Automatic Speech Recognition

Instructor: Lara J. Martin (she/they)

TA: Omkar Kulkarni (he)

Slides adapted from Laurent Besacier

Learning Objectives

Distinguish between speech processing features and language modeling features

Explain how ASR has changed over time

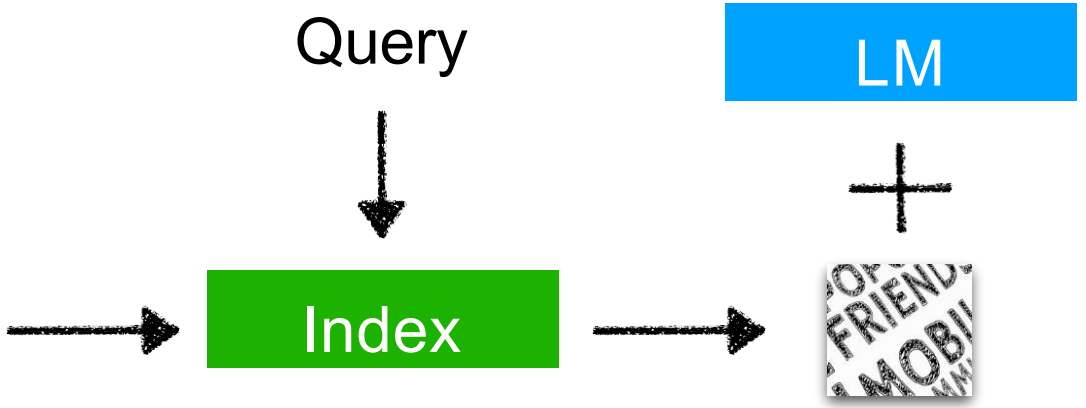
Compare and contrast modern ASR architectures

Determine how LLMs are used in ASR today

Learn modern issues in ASR

Review: RAG at Inference

Datastore



What is speech?

Speech generally conveys a (linguistic) message (that can be reduced to a transcript)

But also much more!

- Prosody: extra information beyond words, which conveys a meaning (e.g., tone, pitch, stress)
- Paralinguistics: speaker identity, speaker mood, speaker health condition, speaker accent, etc.
- Variability at all levels: intra speaker, inter speaker, microphone, phone line, room acoustics, style, etc.

(Main) Speech tasks

Speech compression (solved)

Speaker recognition (strong progresses over the last 10 years but still poor compared to other biometric modalities like fingerprint and iris)

Text-to-speech (TTS)/speech synthesis (can still gain in naturalness but new progresses with DL: Wavenet, Tacotron2, VoiceLoop)

Speech-to-text (this talk)

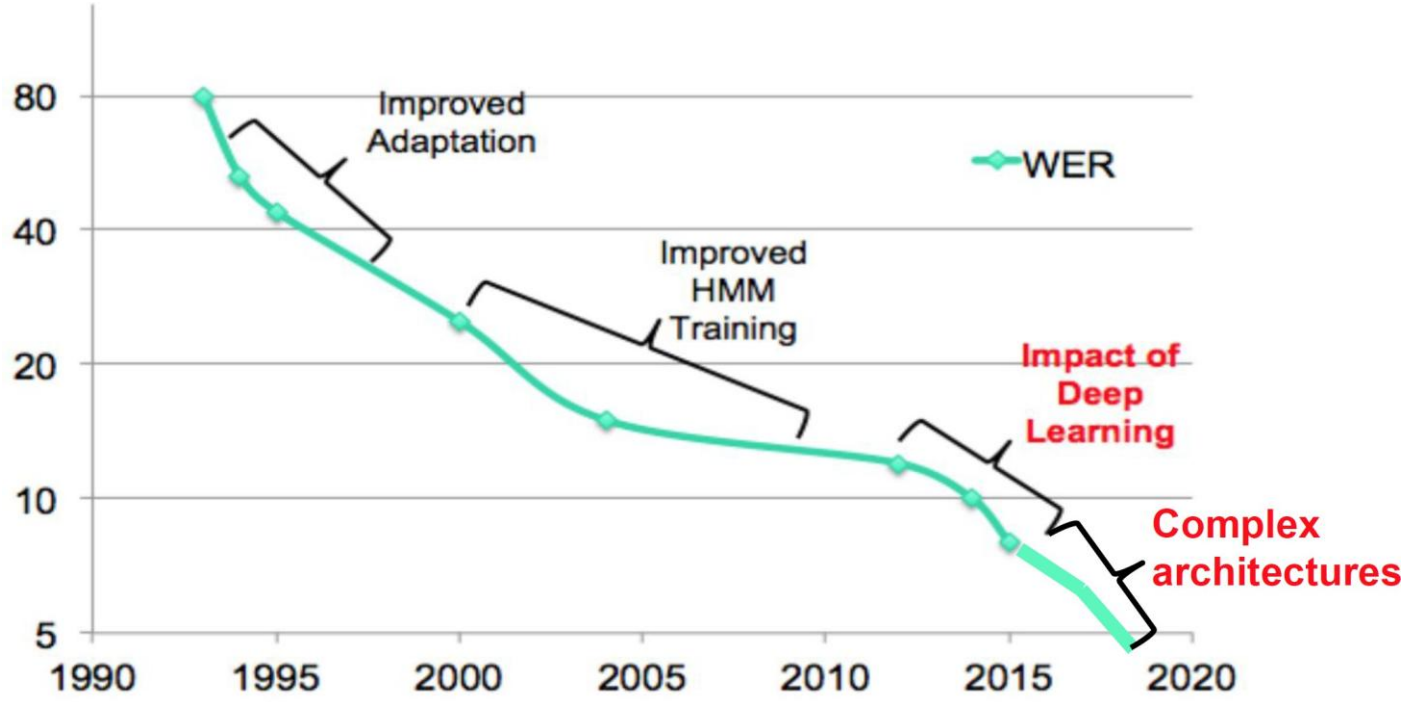
Speech paralinguistics (early days): detection of gender, age, deception, sincerity, nativeness, emotion, sleepiness, cognitive disorders, intoxication, pathologies, etc.

Speech-to-Text (Automatic Speech Recognition)

We want a system that can handle spontaneous speech, multi-speakers, unlimited output vocabulary, any acoustic condition

In reality, performance differs greatly for different contexts (read vs spontaneous speech; small vs large vocabulary; quiet vs noisy; native vs non-native speech)

Progress over the years



ASR Performance on English Conversational Telephony (Switchboard)

Image from Bhuvana Ramabhadran' s presentation at Interspeech 2018

ASR as a partial task in a larger system

ASR for spoken language processing (speech understanding, speech translation, speech summarization, etc.)

Not just a problem of noisy transcripts

No sentence boundaries, punctuation, case

Disfluencies in spontaneous speech: false starts, fillers, repaired utterances

- Should we keep them or remove them?
- Some speech tasks are ill defined (ex: speech translation)

Speech Signal

Speech is a continuous signal (no explicit word boundaries)

May be decomposed into elementary units of sound (phones) that distinguish one word from another in a particular language (minimal pairs)

- Minimal pairs: kill vs kiss - pat vs bat
- Phoneme set is the set of phones that are language dependent
- Acoustic realization of the phoneme is dependent of its left and right neighbors (co-articulation)

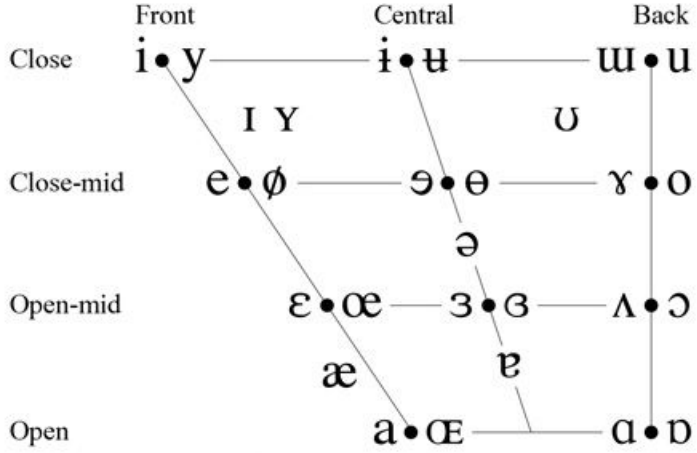
International Phonetic Alphabet

CONSONANTS (PULMONIC) © 2015 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	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			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

VOWELS



Where symbols appear in pairs, the one to the right represents a rounded vowel.

Lexicon (Vocabulary)

For acoustic modelling in large vocabulary speech recognition, we model phones instead of full words

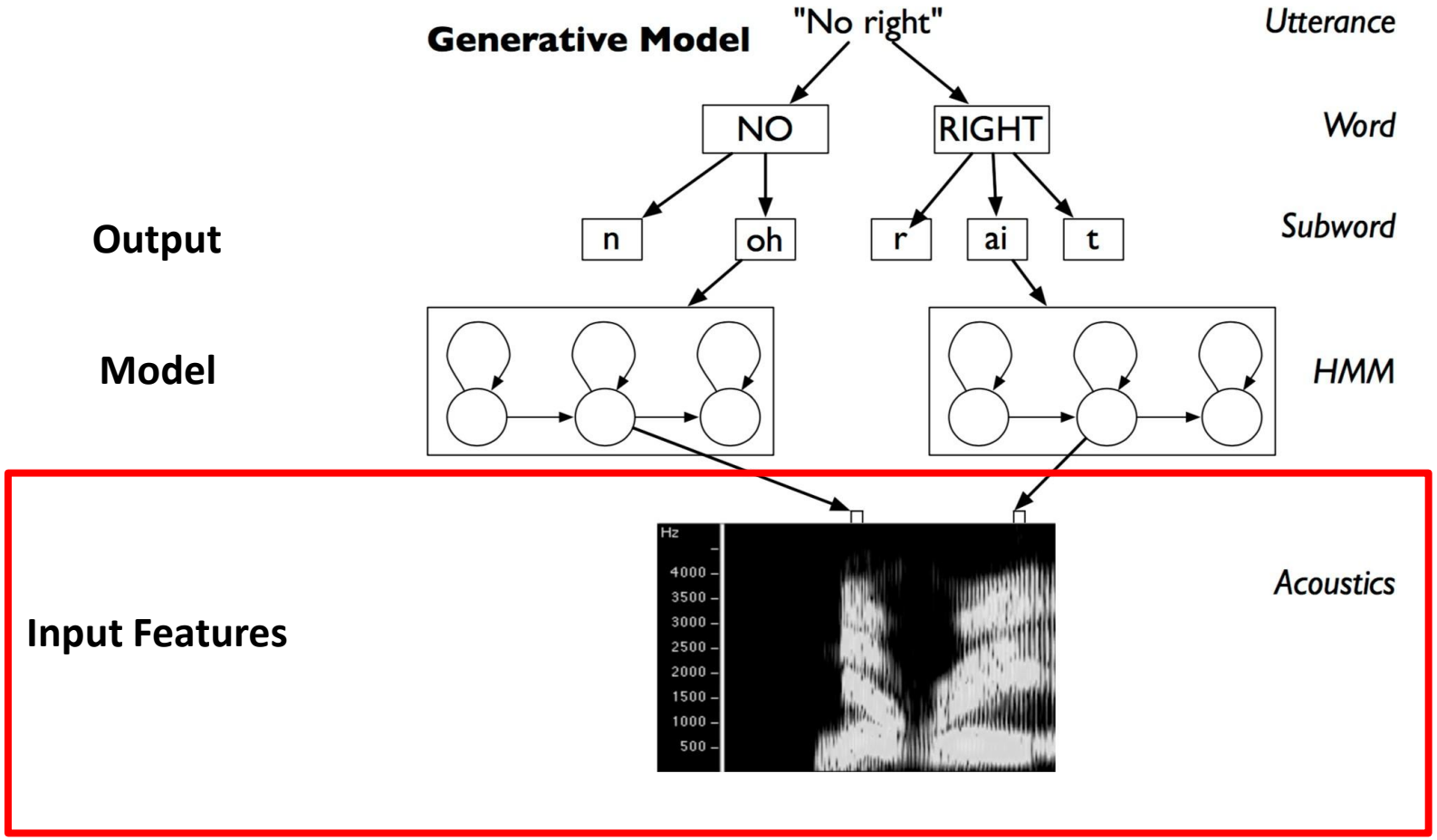
A pronunciation lexicon (e.g., CMU pronouncing dictionary) gives the decomposition of words into phonemes

Adding a new word to the output vocabulary does not require retraining of the acoustic models

- just add an entry to the pronunciation lexicon
- cat /k a t/

Hierarchical modelling of speech (signal/phones/words/utterance)

Hierarchical modelling of speech



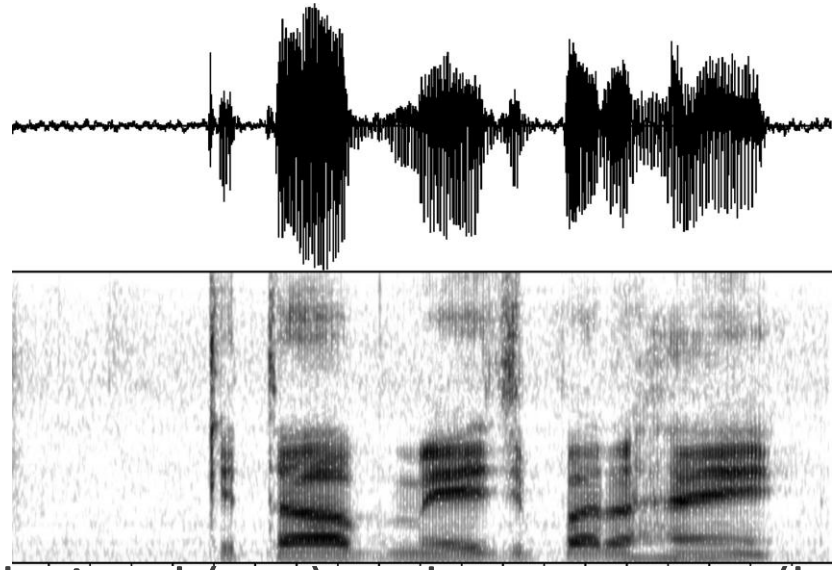
Speech Features

1. Handcrafted feature vectors
   - standard extraction on sliding windows of 20-30ms at a frame rate of 10ms
   - filterbanks (signal energy in different frequency bands)
   - cepstral coefficients (inverse Fourier transform of the logarithm of the estimated spectrum of a signal)
   - linear predictive coding (a sample is predicted as a weighted sum of preceding samples and weights are used as features)
   - prosodic features (pitch, energy)
2. Raw waveform (> 2015)
   - bypass handcrafted preprocessing
   - preprocessing become part of the acoustic modeling and training
   - introducing convolutional layers in the first stages of the NN pipeline

Speech Features

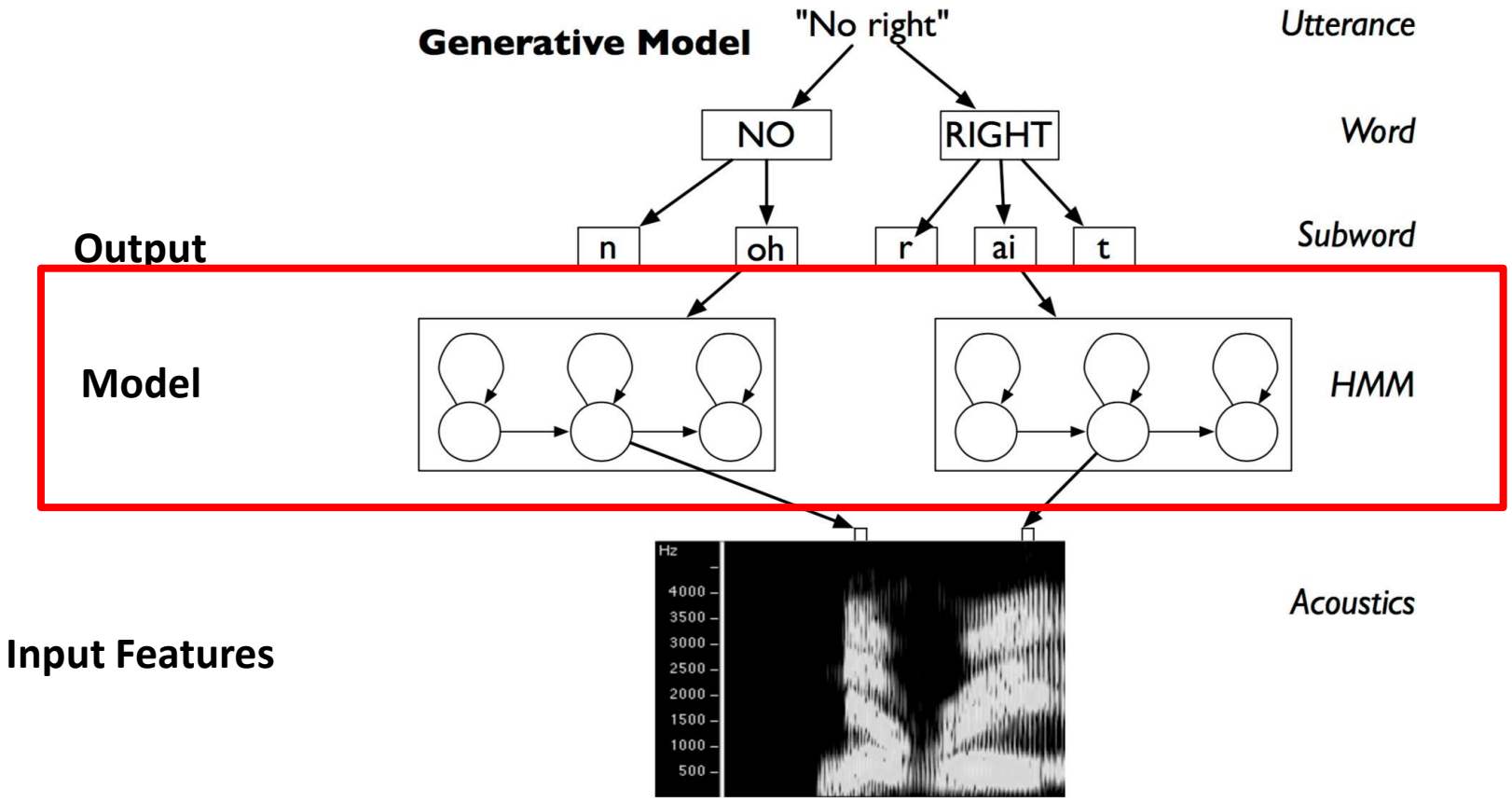
3. Spectrograms (< 1990 and > 2015!)

