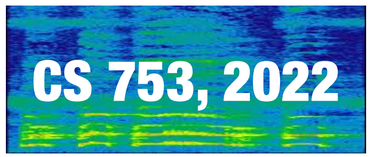
Speech Representations



Instructor: Preethi Jyothi, IITB

wav2vec

- Algorithm that uses raw audio to learn speech representations (“**self-supervised**” approach)

Image from: https://arxiv.org/pdf/1904.05862.pdf

wav2vec

- Algorithm that uses raw audio to learn speech representations ("**self-supervised**" approach)

  - Encoder network embeds raw audio into a latent representation ($f : \mathcal{X} \rightarrow \mathcal{Z}$) and a context network combines multiple encoded representations into a contextualised embedding ($g : \mathcal{Z} \rightarrow \mathcal{C}$)

Image from: https://arxiv.org/pdf/1904.05862.pdf

wav2vec

- Algorithm that uses raw audio to learn speech representations ("**self-supervised**" approach)

  - Encoder network embeds raw audio into a latent representation ($f : \mathcal{X} \rightarrow \mathcal{Z}$) and a context network combines multiple encoded representations into a contextualised embedding ($g : \mathcal{Z} \rightarrow \mathcal{C}$)

  - Train the model to minimize the following contrastive loss:

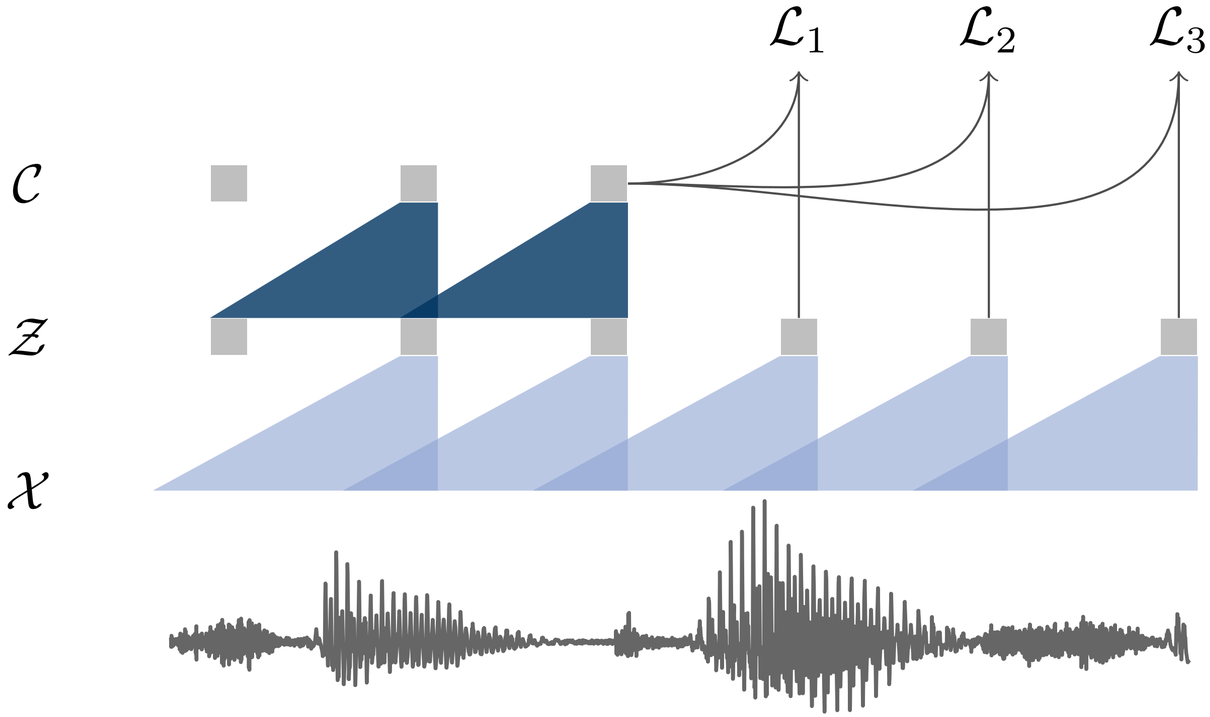


Image from: https://arxiv.org/pdf/1904.05862.pdf

wav2vec

- Algorithm that uses raw audio to learn speech representations (“**self-supervised**” approach)

