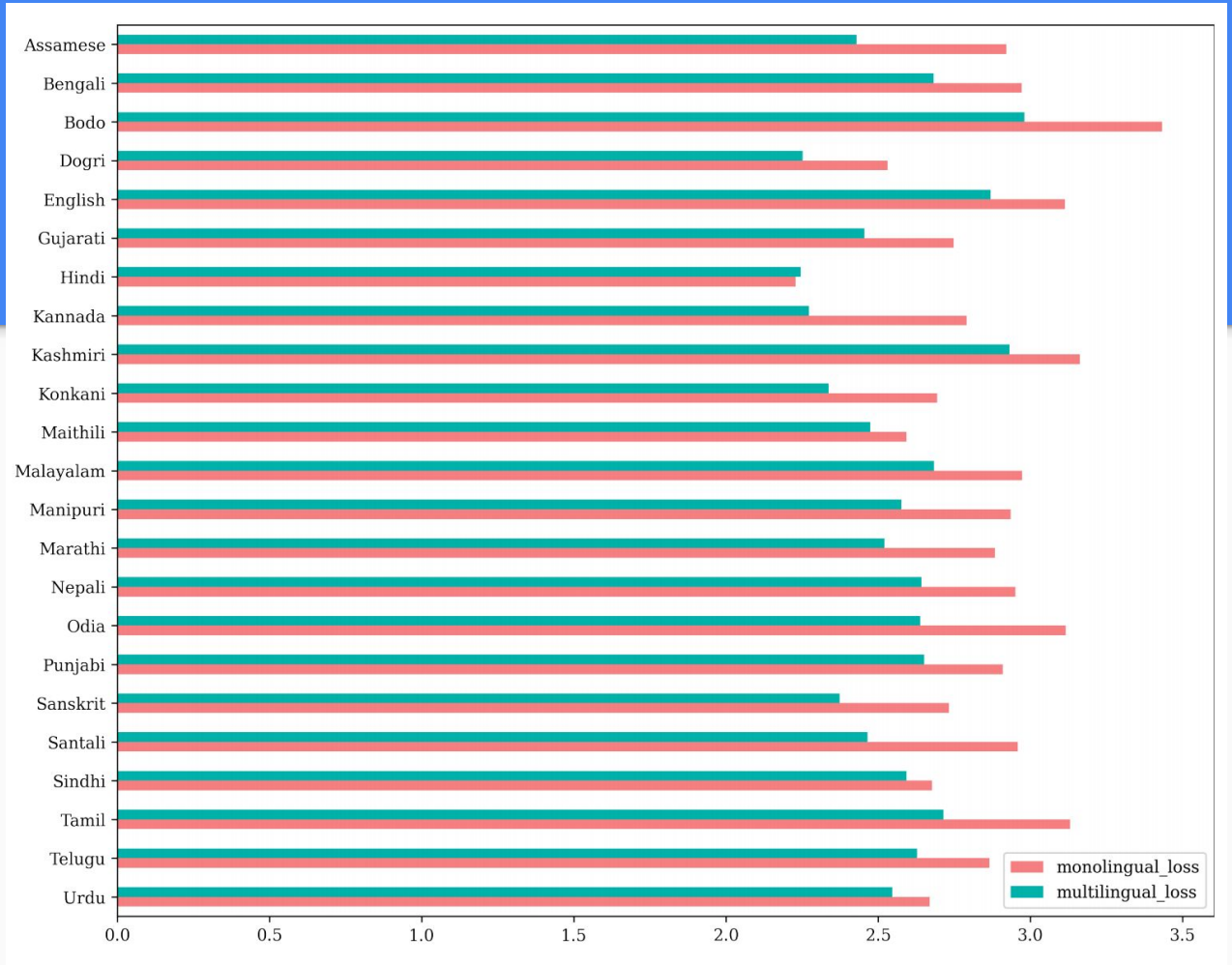
Wav2vec 2.0

- Compensate for amount of fine-tuning data with the amount of pre-training data.