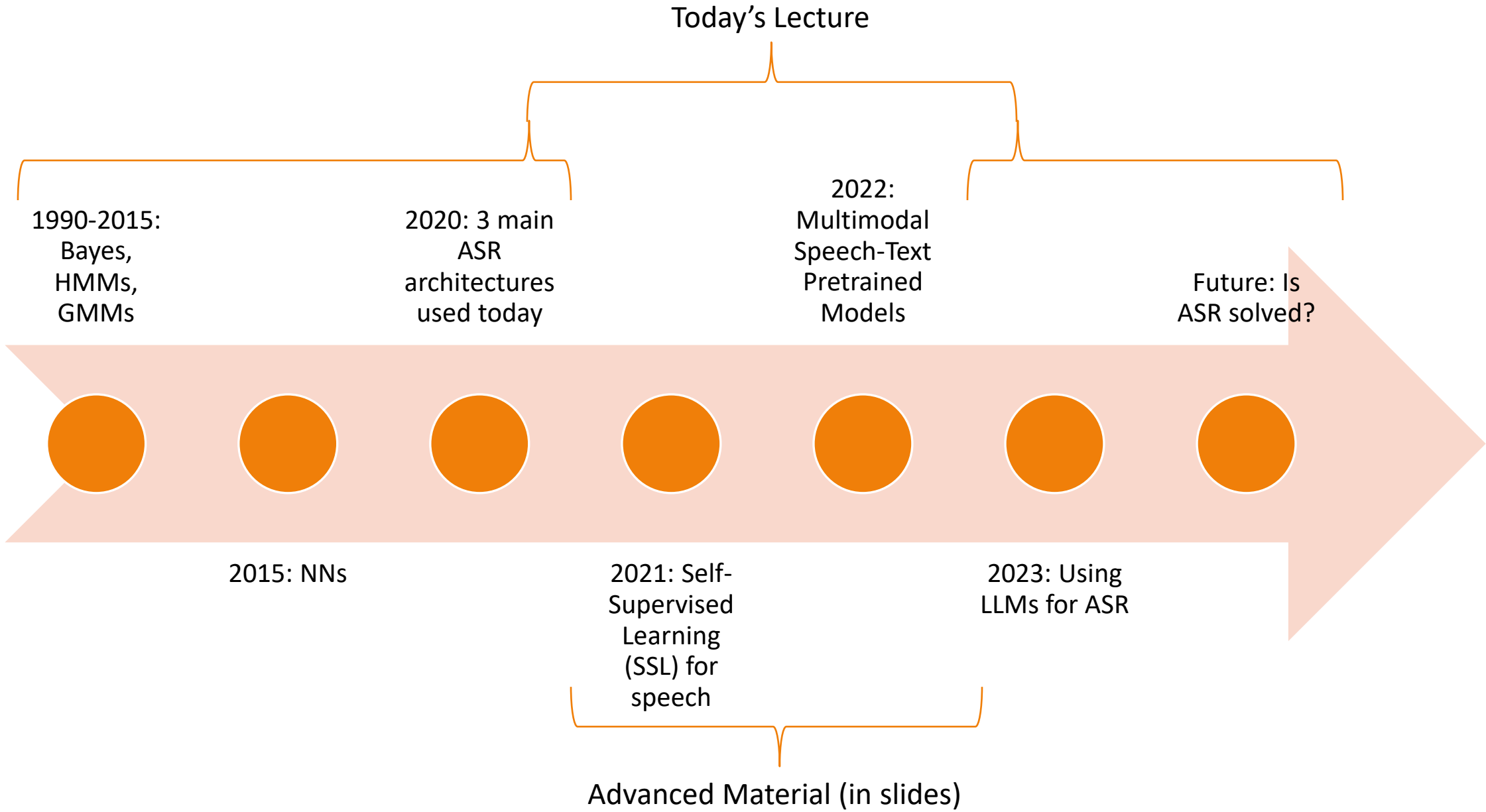
Representation used in phonetics to visualize speech, which captures intensity in addition to pitch over time; can be processed as an image!



Speech signal (top) and spectrogram (bottom)

Hierarchical modelling of speech




Today's Lecture
1990-2015: Bayes, HMMs, GMMs
2015: NNs
2020: 3 main ASR architectures used today
2021: Self-Supervised Learning (SSL) for speech
2022: Multimodal Speech-Text Pretrained Models
2023: Using LLMs for ASR
Future: Is ASR solved?
Advanced Material (in slides)

Review: Goal of Language Modeling

$$p_\theta(\textit{[...text..]})$$

Learn a probabilistic model of text

Accomplished through observing text and updating model parameters to make text more likely

Fundamental equation

x: observation (signal or features)

w: a word sequence

Naïve Bayes!

$$w^* = \text{argmax}_w\ p(\ w\ /\ x\ )$$
$$= \text{argmax}_w\ p(\ x\ /\ w\ )\ p(\ w\ )$$

$p(\ x\ /\ w\ )$: acoustic model

$p(\ w\ )$: language model

Acoustic modeling: HMM/GMM

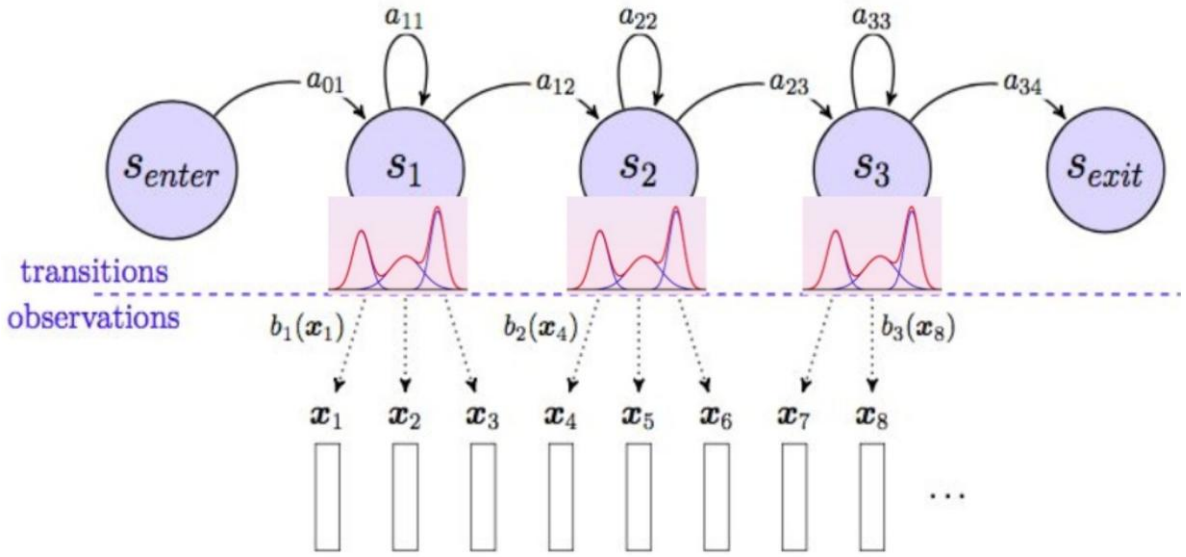
Complex sequential patterns of speech decomposed into piecewise stationary segments

Sequential structure of the data described by a sequence of states

- HMM (Hidden Markov Models) transitions

Local characteristics of the data described by a distribution associated to each state

- GMM (Gaussian Mixture Models) observations (outputs)



HMMs

Well known algorithms for

- *training* the model parameters (Baum-Welch algo.)
- *decoding* the most probable hidden state sequence (Viterbi algo.)
- *evaluating* the likelihood of an observation being generated by a HMM (Forward algo.)

Phonemes are generally modeled in context (1 phoneme = N HMMs)

- triphones or quintphones (model co-articulation)
- state or parameter tying to reduce model complexity

Language models: from N-grams to RNNs

For a sequence of T words $W = w_1, w_2, \ldots, w_T$

$$P(W) = \prod_{k=1}^{T} P(w_k | w_1, w_2, \ldots, w_{k-1})$$

$$P(W) = \prod_{k=1}^{T} P(w_k | h)$$

n-gram LM: $h = w_{k-n+1}, w_{k-n+2}, \ldots, w_{k-1}$

recurrent neural network LM: $h = rnn\ state(E(w_1), E(w_2), \ldots, E(w_{k-1}))$

NNs in the 90s and 00s

Introduced in the 80s and 90s to speech recognition, but extremely slow and poor in performance compared to the state-of-the-art HMM/GMM

Pros: no assumption about a specific data distribution

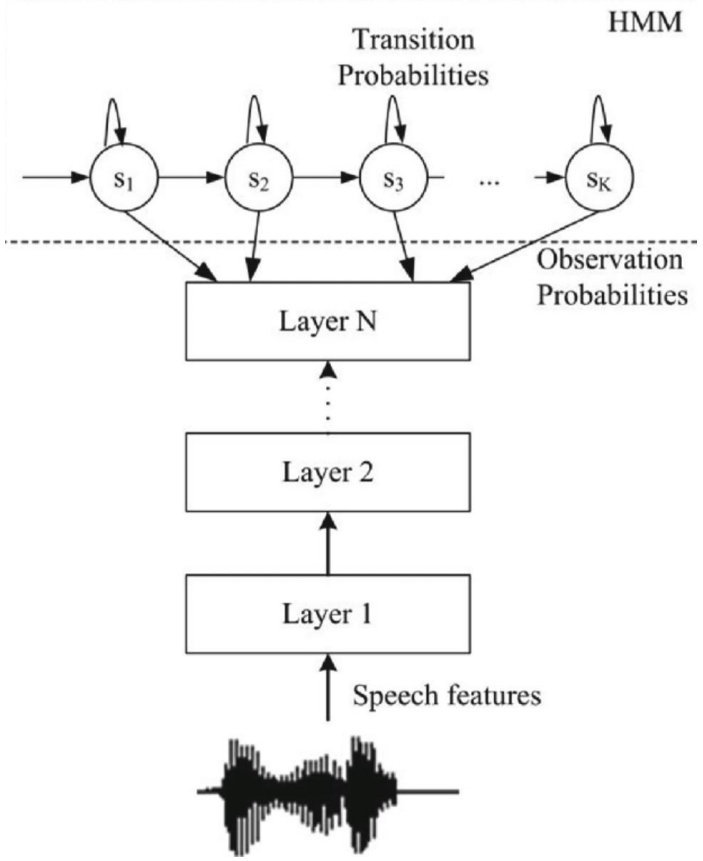
Cons: slow and do not scale to large tasks

NNs for acoustic modeling (1990-2010)

In most approaches, NNs model the posterior probability $p(s|x)$ of an HMM state s given an acoustic observation x

Existing HMM speech recognizers can be used

This model is known as hybrid NN-HMM and was introduced by Renals et al. (1994)



Hybrid NN-HMM

NNs for language modeling (1990-2010)

Rescoring a lattice of output hypotheses using NN LM instead of N-gram

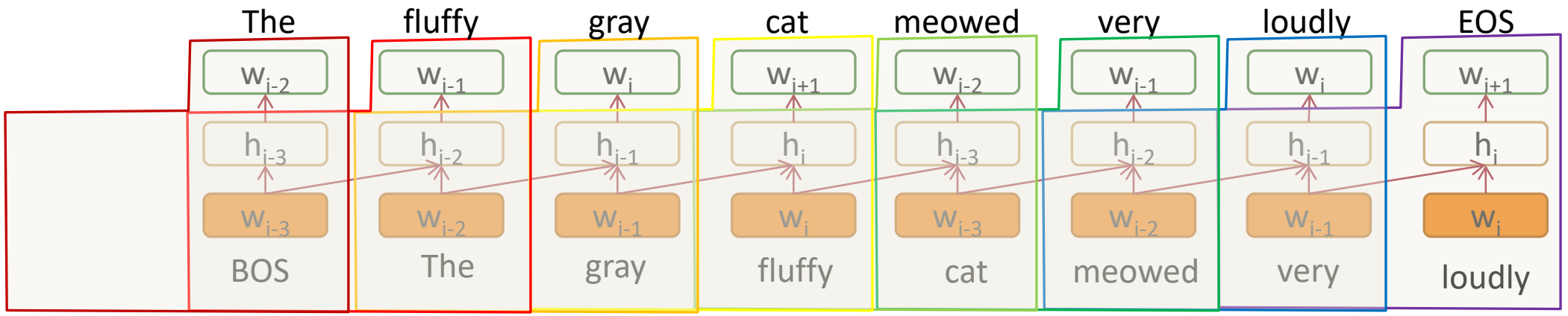
Extended to large vocabulary speech recognition (Schwenk, 2007)

Reducing computational complexity

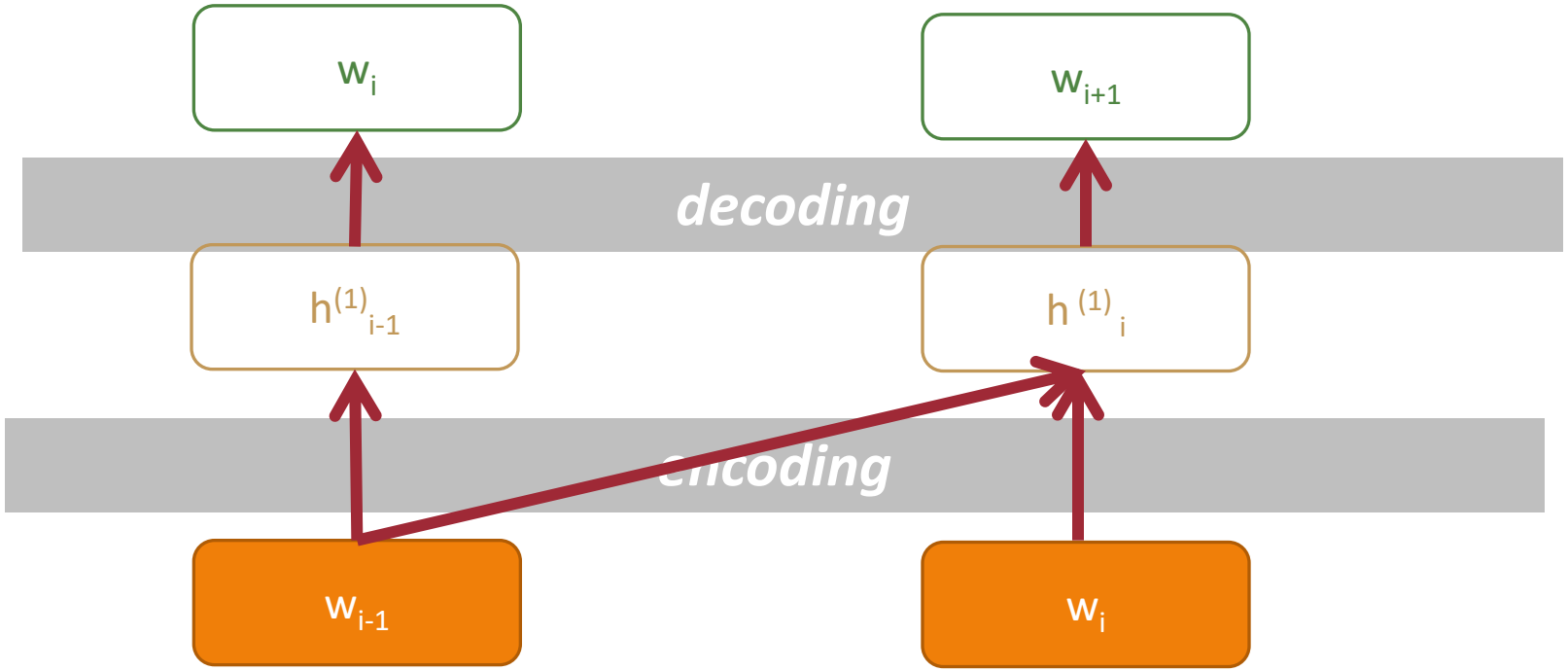
- hierarchical decomposition of output probabilities (Morin and Bengio, 2005; Mnih and Hinton, 2008; Le et al., 2011)

Review: A Neural N-Gram Model (N=3)

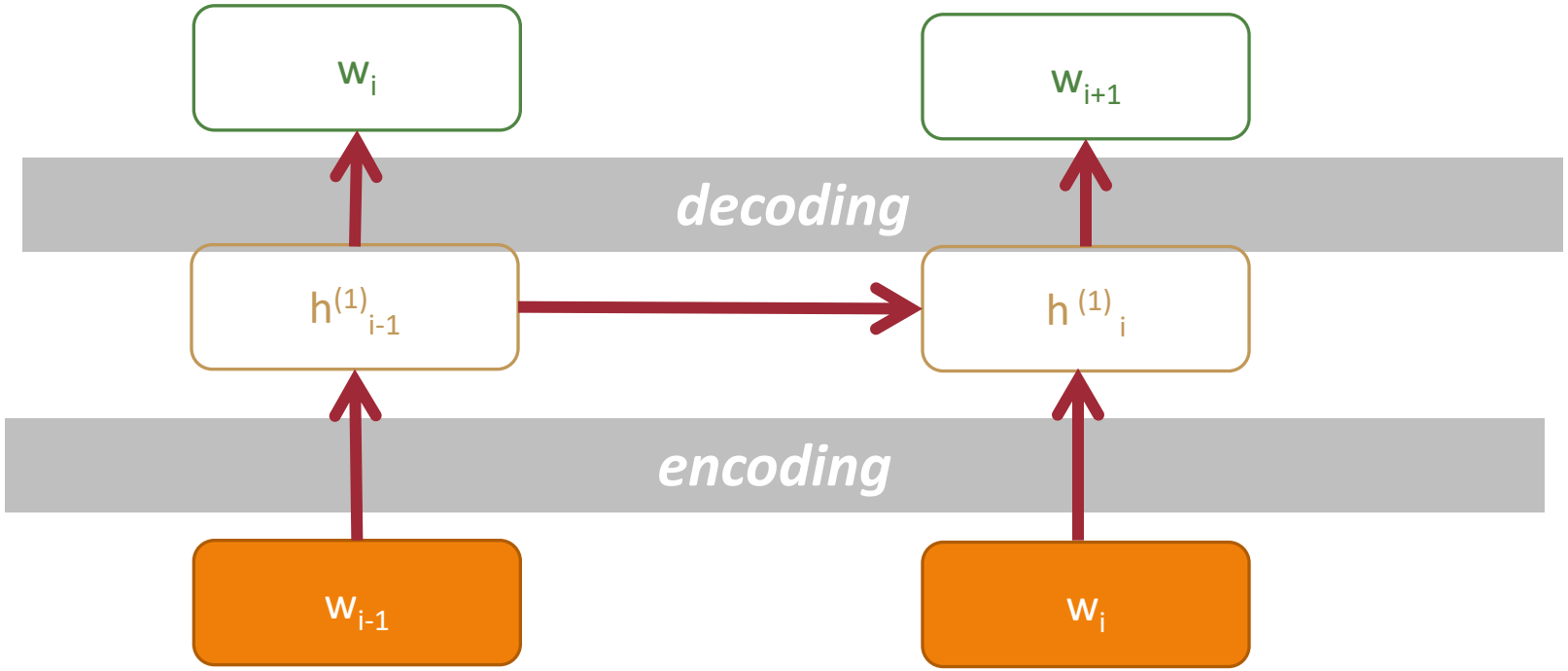
The fluffy gray cat meowed very loudly



Review: Trigram Feedforward NN



Review: RNN



Deep learning breakthrough

Like in vision, due to…

More data

- ex: (2015) Librispeech (en) 1,000h (Panayotov et al., 2015)
- ex: (2016) Baidu Deep Speech 2 (en) 12,000h (Amodei et al., 2016)
- ex: (2017) Google Home (en) 18,000h (from a Google presentation)
- ex: (2018) Google wav2words (en) >100.000h (informal discussion)
- ex: (2021) Meta XLS-R >436,000h (Babu et al., 2021) (self-supervised)
- ex: (2022) OpenAI Whisper model trained on 680,000 hours of multilingual speech3 (Alec Radford, 2022)

Computation (ex: GPU)

Better optimization algorithms and training objectives

ASR Toolkits (ex: Kaldi (Povey et al., 2011)) and DL frameworks (Tensorflow, Pytorch)

End-to-end ASR (get rid of HMMs)

1) Connectionist Temporal Classification (CTC)

- Solves the problem of unaligned input and output sequences by marginalizing the conditional likelihood of the output sequence given the input over all possible alignments

2) Attention Modeling

- Simultaneously optimize alignment and grapheme (or word) decoding using attention weights (linear combination of hidden states) to influence the generated output

3) Transducer-based

- Allow to decouple the acoustic model from the language model; elegant to leverage larger amounts of raw text (for LM)

1) CTC

Learns to align the transcript itself during training (Graves et al., 2006)

Tries to label sequence z (of length M)

Uses blank or _ symbol to allow M-length output sequence to be mapped to a T-length input sequence *x*

z can be represented by a set of all possible CTC paths (sequence of labels, at frame level) that are mapped to z

- ex: M=2 (z = hi) and T=3 (3 frames): possible sequences are ’hhi’, ’hii’, ’ _hi’, ’h_i’, ’hi_’

Probability p(z|x) evaluated as sum of probabilities over all possible CTC paths (using Forward-Backward)

We start with an input sequence, like a spectrogram of audio.

The input is fed into an RNN, for example.

The network gives $p_t(a \mid X)$, a distribution over the outputs {h, e, l, o, ϵ} for each input step.

With the per time-step output distribution, we compute the probability of different sequences

By marginalizing over alignments, we get a distribution over outputs.

CTC loss function

CTC loss can be very expensive to compute

The problem is there can be a massive number of alignments

We can compute the loss much faster with a dynamic programming algorithm

$$p(Y \mid X) = \sum_{A \in \mathcal{A}_{X,Y}} \prod_{t=1}^{T} p_t(a_t \mid X)$$

The CTC conditional **probability**

marginalizes over the set of valid alignments

computing the **probability** for a single alignment step-by-step.

from https://www.assemblyai.com/blog/end-to-end-speech-recognition-pytorch

CTC inference

Greedy decoding

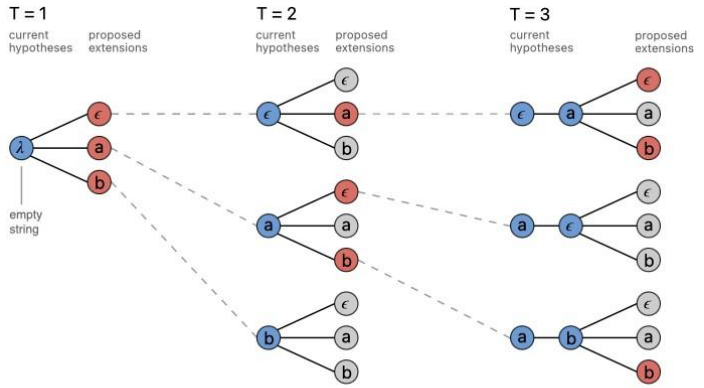
$$Y^* = \underset{Y}{\operatorname{argmax}}\ p(Y \mid X)$$

One heuristic is to take the most likely output at each time-step. This gives us the alignment with the highest probability:

$$A^* = \underset{A}{\operatorname{argmax}} \prod_{t=1}^{T} p_t(a_t \mid X)$$

We can then collapse repeats and remove ϵ tokens to get Y.

Beam-Search decoding



A standard beam search algorithm with an alphabet of $\{\epsilon, a, b\}$ and a beam size of three.

Figure: from https://www.assemblyai.com/blog/end-to-end-speech-recognition-pytorch

Problems with CTC

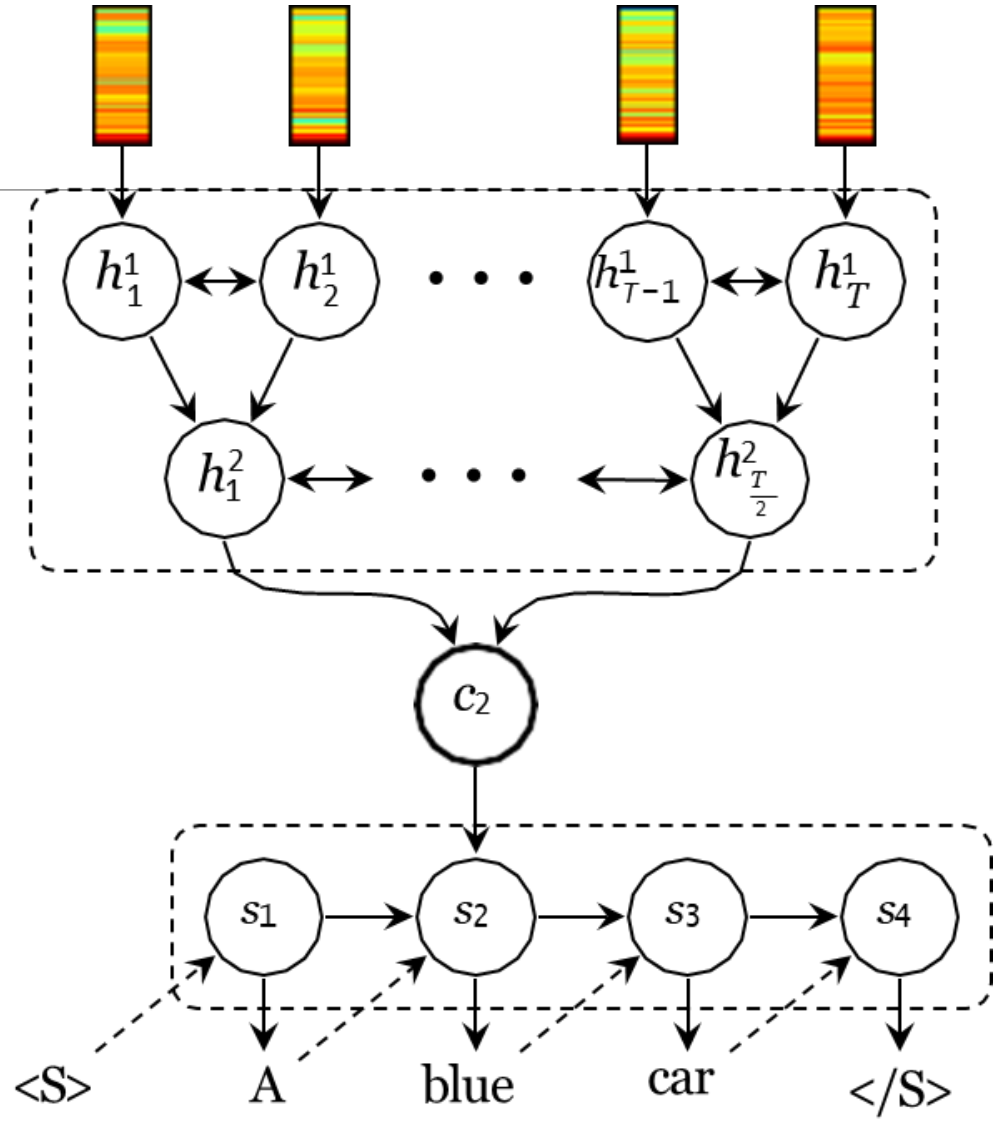
The output sequence length M has to be smaller than the input sequence length T (prevents models that do a lot of input pooling)

The outputs are assumed to be independent of each other. CTC models often produce wrong outputs like “I eight food”

2) Attention modeling

Architecture similar to sequence-to-sequence model

Speech encoder based on CNNs or pyramidal LSTMs?



Attention modeling

Initially proposed for (neural) machine translation (Bahdanau et al., 2014) and introduced for ASR by Chorowski et al. (2015)

- A context (attention) model is a function of the encoder codes and of the previous decoded tokens
- A speech encoder is defined (CNNs, pyramidal LSTMs)
- While CTC generates frame-level predictions, attention models generate L predictions until the end-of-sequence symbol (no posterior for a given frame)
- Well-known issue with attention and CTC models is the thin lattices

Attention modeling (different view)

Also called LAS: **L**isten (encode), **A**ttend (attention), and **S**pell (decode)

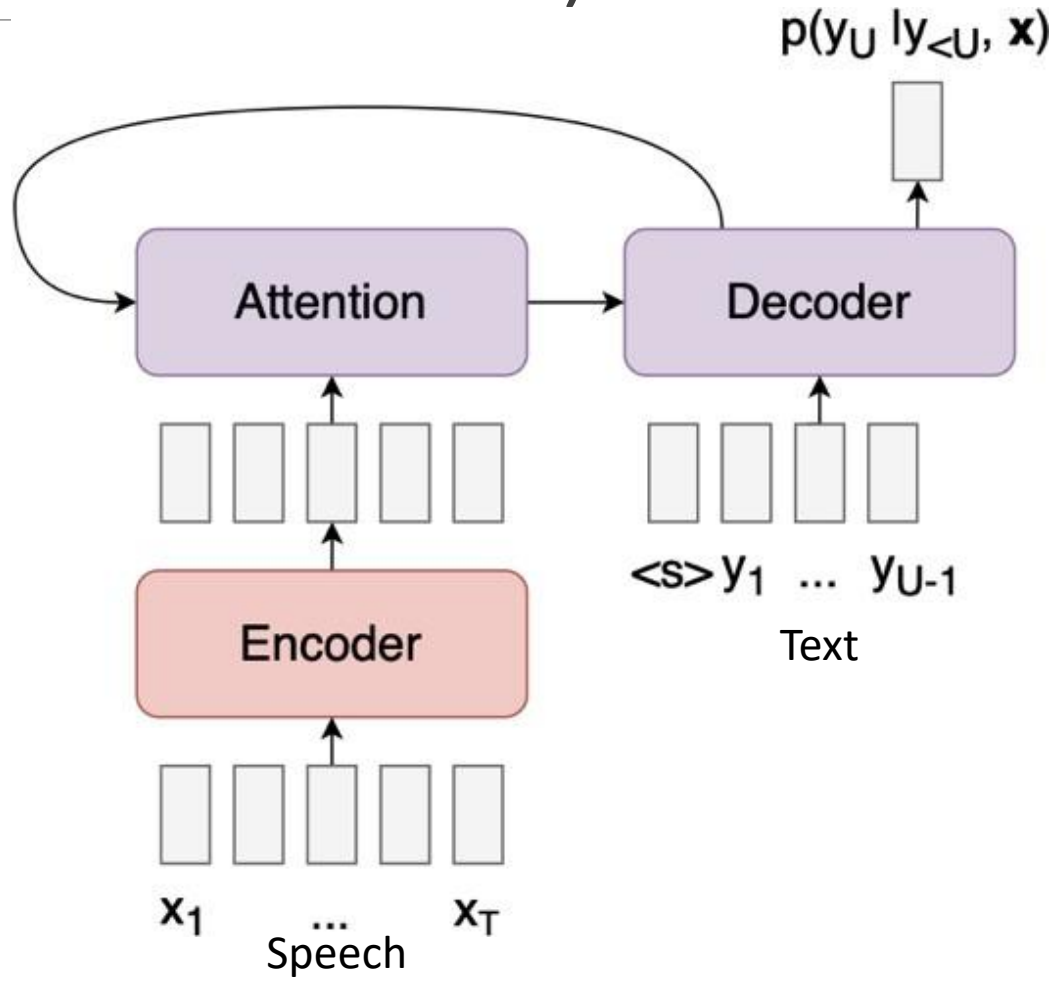


Image from
https://lorenlugosch.github.io/posts/2020/11/transducer/

Attention modeling (alignment)

Allows non-monotonic alignments

As opposed to CTC (monotonic)

Monotonic: one function is always larger than (or equal to) the other

f(x) > f(y) or f(y) > f(x)