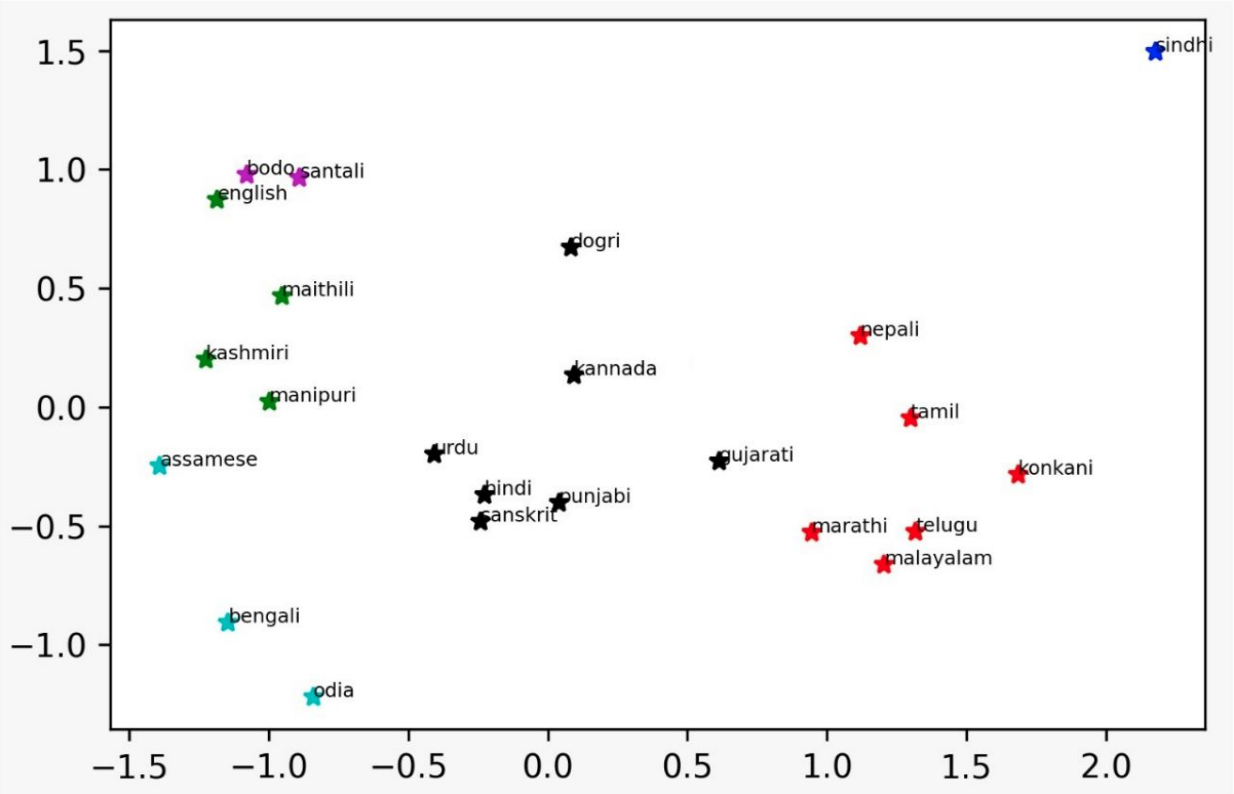
Model	Unlabeled data	LM	dev		test	
			clean	other	clean	other
10 min labeled						
BASE	LS-960	None	46.1	51.5	46.9	50.9
		4-gram	8.9	15.7	9.1	15.6
		Transf.	6.6	13.2	6.9	12.9
LARGE	LS-960	None	43.0	46.3	43.5	45.3
		4-gram	8.6	12.9	8.9	13.1
		Transf.	6.6	10.6	6.8	10.8
LARGE	LV-60k	None	38.3	41.0	40.2	38.7
		4-gram	6.3	9.8	6.6	10.3
		Transf.	4.6	7.9	4.8	8.2
1h labeled						
BASE	LS-960	None	24.1	29.6	24.5	29.7
		4-gram	5.0	10.8	5.5	11.3
		Transf.	3.8	9.0	4.0	9.3
LARGE	LS-960	None	21.6	25.3	22.1	25.3
		4-gram	4.8	8.5	5.1	9.4
		Transf.	3.8	7.1	3.9	7.6
LARGE	LV-60k	None	17.3	20.6	17.2	20.3
		4-gram	3.6	6.5	3.8	7.1
		Transf.	2.9	5.4	2.9	5.8
10h labeled						
BASE	LS-960	None	10.9	17.4	11.1	17.6
		4-gram	3.8	9.1	4.3	9.5
		Transf.	2.9	7.4	3.2	7.8
LARGE	LS-960	None	8.1	12.0	8.0	12.1
		4-gram	3.4	6.9	3.8	7.3
		Transf.	2.9	5.7	3.2	6.1
LARGE	LV-60k	None	6.3	9.8	6.3	10.0
		4-gram	2.6	5.5	3.0	5.8
		Transf.	2.4	4.8	2.6	4.9
100h labeled						
BASE	LS-960	None	6.1	13.5	6.1	13.3
		4-gram	2.7	7.9	3.4	8.0
		Transf.	2.2	6.3	2.6	6.3
LARGE	LS-960	None	4.6	9.3	4.7	9.0
		4-gram	2.3	5.7	2.8	6.0
		Transf.	2.1	4.8	2.3	5.0
LARGE	LV-60k	None	3.3	6.5	3.1	6.3
		4-gram	1.8	4.5	2.3	4.6
		Transf.	1.9	4.0	2.0	4.0