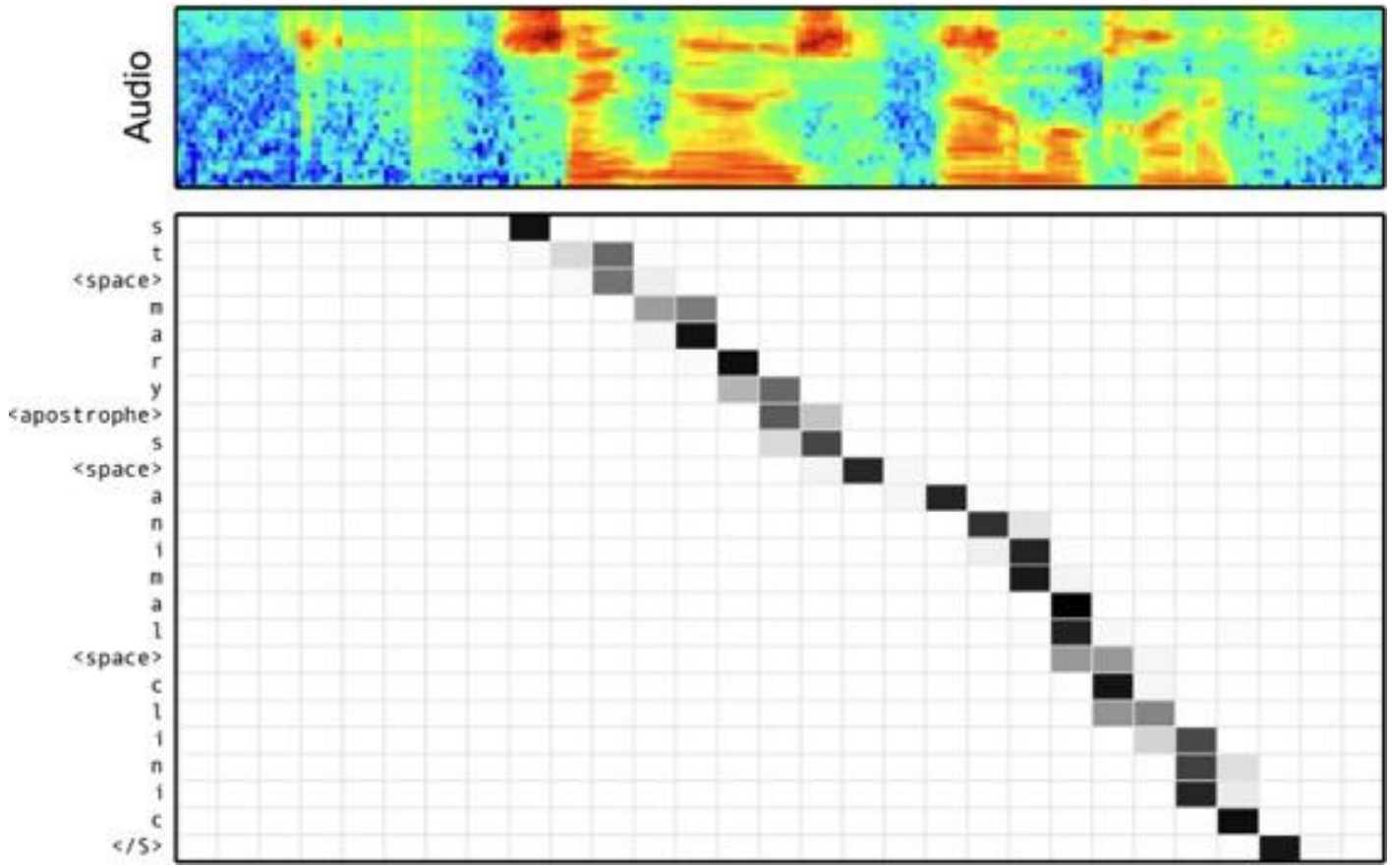


Image from
https://lorenlugosch.github.io/posts/2020/11/transducer/

CTC vs Attention

CTC

Output needs to be smaller than input

Prediction is for every frame

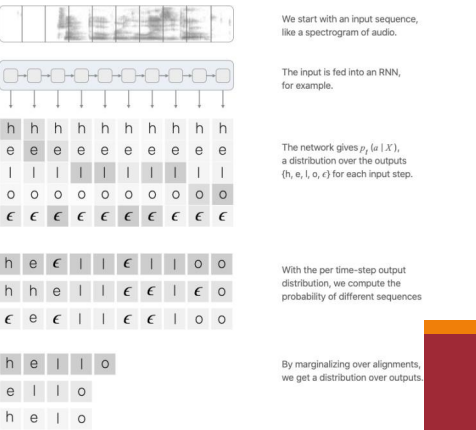
Each prediction is independent

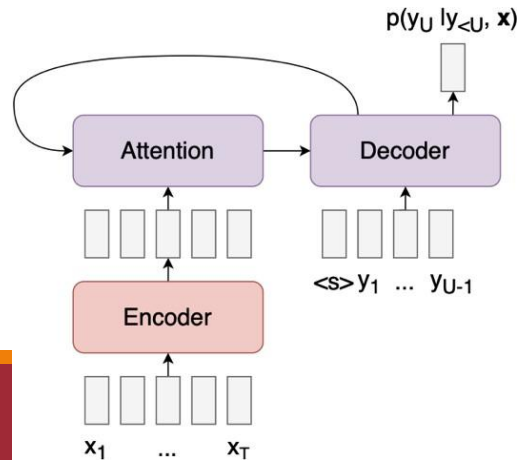
ATTENTION

Alignment can be non-monotonic

Predict until reaches the end symbol

Predictions from “decoder” can take history into consideration





3) Transducer models

Predictor is a language model

Joiner is a simple feed-forward network

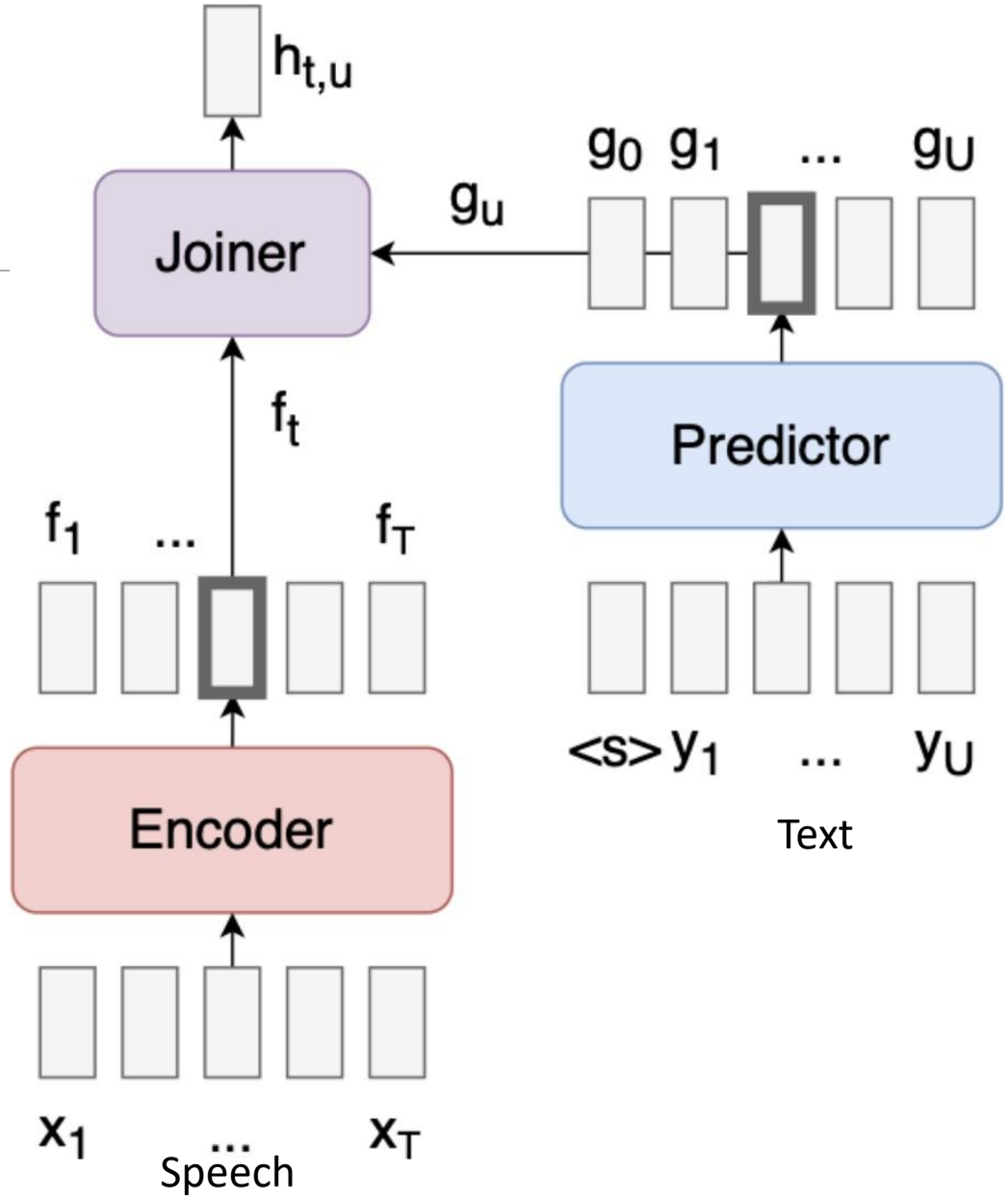


Image from https://lorenlugosch.github.io/posts/2020/11/transducer/

2020: 3 main ASR architectures used today

Transducer: Search

1. Start by setting $t := 1$, $u := 0$, and $\mathbf{y} :=$ an empty list.
2. Compute f_t using $\mathbf{x}$ and g_u using $\mathbf{y}$.
3. Compute $h_{t,u}$ using f_t and g_u.
4. If the argmax of $h_{t,u}$ is a *label*, set $u := u + 1$, and output the label (append it to $\mathbf{y}$ and feed it back into the predictor).

   If the argmax of $h_{t,u}$ is $\varnothing$, set $t := t + 1$ (in other words, just move to the next input timestep and output nothing).
5. If $t = T + 1$, we're done. Else, go back to step 2.

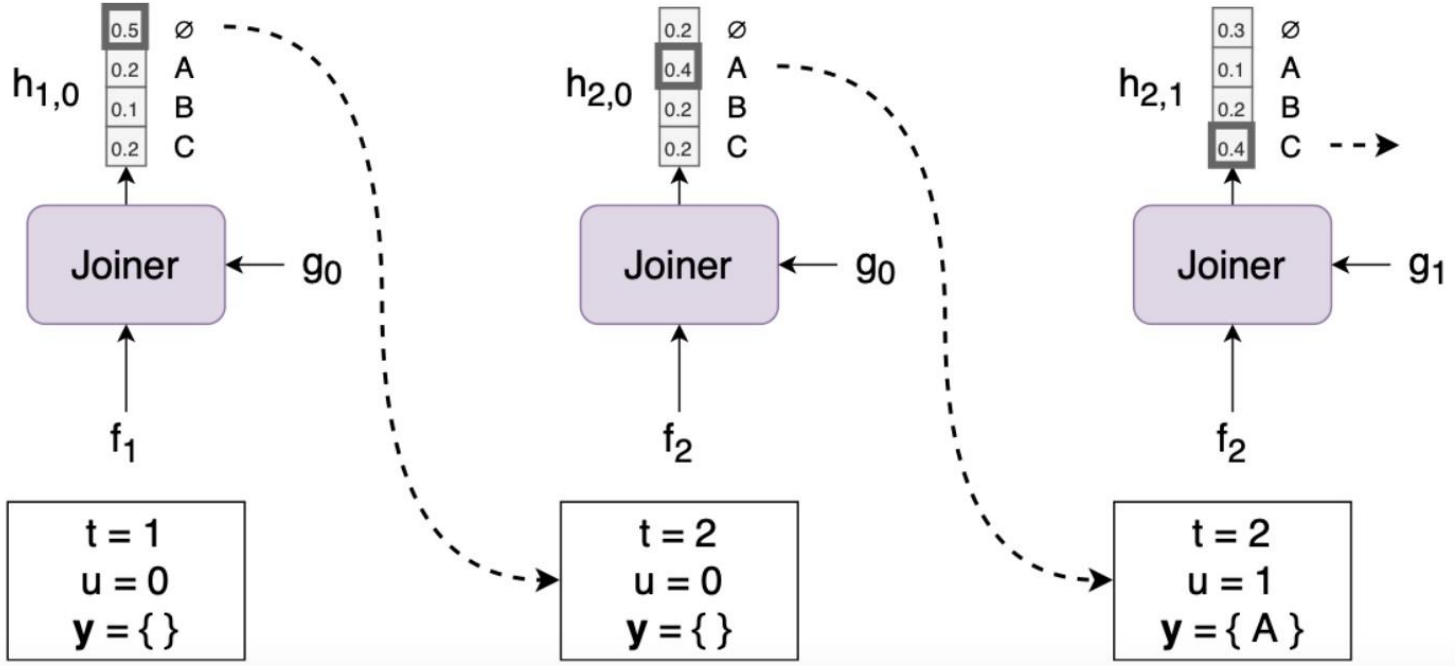


Image from https://lorenlugosch.github.io/posts/2020/11/transducer/

Attributes of Transducer models

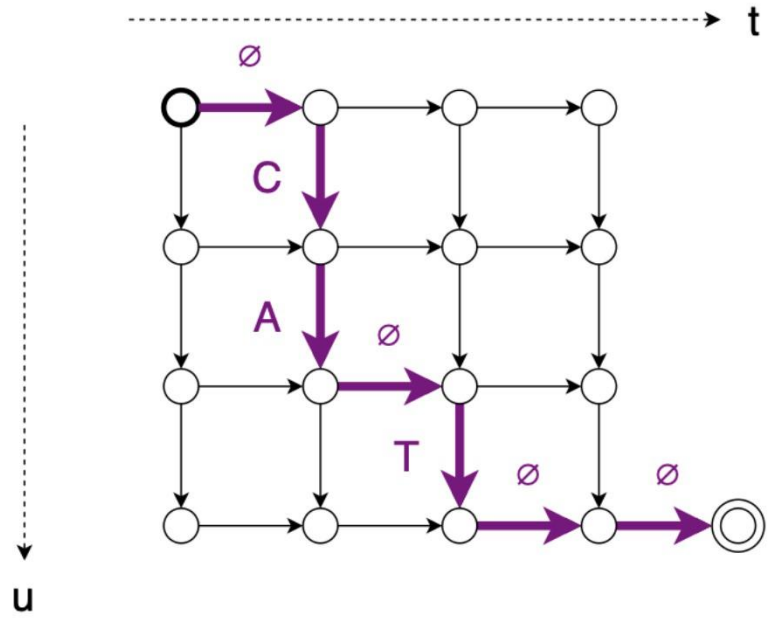
If encoder is causal (not using something like a bidirectional RNN), then search can run in an online/streaming fashion

The predictor only has access to *y* (text) -- not *x* (speech) -- unlike the decoder in an attention model

- we can easily pre-train the predictor on text-only data

Naturally defines alignment between *x* and *y*

Here's one alignment: $\mathbf{z} = \varnothing, C, A, \varnothing, T, \varnothing, \varnothing$



CTC vs Attention vs Transducer

CTC

Output needs to be smaller than input

Prediction is for every frame

Each prediction is independent

ATTENTION

Alignment can be non-monotonic

Predict until reaches the end symbol

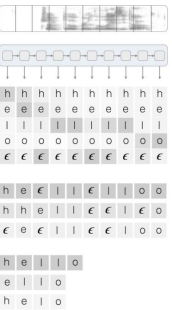
Predictions from “decoder” can take history into consideration

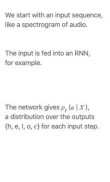
TRANSDUCER

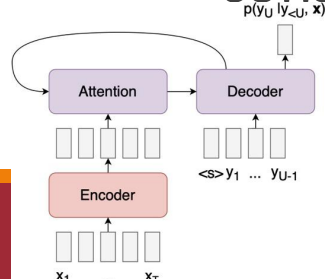
Two separate models both fed into third joiner model

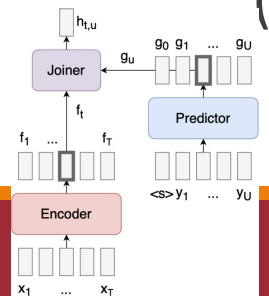
Speech and text are processed separately

Can make predictions continuously (streaming)









Jump to Using LLMs for ASR

Self supervised representation learning

Using huge unlabeled data for training ; targets are computed from the signal itself

- "*learn representations using objective functions similar to those used for supervised learning, but train networks to perform pretext tasks where both the inputs and labels are derived from an unlabeled dataset*" (from Chen et al. (2020))

Introduced for vision: see for instance (Chen et al., 2020)

- learn representations by contrasting positive pairs against negative pairs

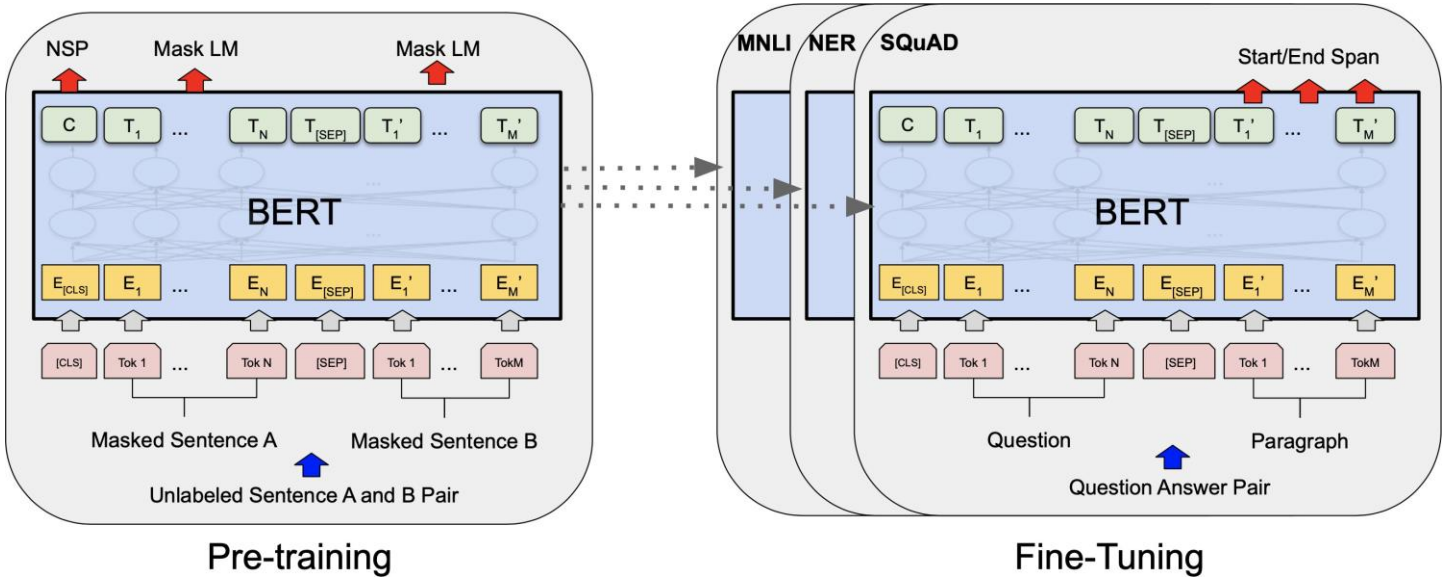
Introduced also in NLP: see for instance (Devlin et al., 2018)

learn representations by predicting tokens that were masked in an input sequence

Pre-trained language models

Leverage large amount of freely available unlabeled text to facilitate transfer learning in NLP

Yield state-of-the-art results on a wide range of NLP tasks + save time and computational resources



Example of BERT (Devlin et al., 2018) based on the Transformer model (Vaswani et al., 2017)

Self supervised representation learning from speech

Autoregressive predictive coding (APC) (Chung et al., 2019; Chung and Glass, 2020)

- Considers the sequential structure of speech and predicts information about a future frame

Contrastive Predictive Coding (CPC) (Baevski et al., 2019; Schneider et al., 2019a; Kahn et al., 2019)

- Easier learning objective which consists in distinguishing a true future audio frame from negatives

Other approaches for feature representation learning using multiple self supervised tasks (Pascual et al., 2019; Ravanelli et al., 2020) or bidirectional encoders (Song et al., 2019; Liu et al., 2020; Wang et al., 2020)

Autoregressive predictive coding (APC)

Predicting the spectrum of a future frame (rather than a wave sample) (Chung et al., 2019)

Somewhat inspired by language models (LMs) for text, which are typically a probability distribution over sequences of T tokens (t_1, t_2, ..., t_T)

$$P(sequence) = \prod_{k=1}^{T} P(t_k | t_1, t_2, \ldots, t_{k-1})$$

$$P(sequence) = \prod_{k=1}^{T} P(t_k | h)$$

Recurrent neural network LM:

- *h = rnn_state(E(t_1), E(t_2), ..., E(t_{k-1}))*

For speech, each token t_k corresponds to a frame rather than a word or character token

Autoregressive predictive coding (APC)

No final set of target tokens (softmax layer replaced by a regression layer)

Learnable parameters in APC are the RNN parameters θ_{RNN} and the regression layer parameters θ_r

Encourage APC to infer more global structures rather than the local information in the signal

- ask the model to predict a frame n steps ahead of the current one

Model is optimized by minimizing the L1 loss between sequence (x_1, x_2, ..., x_T) and the predicted sequence (y_1, y_2, ..., y_T):

$$\sum_{i=1}^{T-n} |t_i - y_i|, t_i = x_{i+n}$$

Autoregressive predictive coding (APC)

Chung et al. (2019) models APC with a multi-layer unidirectional LSTM with residual connections

After training, RNN hidden states are taken as the learned representations

A follow-up work (Chung and Glass, 2020) adds as regularization to improve generalization

Feature	ASR (WER ↓)	ST (BLEU ↑)
log Mel	18.3	12.9
APC w/ L_f	15.2	13.8
APC w/ L_m	14.2	14.5

Table 2: Automatic speech recognition (ASR) and speech translation (ST) results using different types of features as input to a seq2seq with attention model. Word error rates (WER, ↓) and BLEU scores (↑) are reported for the two tasks, respectively.

Differences between speech and text SSL

Input speech representations (MFCCs for instance) are already in a vector form (no embedding layer)

More uncertainty

- text (discrete): finite number of possible outcomes (target tokens)
- speech and video (continuous): infinite number of frames that can plausibly follow a given audio (or video) clip

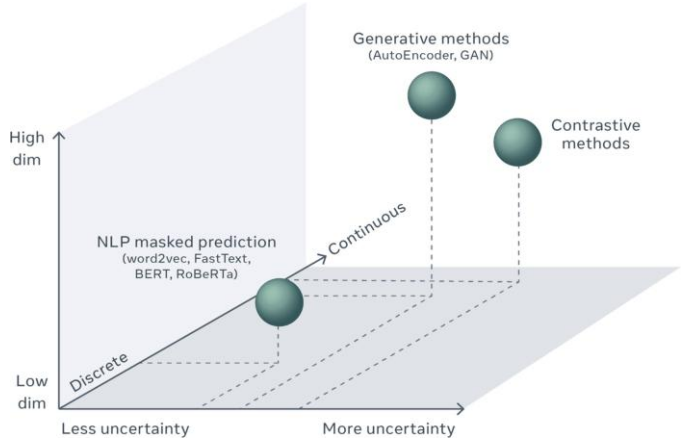


 Figure from https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence

Contrastive Predictive Coding (CPC)

Idea proposed by van den Oord et al. (2018)

Maybe an easier learning objective (classification instead of regression) Use of a contrastive loss that distinguishes a true future audio sample from negatives

Example of wav2vec (Schneider et al., 2019b) that relies on a fully convolutional architecture

Applied the learned representations to improve a supervised ASR system

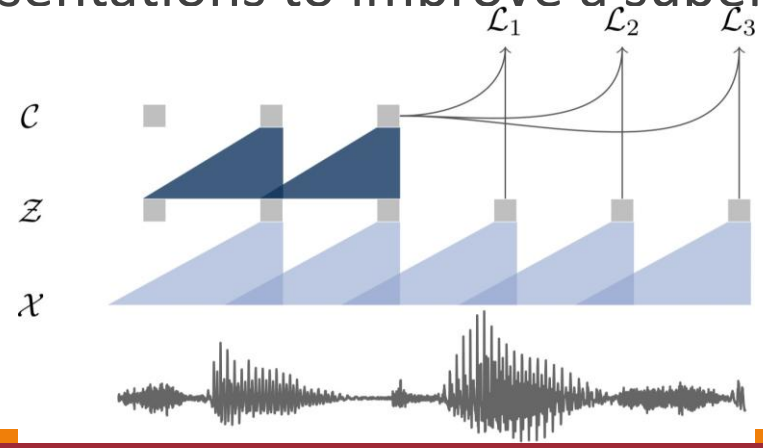


Figure from (Schneider et al. ,2019b)

Contrastive Predictive Coding (CPC)

Encoder network $Z = f(X)$; 5 (causal) convolution layers ; local feature representations z_i encode 30 ms of audio every 10ms

Context network $C = g(Z)$; 9 (causal) convolution layers ; mix multiple z_i (receptive field of dimension v corresponding to 210ms) into a single contextualized representation c_i

Model trained to distinguish a sample z_{i+k} that is k steps in the future from distractor samples $\tilde{z}$ drawn from a proposal distribution p_n by minimizing a contrastive loss for each step $k = 1, ..., K$

Negatives examples sampled by uniformly choosing distractors from each audio sequence: is $p_n(z) = 1/T$ where T is the sequence length

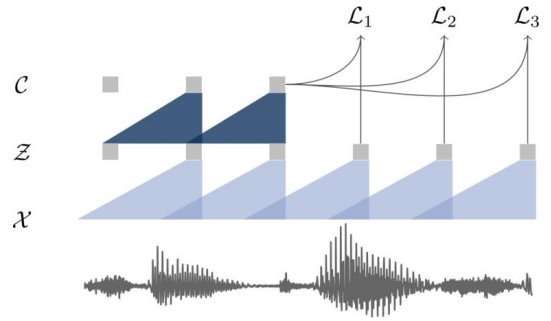


Figure from (Schneider et al., 2019b)

Representation learning with multiple self-supervised tasks

Problem-agnostic speech encoder (PASE) (Pascual et al., 2019)

PASE+: robust speech recognition in noisy and reverberant environments (Ravanelli et al., 2020)

Representation learning with multiple self-supervised tasks

Problem-agnostic speech encoder (PASE) (Pascual et al., 2019)

Jointly tackle multiple self-supervised tasks using an ensemble of neural networks that cooperate to discover good speech representations

Approach requires consensus across tasks, more likely to learn general, robust, and transferable features

Authors find that such representations outperform more traditional hand-crafted features in different speech classification tasks such as speaker identification, emotion classification, and ASR

Problem-agnostic speech encoder (PASE)

Encoder: SincNet (Ravanelli and Bengio, 2018) + Convblocks (receptive field 150ms)

Workers: one for each task (see next slide)

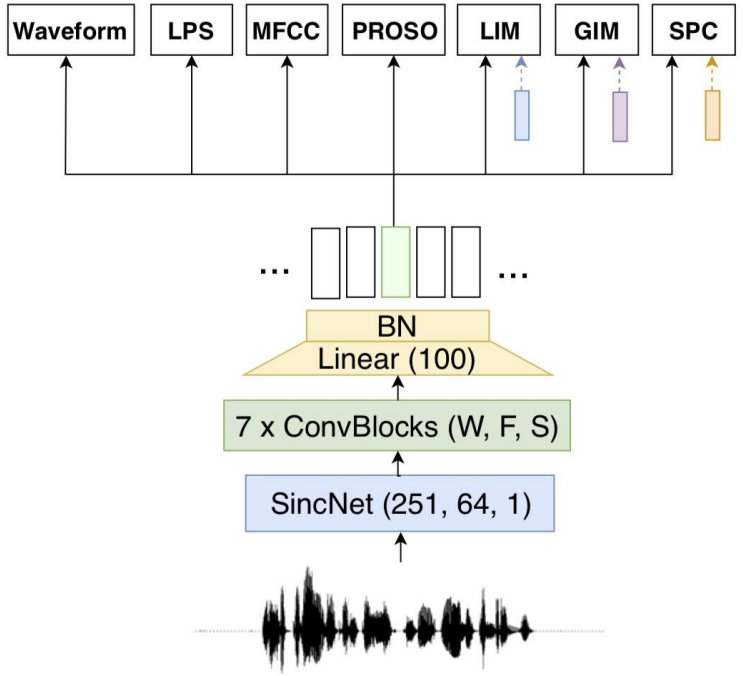


Figure 1: *The PASE architecture, with the considered workers.*

Figure from (Pascual et al., 2019)

Problem-agnostic speech encoder (PASE)

Regression workers that solve 7 self-supervised tasks

Trained to minimize the mean squared error (MSE) between the target features and the network predictions

- **Waveform** learns to reconstruct waveforms
- **LPS** reconstruct log power spectrum
- **MFCC** reconstruct mel-frequency cepstral coefficients
- **Prosody** predicts 4 basic prosodic features per frame
- **LIM** (local info max) contrastive task where positive sample is drawn from the same utterance and a negative sample is drawn from another random utterance (that likely belongs to a different speaker)
- **GIM** (global info max) similar to LIM using global representations (averaged over 1s) instead of local ones
- **SPC** sequence predicting coding: similar to contrastive predictive coding (CPC) introduced earlier

Problem-agnostic speech encoder (PASE)

Experiments on speaker identification, emotion recognition and ASR

Table 2: *Accuracy comparison on the considered classification tasks using MLPs and RNNs as classifiers.*

Model	Classification accuracy [%]					
	Speaker-ID *(VCTK)*		Emotion *(INTERFACE)*		ASR *(TIMIT)*	
	MLP	RNN	MLP	RNN	MLP	RNN
MFCC	96.9	72.3	90.8	91.1	81.1	84.8
FBANK	98.4	75.1	94.1	92.8	80.9	85.1
PASE-Supervised	97.0	80.5	93.8	92.8	82.1	84.7
PASE-Frozen	97.3	82.5	91.5	92.8	81.4	84.7
PASE-FineTuned	**99.3**	**97.2**	**97.7**	**97.0**	**82.9**	**85.3**

Table 3: *Word error rate (WER) obtained on the DIRHA corpus.*

	WER [%]
MFCC	35.8
FBANK	34.0
PASE-Supervised	33.5
PASE-Frozen	32.5
PASE-FineTuned	**29.8**

Many follow-up approaches

Speech-XLNet: Unsupervised Acoustic Model Pretraining For Self-Attention Networks (Song et al., 2019)

Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders (Liu et al., 2020)

Unsupervised pre-training of bidirectional speech encoders via masked reconstruction (Wang et al., 2020)

Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations (Baevski et al., 2020)

HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units (?)

Speech-XLNet (Song et al., 2019)

Learn speech representations with self-attention networks

BERT-like autoencoding (AE) scheme to train a bi-directional speech representation model (not only left-to-right)

Mask and reconstruct speech frames rather than word tokens (regression instead of classification task)

Encourage network to learn global structures by shuffling speech frame orders (can be also seen as dynamic data augmentation)

Training using a Mean Absolute Error (MAE) loss over several permutations of the input frames

(Unfortunately) not compared with previous APC and CPC approaches

Speech-XLNet

Experiments on Hybrid and end-to-end ASR

Results of Hybrid ASR on TIMIT are reported below

Table 2: *PER comparison with previous pretrain methods. We approximate the number of parameters based on the description in the previous studies.*

Pretrain Method	Pretrain Data	Pretrain Params	Dev/Test PER(%)
VQ-Wav2vec ([8])	libri (960h)	34M	15.34 / 17.78
RBM-DBN ([21])	timit (8h)	≈ 34.2M	15.90 / 16.80
Ours (Randomly Init)	-	19.9M	13.20 / 15.10
Wav2vec ([7])	libri+wsj (1041h)	34M	12.90 / 14.70
Ours (Pretrained)	libri+wsj+ted (1248h)	19.9M	11.70 / 12.80
VQ-Wav2vec+BERT ([8])	libri (960h)	≈ 71.8M	9.64 / 11.64

Table from (Song et al., 2019)

Unsupervised speech representation learning with deep bidirectional transformer encoders

Predict the current frame through jointly conditioning on both past and future contexts (Mockingjay (Liu et al., 2020))

Masked acoustic modeling task (rand. mask 15% of input frames)7 Use multi-layer transformer encoders and multi-head self-attention Add a prediction head (2 layers of feed-forward network with

layer-norm) using last encoder layer as input

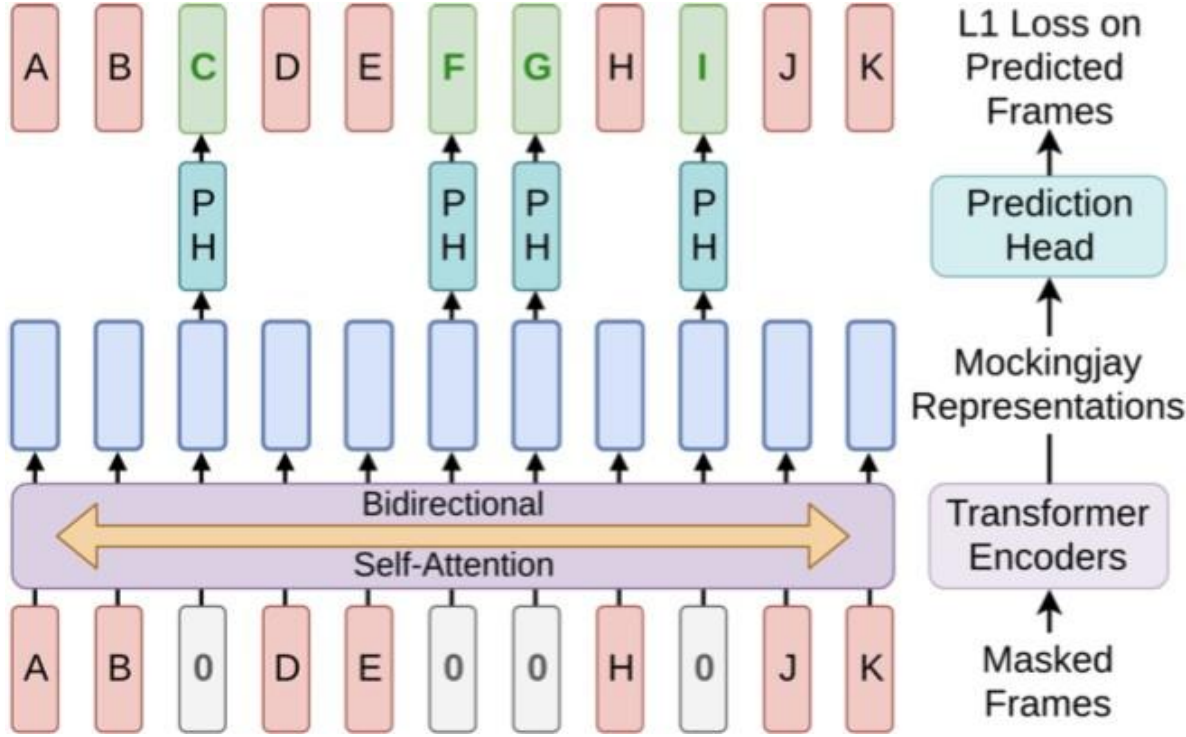


Figure from (Liu et al., 2020)

Unsupervised speech representation learning with deep bidirectional transformer encoders