Cross lingual representations

- Pre-training with multiple languages facilitates learning cross lingual speech representations.
- Data with high resource languages can be used to aid learning of low resource languages.
- Cross lingual pre-training significantly outperforms monolingual pre-training in speech recognition tasks.


Assamese
Bengali
Bodo
Dogri
English
Gujarati
Hindi
Kannada
Kashmiri
Konkani
Maithili
Malayalam
Manipuri
Marathi
Nepali
Odia
Punjabi
Sanskrit
Santali
Sindhi
Tamil
Telugu
Urdu
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
monolingual_loss
multilingual_loss

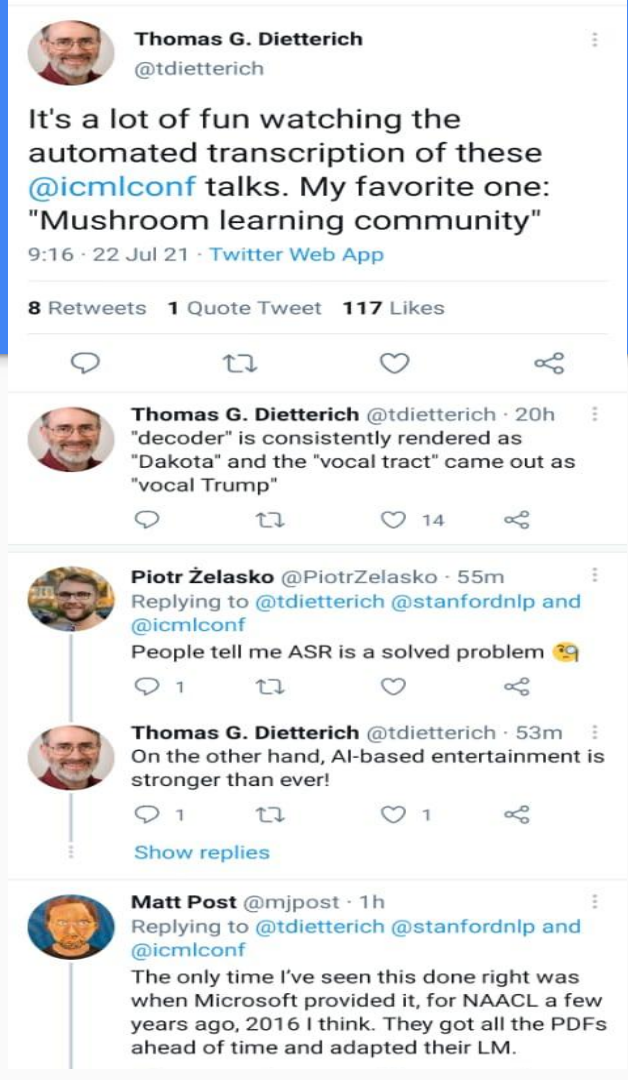
source


sindhi
bodo
santali
english
dogri
maithili
nepali
kashmiri
kannada
manipuri
tamil
urdu
assamese
gujarati
konkani
hindi
punjabi
sanskrit
marathi
telugu
malayalam
bengali
odia
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

source

Still not solved!

- Conversational speech with multiple speakers.
- Performance in noisy environments.
- Performance on esoteric accents and model bias.
- Goal: works on some speakers sometimes -----> works for all people all the time.
- Looking into semantic errors as well while evaluating WER.
- Human-error rate on benchmarks is often quite high since the evaluation is done in a context free environment.



Thank You