Experiments on phoneme classification

With different amount of annotated data for training

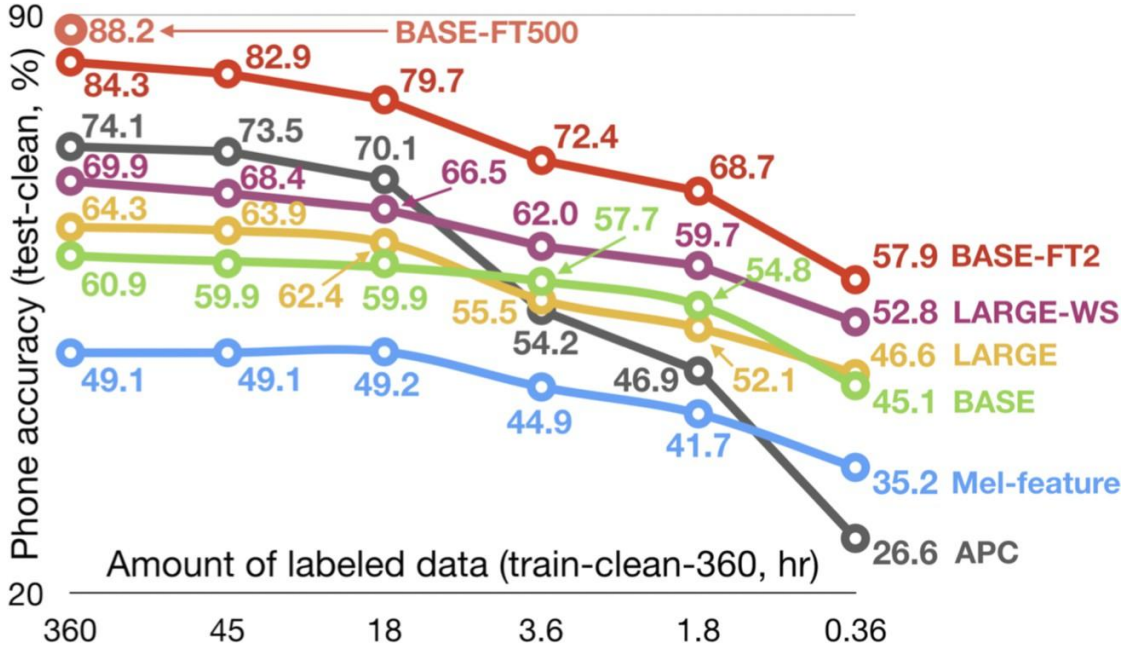


Figure from (Liu et al., 2020)

Unsupervised Pre-training of Bidirectional Speech Encoders via Masked Reconstruction

Pre-training speech representations via a masked reconstruction loss (Wang et al., 2020)

Masking in both frequency and time to encourage model to exploit spatio-temporal info

Elegant extension of data augmentation technique SpecAugment (Park et al., 2019)

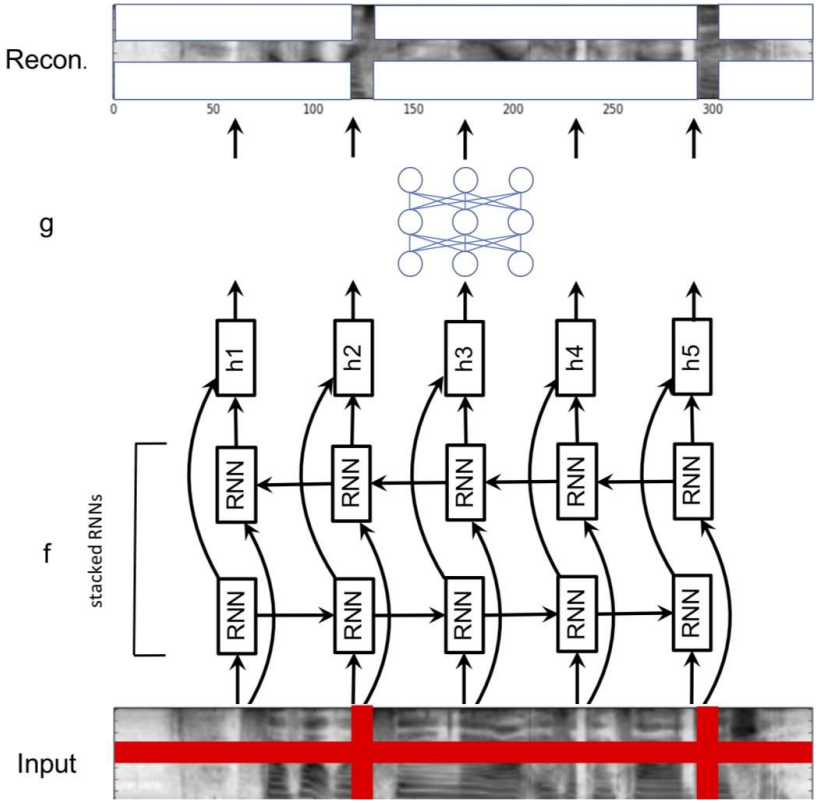


Figure from (Wang et al., 2020)

Unsupervised Pre-training of Bidirectional Speech Encoders via Masked Reconstruction

Table 3: Dev set %CERs of character-based systems pre-trained on LibriSpeech, and fine-tuned with different amounts of supervised data.

	Baseline	Pre-train *Libri.* w/o LIN	Pre-train *Libri.* w/ LIN
si84	15.23	14.02	**13.29**
+ SpecAug	12.98	12.26	**11.70**
si284	7.01	6.90	**6.48**
+ SpecAug	6.29	6.19	**5.61**

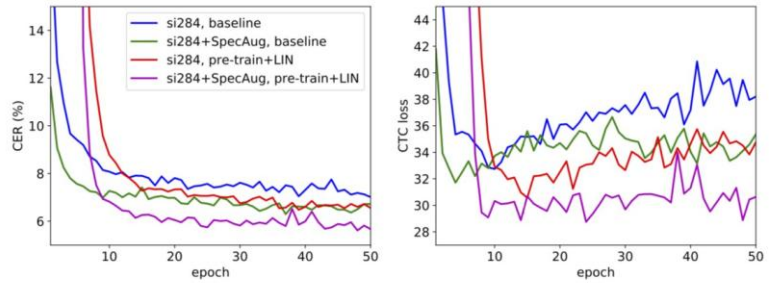


Fig. 2: Dev set learning curves (%CER and CTC loss) of different systems pre-trained on *LibriSpeech*. The first 5 epochs of fine-tuning update only the LIN and softmax layers.

From (Wang et al., 2020)

Wav2vec 2.0 (Baevski et al., 2020)

Encode speech with CNN layers and then mask spans of the resulting latent speech representations (cf masked LM)

Learn discrete speech units as latent representations8

Latent representations fed to a Transformer network to build contextualized representations

Model trained with a contrastive task (true latent to be distinguished from distractors)

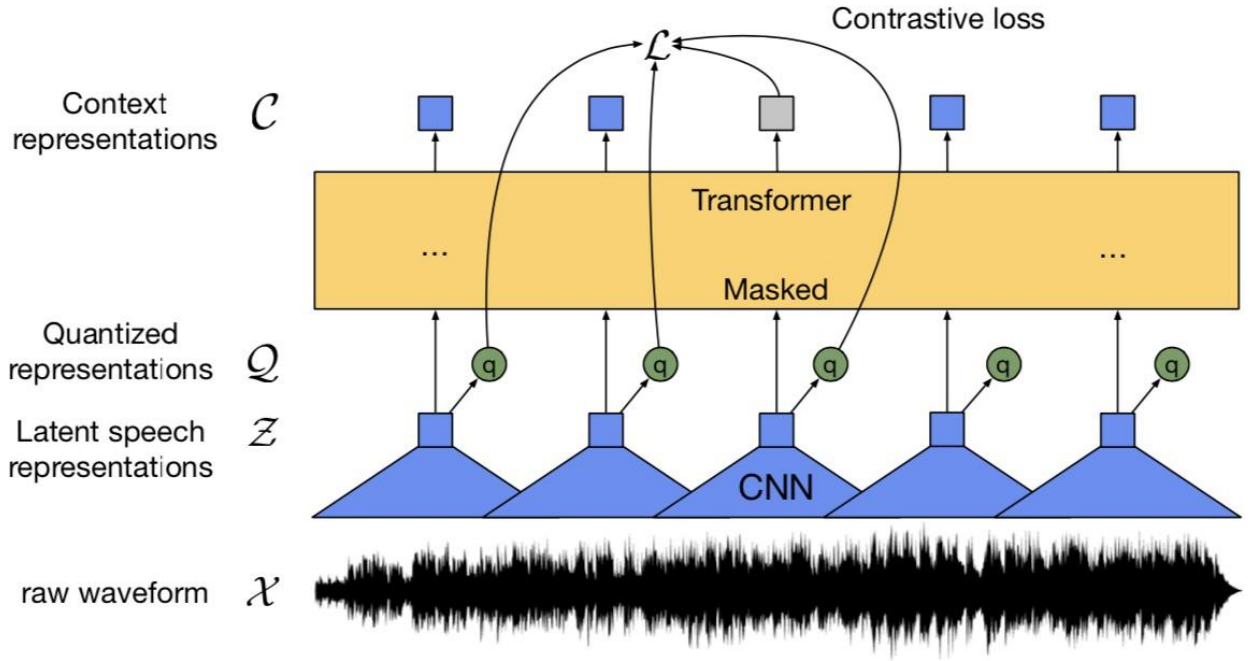


Figure from (Baevski et al., 2020)

HuBERT (Hsu et al., 2021)

Similar Conv+Transf encoder but

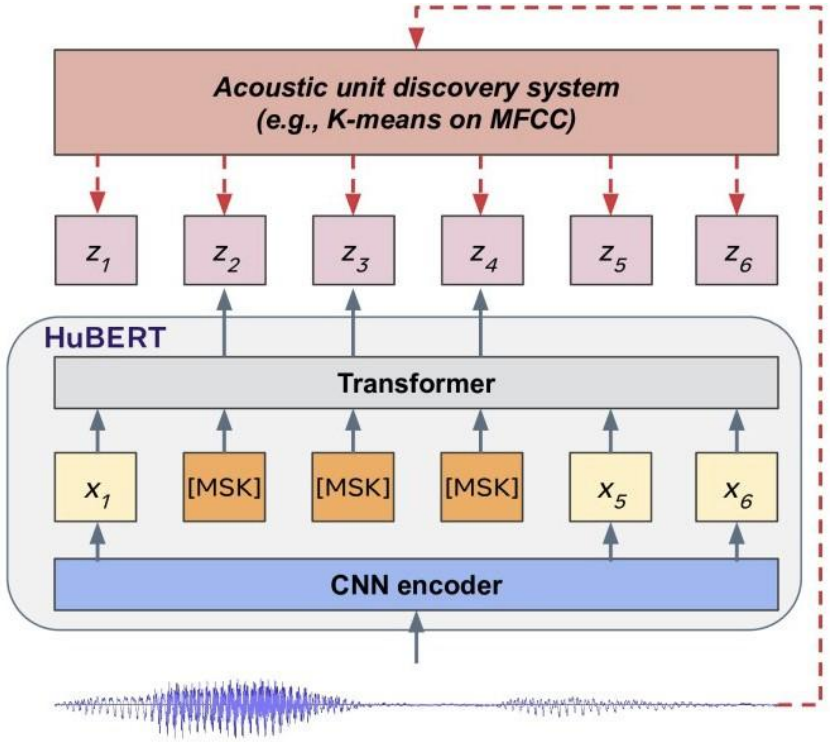
Uses cross-entropy loss (same as BERT) instead of contrastive loss

Discrete targets are built through a separate clustering process

Learnt discrete speech units are refined at each iteration (3 iterations for large models)

X-LARGE version of HuBERT as 1 billion parameters

Model recently outperformed SOTA techniques for speech recognition, generation, and compression



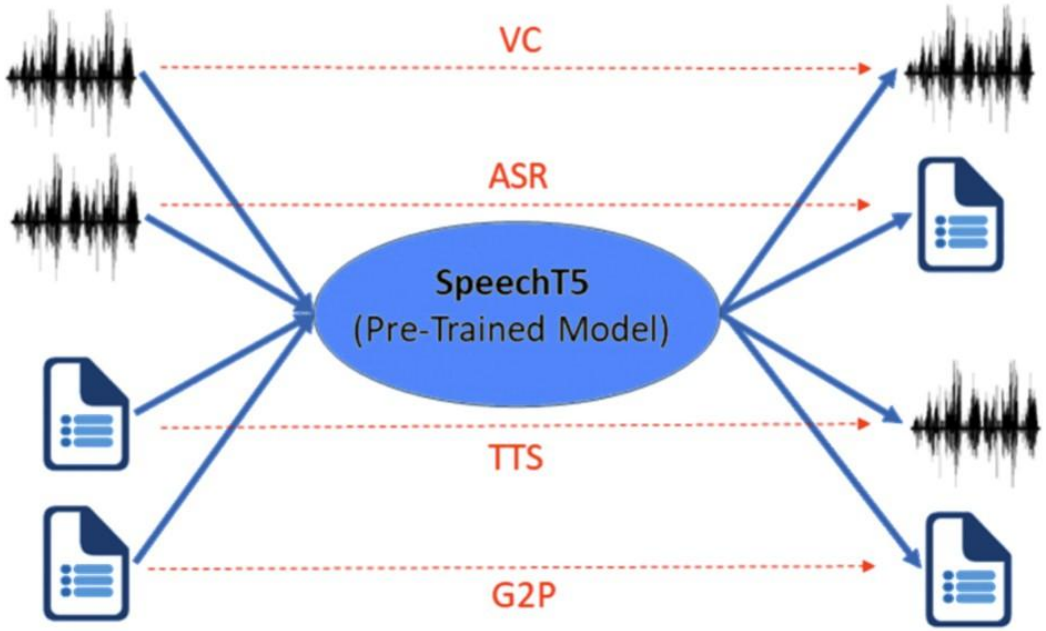
SpeechT5 (Ao et al., 2021)

A multimodal extension of transformer encoder-decoder models such as T5

Encode or decode both speech and text with a single model

Maps both acoustic and text information in a shared vector space

Used to initialize ASR (speech-to-text), TTS (text-to-speech), Voice Conversion (VC – speech-to-speech), etc.

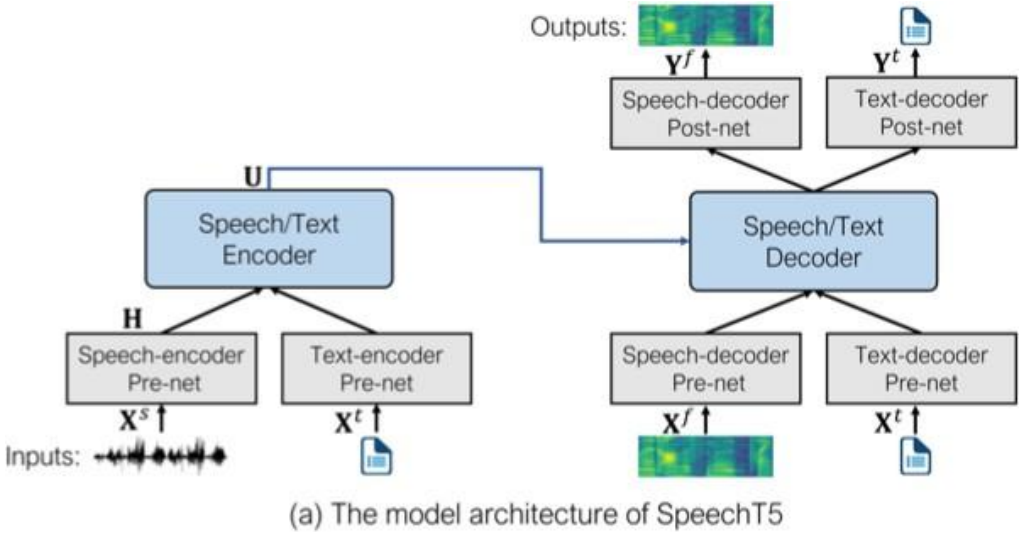


SpeechT5 (Ao et al., 2021)

A single transformer encoder/decoder backbone

Several modality-specific pre-post nets

- standard for text
- encoder pre-net for speech similar to the CNN blocks of wav2vec2.0
- decoder pre-net for speech is different (fully
- connected net + ReLU) as the model will output slices of filterbank features (no speech directly)
- a speaker embedding is concatenated to the
- output of the speech-decoder pre-net to support voice conversion and multi-speaker TTS



(a) The model architecture of SpeechT5

SpeechT5 (Ao et al., 2021)

A composite loss with multiple pre-training objectives

- a masked language modeling (MLM) loss on discrete latent speech representations (with HuBERT)
- a speech reconstruction L1 loss (in the continuous filterbank space)
- a cross-entropy loss specific to the prediction of the stop token
- a text denoising objective (with BART)
- a cross-modal objective to better
- align speech and text representations (unclear in the paper)

The final pre-training loss with unlabeled speech and text data can be formulated as

$$\mathcal{L} = \mathcal{L}^s_{mlm} + \mathcal{L}^s_1 + \mathcal{L}^s_{bce} + \mathcal{L}^t_{mle} + \gamma \mathcal{L}_d. \quad (6)$$

where γ is set to 0.1 during pre-training.

SpeechT5 (Ao et al., 2021)

Experiments on several downstream speech tasks (ASR, VC, TTS, speaker id.) show slightly better results than speech-only pre-training

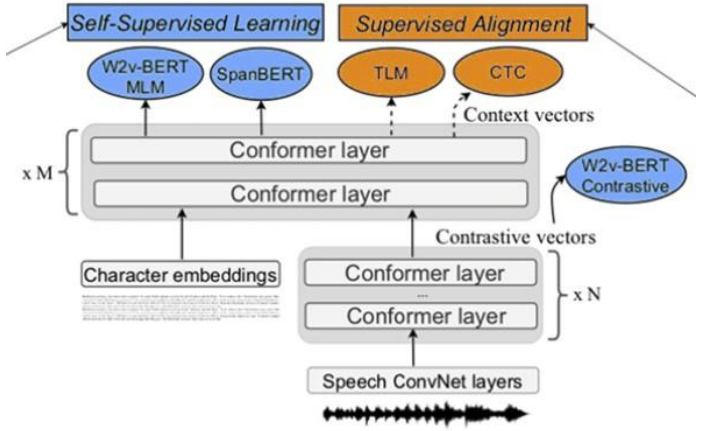
Model	LM	dev-clean	dev-other	test-clean	test-other
wav2vec 2.0 BASE (Baevski et al., 2020)	-	6.1	13.5	6.1	13.3
HuBERT BASE (Hsu et al., 2021) †	-	5.5	13.1	5.8	13.3
Baseline (w/o CTC)	-	5.8	12.3	6.2	12.3
Baseline	-	4.9	11.7	5.0	11.9
SpeechT5 (w/o CTC)	-	5.4	10.7	5.8	10.7
SpeechT5	-	**4.3**	**10.3**	**4.4**	**10.4**
DiscreteBERT (Baevski et al., 2019)	4-gram	4.0	10.9	4.5	12.1
wav2vec 2.0 BASE (Baevski et al., 2020)	4-gram	2.7	7.9	3.4	8.0
HuBERT BASE (Hsu et al., 2021)	4-gram	2.7	7.8	3.4	8.1
wav2vec 2.0 BASE (Baevski et al., 2020)	Transf.	2.2	6.3	2.6	6.3
Baseline	Transf.	2.3	6.3	2.5	6.3
SpeechT5	Transf.	**2.1**	**5.5**	**2.4**	**5.8**

Table 1: Results of ASR (speech to text) on the LibriSpeech dev and test sets when training on the 100 hours subset of LibriSpeech. † indicates that results are not reported in the corresponding paper and evaluated by ourselves.

mSLAM (Bapna et al., 2022)

Extension of SLAM (Bapna et al., 2021) architecture where speech and text unified in a common encoder model

- Use of convolution-augmented transformer (conformer) blocks, introduced earlier for ASR Gulati et al. (2020)
- Input is speech, text, or concatenated speech-text
- Speech-text pre-training is a mix of self-supervised learning objectives (rather similar to SpeechT5) and supervised cross-modal learning objectives (which leverage aligned speech-text pairs)

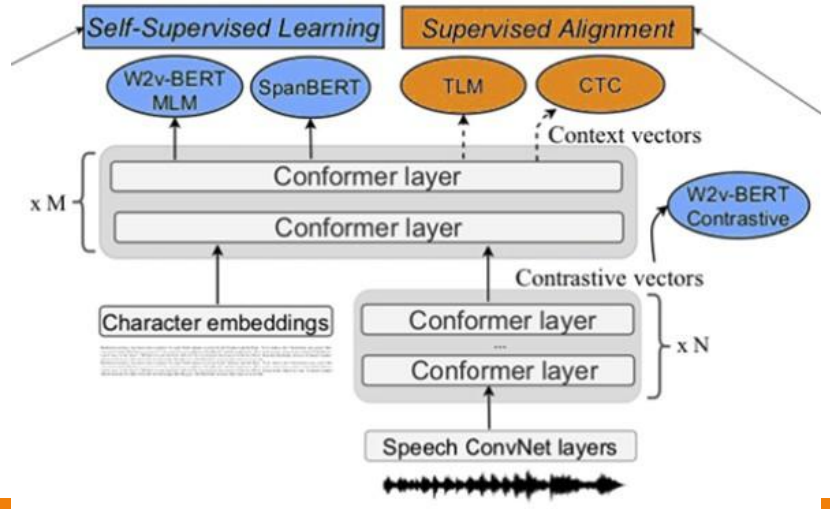


mSLAM (Bapna et al., 2022)

SSL learning objectives for speech (like HuBERT) and text (BERT)

Speech-text objectives

- translation language modeling (TLM): predicts masked text or speech spans from a concatenated speech-text input (to encourage use of cross-modal context)
- Connectionist Temporal Classification (CTC) loss is applied on the speech part of concatenated speech-text using character transcript as a target (ASR loss to learn better speech-text

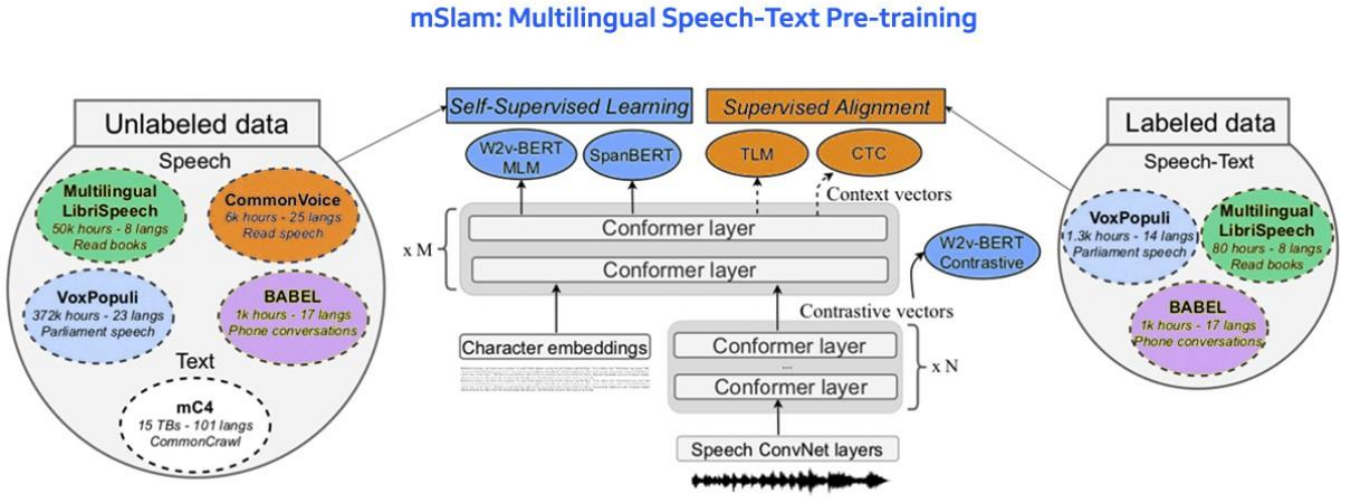


mSLAM (Bapna et al., 2022)

Massively multilingual (51 lang. speech; 101 lang. text), 2B param

Downstream task experiments

- ASR, speech translation, spoken lang. id., spoken intent classification and text classification
- Speech-text pre-training better than speech-only pre-training for multilingual ASR and translation



mSLAM (Bapna et al., 2022)

Zero shot cross-modal properties

Zero shot text translation from a fine-tuned speech translation model

The model has never seen src-txt/tgt-txt parallel data but it has seen src-speech/tgt-text + monolingual src-txt

... but a system fine-tuned only on text cannot translate speech

Table 6: Zero-shot Performance - CoVoST 2 translation results with X→Y indicating X as the fine tuning modality and Y as the testing modality: S=Speech, T=Text. CAE is our CTC zero-shot character auto-encoding probe.

	Hours	BLEU ↑			CER ↓
Lang	Paired	S→S	S→T	T→S	S→T CAE
ar	0	13.3	0.0	0.0	82.6
fa	0	6.2	0.0	0.0	80.0
ja	0	1.6	0.0	0.0	100.0
zh	0	8.7	0.0	0.0	100.0
cy	0	6.1	0.1	0.0	24.3
mn	0	0.5	0.1	0.0	78.4
id	0	3.9	5.1	0.0	10.4
lv	0	19.4	8.2	0.0	18.4
et	0	17.2	8.3	0.0	16.5
sv	0	33.1	15.2	0.0	13.9
ca	0	33.4	16.7	0.0	10.0
ru	0	41.7	21.9	0.0	85.9
sl	6	24.9	7.8	0.0	10.6
pt	10	34.2	17.2	0.0	9.0
nl	41	32.6	16.8	0.0	11.3
ta	63	0.3	0.0	0.0	91.2
tr	69	11.7	1.7	0.0	12.6
it	79	35.0	19.7	0.0	11.2
es	140	39.1	21.2	0.0	7.9
fr	179	36.7	20.0	0.0	9.4
de	197	32.7	16.8	0.0	8.3

Data2Vec (Baevski et al., 2022)

Data2vec is a framework for general self-supervised learning

- works for all three modalities: images, speech and text
- with a learning objective identical in each modality
- based on masked predictions and latent representation learning

Data2vec unifies the learning algorithm but still learns representations individually for each modality

- a.k.a Data2vec Text model (-BERT), Data2vec Speech model (-HuBERT), Data2vec Image model (-BEiT)

Data2Vec (Baevski et al., 2022)

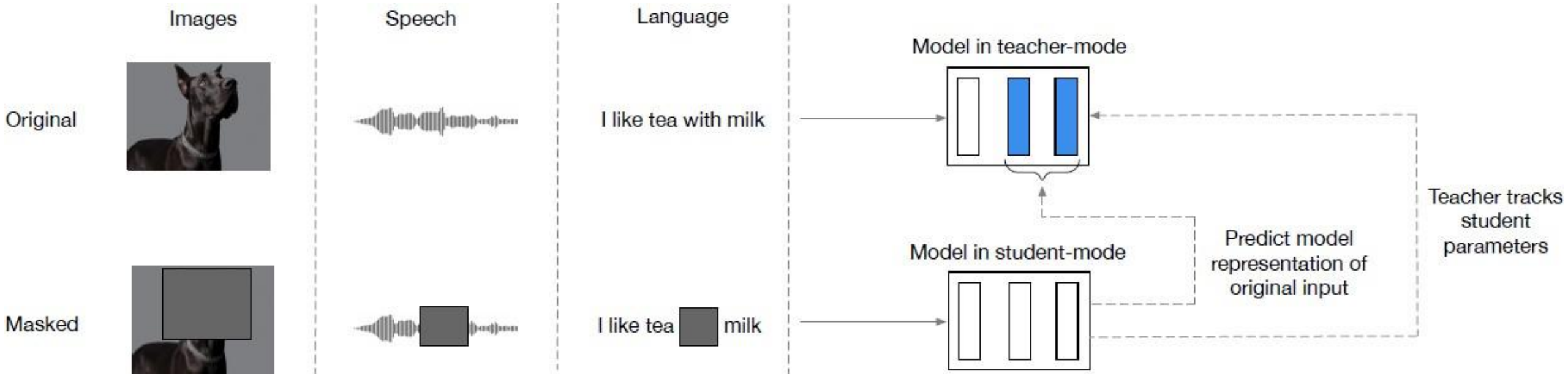
The only modality-specific elements in the architecture are the feature encoders and masking strategies on the input data

The model is an off-the-shelf Transformer network (Vaswani et al., 2017) working in a student/teacher mode

The learning task for the student is to predict masked latent representations of the teacher

- representations of the full input data are build to serve as targets in the learning task (teacher mode)
- we encode a masked version of the input sample with which we predict the full data representations (student mode).

Data2Vec (Baevski et al., 2022)



Data2Vec (Baevski et al., 2022)

The model is trained to predict the model representations of the original unmasked training sample based on an encoding of the masked sample.

only the masked time-steps are predicted

these targets are the average of the top K blocks of the teacher

given contextualized training targets y_t, we use a Smooth L1 loss to regress these targets:

$$\mathcal{L}(y_t, f_t(x)) = \begin{cases} \frac{1}{2}(y_t - f_t(x))^2/\beta & |y_t - f_t(x)| \leq \beta \\ (|y_t - f_t(x)| - \frac{1}{2}\beta) & \text{otherwise} \end{cases}$$

Data2Vec (Baevski et al., 2022)

Table 1. Computer vision: top-1 validation accuracy on ImageNet-1K with ViT-B (86M parameters) and ViT-L (307M parameters) models. Our results are based on training for 800 epochs while as several other well-performing models were trained for 1,600 epochs (MAE, MaskFeat).

	ViT-B	ViT-L
MoCo v3 (Chen et al., 2021b)	83.2	84.1
DINO (Caron et al., 2021)	82.8	-
BEiT (Bao et al., 2021)	83.2	85.2
MAE (He et al., 2021)	83.6	85.9
SimMIM (Xie et al., 2021)	83.8	-
MaskFeat (Wei et al., 2021)	84.0	85.7
data2vec	84.2	86.2

Vision Transformer Base and Large are tested for image classification on ImageNet-1k

Data2Vec (Baevski et al., 2022)

Speech: Data2vec is trained on the Librispeech dataset and evaluated on a ASR task

Table 2. Speech processing: word error rate on the Librispeech test-other test set when fine-tuning pre-trained models on the Libri-light low-resource labeled data setups (Kahn et al., 2020) of 10 min, 1 hour, 10 hours, the clean 100h subset of Librispeech and the full 960h of Librispeech. Models use the 960 hours of audio from Librispeech (LS-960) as unlabeled data. We indicate the language model used during decoding (LM). Results for all dev/test sets and other LMs can be found in the supplementary material (Table 5).

	Unlabeled data	LM	Amount of labeled data				
			10m	1h	10h	100h	960h
Base models							
wav2vec 2.0 (Baevski et al., 2020b)	LS-960	4-gram	15.6	11.3	9.5	8.0	6.1
HuBERT (Hsu et al., 2021)	LS-960	4-gram	15.3	11.3	9.4	8.1	-
WavLM (Chen et al., 2021a)	LS-960	4-gram	-	10.8	9.2	7.7	-
data2vec	LS-960	4-gram	12.3	9.1	8.1	6.8	5.5

Data2Vec (Baevski et al., 2022)

Text: Data2vec is trained on the same setting as BERT and tested on the GLUE benchmark

Table 3. Natural language processing: GLUE results on the development set for single-task fine-tuning of individual models. For MNLI we report accuracy on both the matched and unmatched dev sets, for MRPC and QQP, we report the unweighted average of accuracy and F1, for STS-B the unweighted average of Pearson and Spearman correlation, for CoLA we report Matthews correlation and for all other tasks we report accuracy. BERT Base results are from Wu et al. (2020) and our baseline is RoBERTa re-trained in a similar setup as BERT. We also report results with wav2vec 2.0 style masking of spans of four BPE tokens with no unmasked tokens or random targets.

	MNLI	QNLI	RTE	MRPC	QQP	STS-B	CoLA	SST	Avg.
Base models									
BERT (Devlin et al., 2019)	84.0/84.4	89.0	61.0	86.3	89.1	89.5	57.3	93.0	80.7
Baseline (Liu et al., 2019)	84.1/83.9	90.4	69.3	89.0	89.3	88.9	56.8	92.3	82.5
data2vec	83.2/83.0	90.9	67.0	90.2	89.1	87.2	62.2	91.8	82.7
+ wav2vec 2.0 masking	82.8/83.4	91.1	69.9	90.0	89.0	87.7	60.3	92.4	82.9

Data2Vec (Baevski et al., 2022)

The method can be seen as a kind of generalization of the concept of masked language/speech/pixels modelling. It's a masked latent representation pretraining task.

Experimental results show data2vec to be effective in all three modalities

To explain these results the authors say that the fact the "Target representations being continuous and contextualized makes them richer than a fixed set of targets"

Potential improvements and limitations: modality specific feature extractors amd masking strategies, potential representation collapse

Using LLMs for ASR

GPT-2 for rescoring N-best ASR hypotheses (Sun et al., 2023)

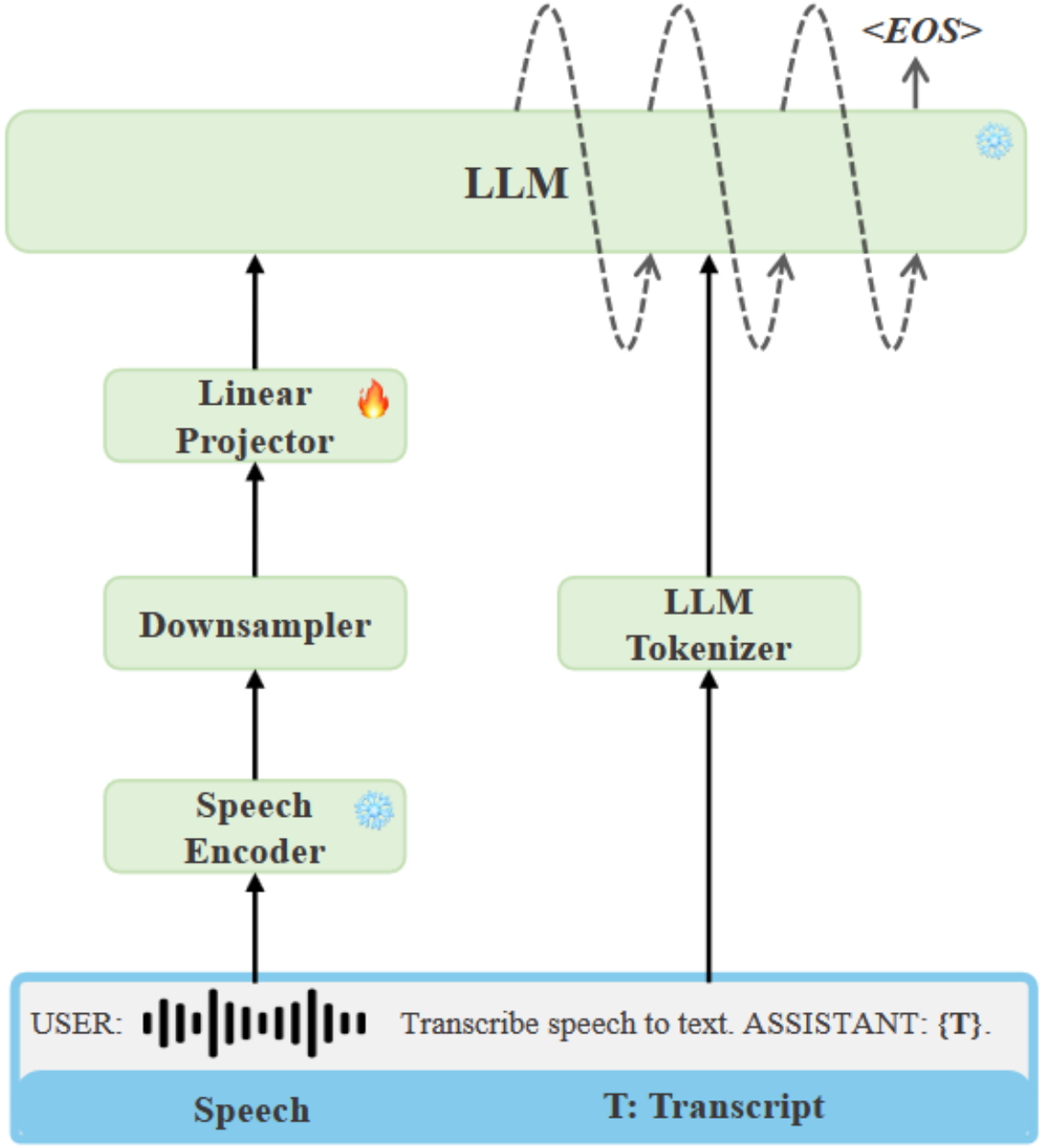
Deep integration between speech models and LLMs

- Architecture of prepending continuous audio embed- dings to the text embeddings before feeding to a decoder-only LLM
- SpeechGPT (Zhang et al., 2023)
- Speech-LLaMA (Wu et al., 2023; Fathullah et al., 2023)

Bridging speech input and instruction following using LoRA (Chen et al., 2023)

SLAM-ASR

Integrating speech as multimodal input and instructing model to do ASR



Ma, Z., Yang, G., Yang, Y., Gao, Z., Wang, J., Du, Z., Yu, F., Chen, Q., Zheng, S., Zhang, S., & Chen, X. (2024). An Embarrassingly Simple Approach for LLM with Strong ASR Capacity. https://arxiv.org/abs/2402.08846

Conclusion

Training such models requires access to powerful computing platforms

Simpler and more efficient pre-training approaches needed

Standardization in the evaluation process also needed (need for a multimodal and multilingual GLUE)

Only scratched the surface of zero-shot capabilities of these models (transfer from text to speech tasks)

More research needed on the decoder side (especially to generate expressive speech with adequate prosody)

On par with human transcription?

Read Speech			
Test set	DS1	DS2	Human
WSJ eval'92	4.94	3.60	5.03
WSJ eval'93	6.94	4.98	8.08
LibriSpeech test-clean	7.89	5.33	5.83
LibriSpeech test-other	21.74	13.25	12.69

Comparison of WER for two speech systems and human level performance on **read** speech (from (Amodei et al., 2016)

Accented Speech			
Test set	DS1	DS2	Human
VoxForge American-Canadian	15.01	7.55	4.85
VoxForge Commonwealth	28.46	13.56	8.15
VoxForge European	31.20	17.55	12.76
VoxForge Indian	45.35	22.44	22.15

Comparison of WER for two speech systems and human level performance on **accented** speech (from (Amodei et al., 2016)

On par with human transcription?

Noisy Speech			
Test set	DS1	DS2	Human
CHiME eval clean	6.30	3.34	3.46
CHiME eval real	67.94	21.79	11.84
CHiME eval sim	80.27	45.05	31.33

Comparison of WER for two speech systems and human level performance on **noisy** speech (from (Amodei et al., 2016)

Language coverage

Google addresses (only) 100 languages (ASR)

Language technology issues: 300 languages (95 % population) Language coverage / revitalisation / documentation issues: > 6000 languages !

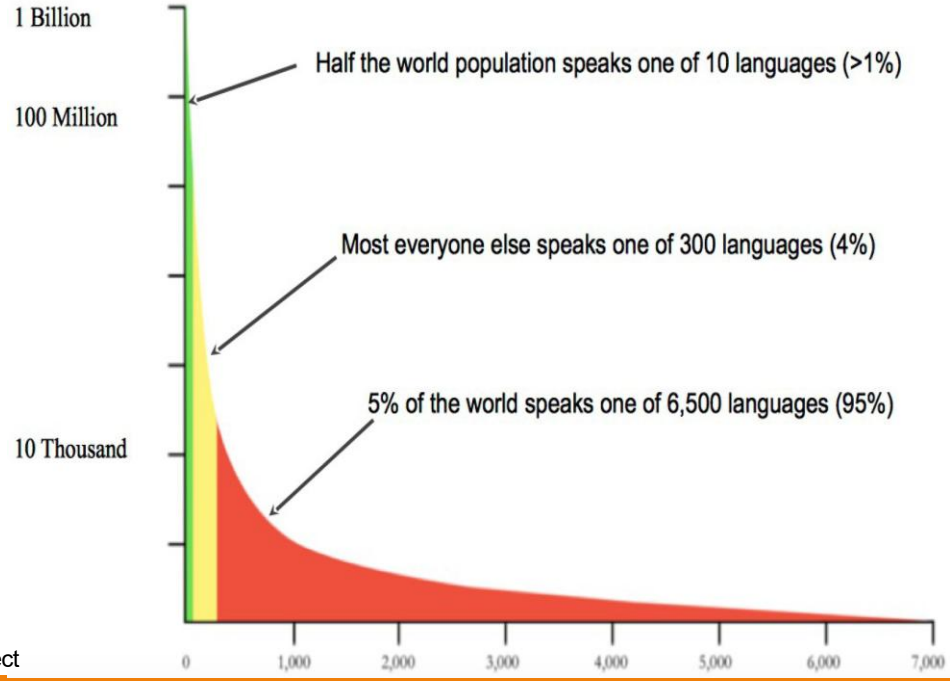


Figure: from Laura Welcher - Big Data for Small Languages The Rosetta Project

Low resource ASR

Rapid development of ASR for new languages In low resource conditions

For languages poorly described ex: DARPA Babel program

Language	ML-24		CTC		Attention	
	XE	ST	XE	ST	XE	ST
Pashto	54.5	51.5	51.5	49.5	50.6	50.1
Guarani	52.1	48.5	--	47.8	47.3	46.3
Igbo	63.4	60.2	61.4	59.2	59.6	58.8
Amharic	49.3	44.5	--	44.6	45.9	44.5
Mongolian	59.1	55.1	59.6	53.0	54.5	53.5
Javanese	60.1	55.3	57.5	54.1	55.4	54.2
Dholuo	43.7	40.5	--	40.0	40.7	39.9
Georgian	45.0	40.8	44.3	40.6	45.6	43.9

Performance of end-2-end models on Babel languages

from Bhuvana Ramabhadran's presentation at Interspeech 2018

Zero resource ASR

In an unknown language, from unannotated raw speech, discover:

- Invariant subword units (phone units?)
- Words/terms (lexicon/semantic units?)

Technological challenge

- Can we build useful speech technologies without any textual resources?
- Unsupervised ASR / autonomous systems

Scientific challenge

- Can we build algorithms that learn languages like infants do?
- Can we build algorithms that extract meaningful units from unknown languages?

The zero resource challenge: http://zerospeech.com (Dunbar et al., 2017)

Multilingual ASR

1 system - N languages

- In end-to-end ASR, acoustic, pronunciation and language model are integrated into a single neural network
- Makes them very suitable for truly multilingual ASR
- First attempts using hybrid CTC/attention ASR approaches (Watanabe et al., 2017)
- Similar in spirit to multilingual NMT (Johnson et al., 2016)
- Further propositions using a transformer network (Zhou et al., 2018)

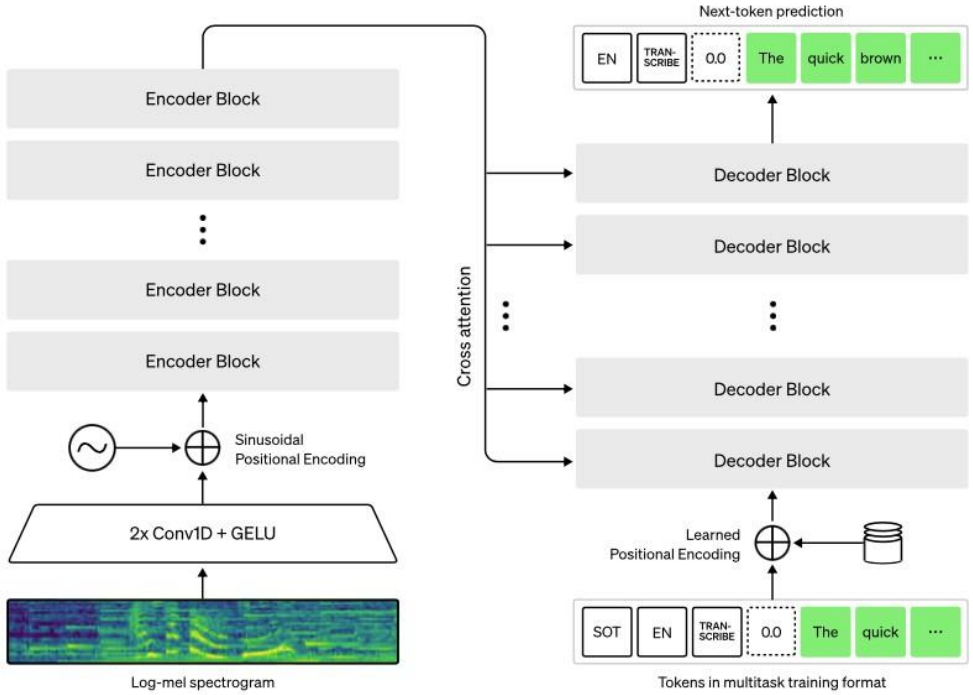
Whisper

A massively multilingual ASR system based on weakly-supervised learning trained on 680,000 hours of multilingual and multitask supervised data collected from the web

use of such a large and diverse dataset leads to improved robustness to accents, background noise and technical language

enables transcription in multiple languages, as well as translation from those languages into English

whisper architecture is a simple end-to-end approach, implemented as an encoder-decoder Transformer



Other ASR challenges

Can we leverage multiple sensors to design noise robust approaches? (Barker et al., 2017)

What do NNs learn? (Belinkov and Glass, 2017)

How can we can exploit adversarial examples to improve overall robustness?

Can we analyze (and deal with) biases between genders, dialects, regional accents?

How to deal with code-switching phenomena?

References I

Alec Radford, Jong Wook Kim, T. X. G. B. C. M. I. S. (2022). Robust speech recognition via large-scale weak supervision.

Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., Elsen, E., Engel, J., Fan, L., Fougner, C., Hannun,

A. Y., Jun, B., Han, T., LeGresley, P., Li, X., Lin, L., Narang, S., Ng, A. Y., Ozair, S.,

Prenger, R., Qian, S., Raiman, J., Satheesh, S., Seetapun, D., Sengupta, S., Wang, C., Wang, Y., Wang, Z., Xiao, B., Xie, Y., Yogatama, D., Zhan, J., and Zhu, Z. (2016). Deep speech 2 : End-to-end speech recognition in english and mandarin. In Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016, pages 173–182.

Ao, J., Wang, R., Zhou, L., Liu, S., Ren, S., Wu, Y., Ko, T., Li, Q., Zhang, Y., Wei, Z., et al. (2021). Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. arXiv preprint arXiv:2110.07205.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., and Auli, M. (2021). Xls-r: Self-supervised cross-lingual speech representation learning at scale. arXiv, abs/2111.09296.

Baevski, A., Auli, M., and Mohamed, A. (2019). Effectiveness of self-supervised pre-training for speech recognition.

References II

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Larochelle, H., Ranzato, M., Hadsell,

R., Balcan, M., and Lin, H., editors, Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473.

Bapna, A., Cherry, C., Zhang, Y., Jia, Y., Johnson, M., Cheng, Y., Khanuja, S., Riesa, J., and Conneau, A. (2022). mslam: Massively multilingual joint pre-training for speech and text. CoRR, abs/2202.01374.

Bapna, A., Chung, Y., Wu, N., Gulati, A., Jia, Y., Clark, J. H., Johnson, M., Riesa, J., Conneau, A., and Zhang, Y. (2021). SLAM: A unified encoder for speech and language modeling via speech-text joint pre-training. CoRR, abs/2110.10329.

Barker, J., Marxer, R., Vincent, E., and Watanabe, S. (2017). Multi-microphone speech recognition in everyday environments. Computer Speech & Language, 46:386–387.

Belinkov, Y. and Glass, J. R. (2017). Analyzing hidden representations in end-to-end automatic speech recognition systems. In NIPS, pages 2438–2448.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. J. Mach. Learn. Res., 3:1137–1155.

References III

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations.

Chen, Z., Huang, H., Andrusenko, A., Hrinchuk, O., Puvvada, K. C., Li, J., Ghosh, S., Balam, J., and Ginsburg, B. (2023). Salm: Speech-augmented language model with in-context learning for speech recognition and translation.

Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015). Attention-based models for speech recognition. CoRR, abs/1506.07503.

Chung, Y., Hsu, W., Tang, H., and Glass, J. R. (2019). An unsupervised autoregressive model for speech representation learning. CoRR, abs/1904.03240.

Chung, Y.-A. and Glass, J. (2020). Improved speech representations with multi-target autoregressive predictive coding.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805.

Dunbar, E., Cao, X., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., Anguera, X., and Dupoux, E. (2017). The zero resource speech challenge 2017. CoRR, abs/1712.04313.

Fathullah, Y., Wu, C., Lakomkin, E., Jia, J., Shangguan, Y., Li, K., Guo, J., Xiong, W., Mahadeokar, J., Kalinli, O., Fuegen, C., and Seltzer, M. (2023). Prompting large language models with speech recognition abilities.

References IV

Graves, A., Fern´andez, S., Gomez, F. J., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In ICML, volume 148 of ACM International Conference Proceeding Series, pages 369–376. ACM.

Gulati, A., Qin, J., Chiu, C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. In Meng, H., Xu, B., and Zheng, T. F., editors, Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020, pages 5036–5040. ISCA.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Vi´egas, F. B., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2016). Google's multilingual neural machine translation system: Enabling zero-shot translation. CoRR, abs/1611.04558.

Kahn, J., Rivi`ere, M., Zheng, W., Kharitonov, E., Xu, Q., Mazar´e, P.-E., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., Likhomanenko, T., Synnaeve, G., Joulin, A., Mohamed, A., and Dupoux, E. (2019). Libri-light: A benchmark for asr with limited or no supervision.

Le, H. S., Oparin, I., Allauzen, A., Gauvain, J., and Yvon, F. (2011). Structured output layer neural network language model. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, May 22-27, 2011, Prague Congress Center, Prague, Czech Republic, pages 5524–5527.

References V

Liu, A. T., Yang, S.-w., Chi, P.-H., Hsu, P.-c., and Lee, H.-y. (2020). Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Mikolov, T., Karafi´at, M., Burget, L., C˘ernock´y, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In Interspeech.

Mnih, A. and Hinton, G. (2008). A scalable hierarchical distributed language model. In In NIPS. Morin, F. and Bengio, Y. (2005). Hierarchical probabilistic neural network language model. In AISTATS'05, pages 246–252.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In ICASSP, pages 5206–5210. IEEE.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (2019).

Specaugment: A simple data augmentation method for automatic speech recognition. Interspeech 2019.

Pascual, S., Ravanelli, M., Serr`a, J., Bonafonte, A., and Bengio, Y. (2019). Learning problem-agnostic speech representations from multiple self-supervised tasks. CoRR, abs/1904.03416.

References VI

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Ravanelli, M. and Bengio, Y. (2018). Speaker recognition from raw waveform with sincnet.

Ravanelli, M., Zhong, J., Pascual, S., Swietojanski, P., Monteiro, J., Trmal, J., and Bengio, Y. (2020). Multi-task self-supervised learning for robust speech recognition.

Renals, S., Morgan, N., Bourlard, H., Cohen, M., and Franco, H. (1994). Connectionist probability estimators in HMM speech recognition. IEEE Trans. Speech and Audio Processing, 2(1):161–174.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019a). wav2vec: Unsupervised Pre-Training for Speech Recognition. In Proc. Interspeech 2019, pages 3465–3469.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019b). wav2vec: Unsupervised pre-training for speech recognition. CoRR, abs/1904.05862.

Schwenk, H. (2007). Continuous space language models. Computer Speech & Language, 21(3):492–518.

Song, X., Wang, G., Wu, Z., Huang, Y., Su, D., Yu, D., and Meng, H. (2019). Speech-xlnet: Unsupervised acoustic model pretraining for self-attention networks.

References VII

Sun, G., Zheng, X., Zhang, C., and Woodland, P. C. (2023). Can contextual biasing remain effective with whisper and gpt-2?

van den Oord, A., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. CoRR, abs/1807.03748.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. CoRR, abs/1706.03762.

Wang, W., Tang, Q., and Livescu, K. (2020). Unsupervised pre-training of bidirectional speech encoders via masked reconstruction.

Watanabe, S., Hori, T., and Hershey, J. R. (2017). Language independent end-to-end architecture for joint language identification and speech recognition. 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 265–271.

Wu, J., Gaur, Y., Chen, Z., Zhou, L., Zhu, Y., Wang, T., Li, J., Liu, S., Ren, B., Liu, L., and Wu, Y. (2023). On decoder-only architecture for speech-to-text and large language model integration.

Zhang, D., Li, S., Zhang, X., Zhan, J., Wang, P., Zhou, Y., and Qiu, X. (2023). Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities.

Zhou, S., Xu, S., and Xu, B. (2018). Multilingual end-to-end speech recognition with A single transformer on low-resource languages. CoRR, abs/1806.05059.