  - Encoder network embeds raw audio into a latent representation ($f : \mathcal{X} \rightarrow \mathcal{Z}$) and a context network combines multiple encoded representations into a contextualised embedding ($g : \mathcal{Z} \rightarrow \mathcal{C}$)

  - Train the model to minimize the following contrastive loss:

$$\mathcal{L}_k = -\sum_{i=1}^{T-k} \log \frac{\exp(\mathrm{sim}(c_i, z_{i+k}))}{\sum_{\tilde{z}} \exp(\mathrm{sim}(c_i, \tilde{z}))}$$

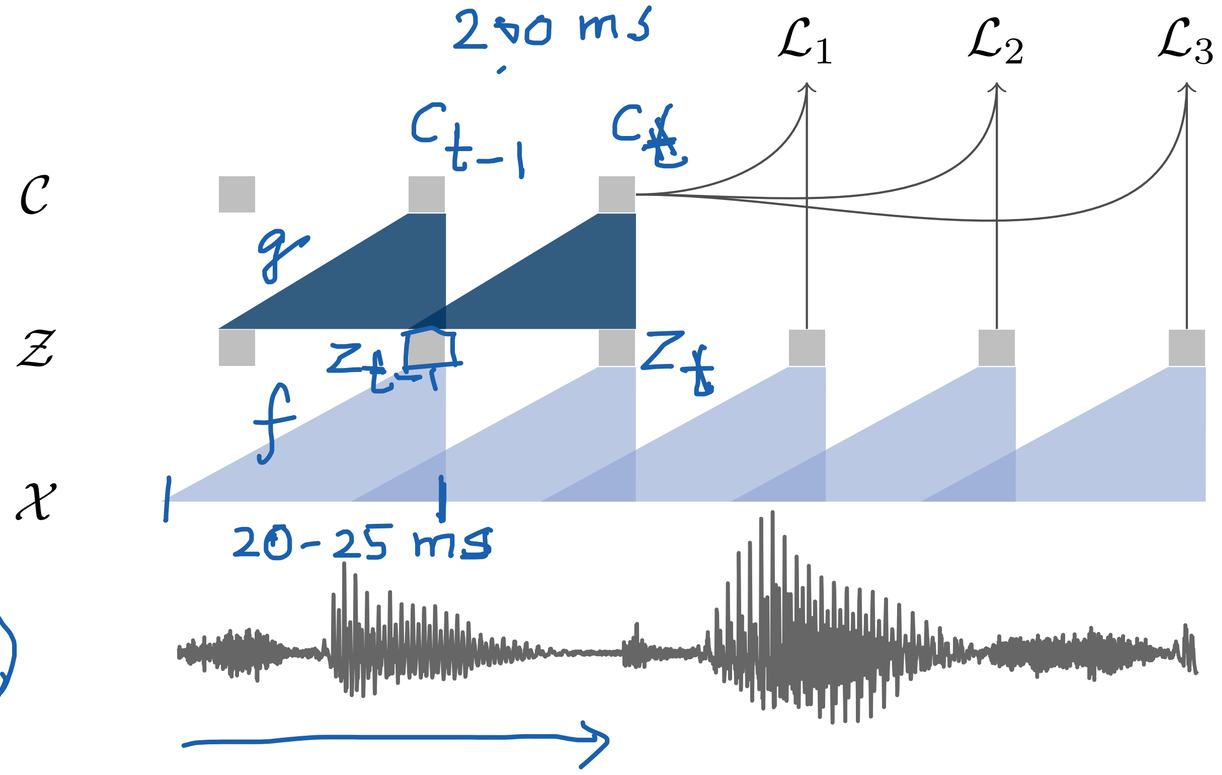
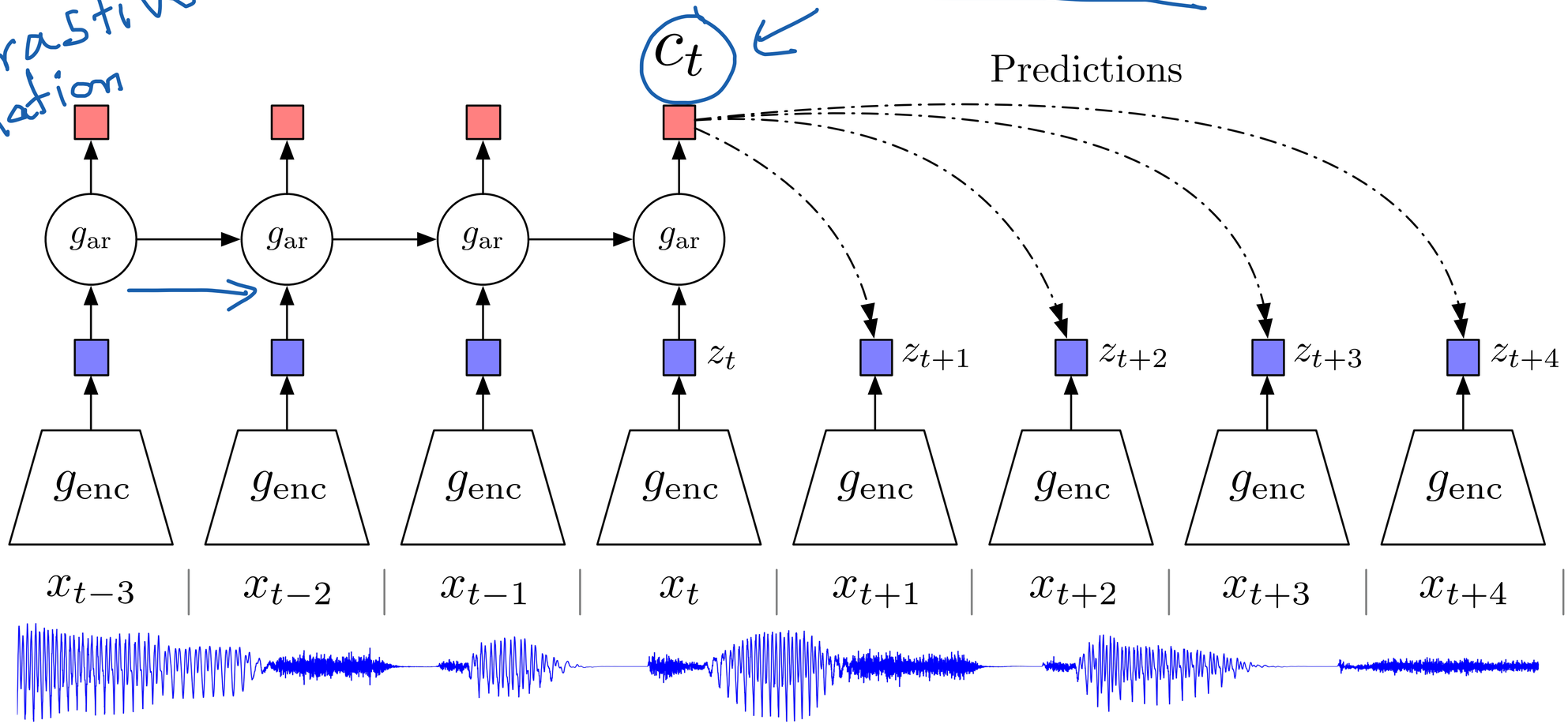


Image from: https://arxiv.org/pdf/1904.05862.pdf

Contrastive Predictive Coding

Noise Contrastive Estimation



$$\mathcal{L}_k = -\sum_{i=1}^{T-k} \log \frac{\exp(\text{sim}(c_i, z_{i+k}))}{\sum_{\tilde{z}} \exp(\text{sim}(c_i, \tilde{z}))}$$

Contrastive Loss

Image/table from: https://arxiv.org/pdf/1807.03748v2.pdf (van den Oord et al., 2019)

wav2vec

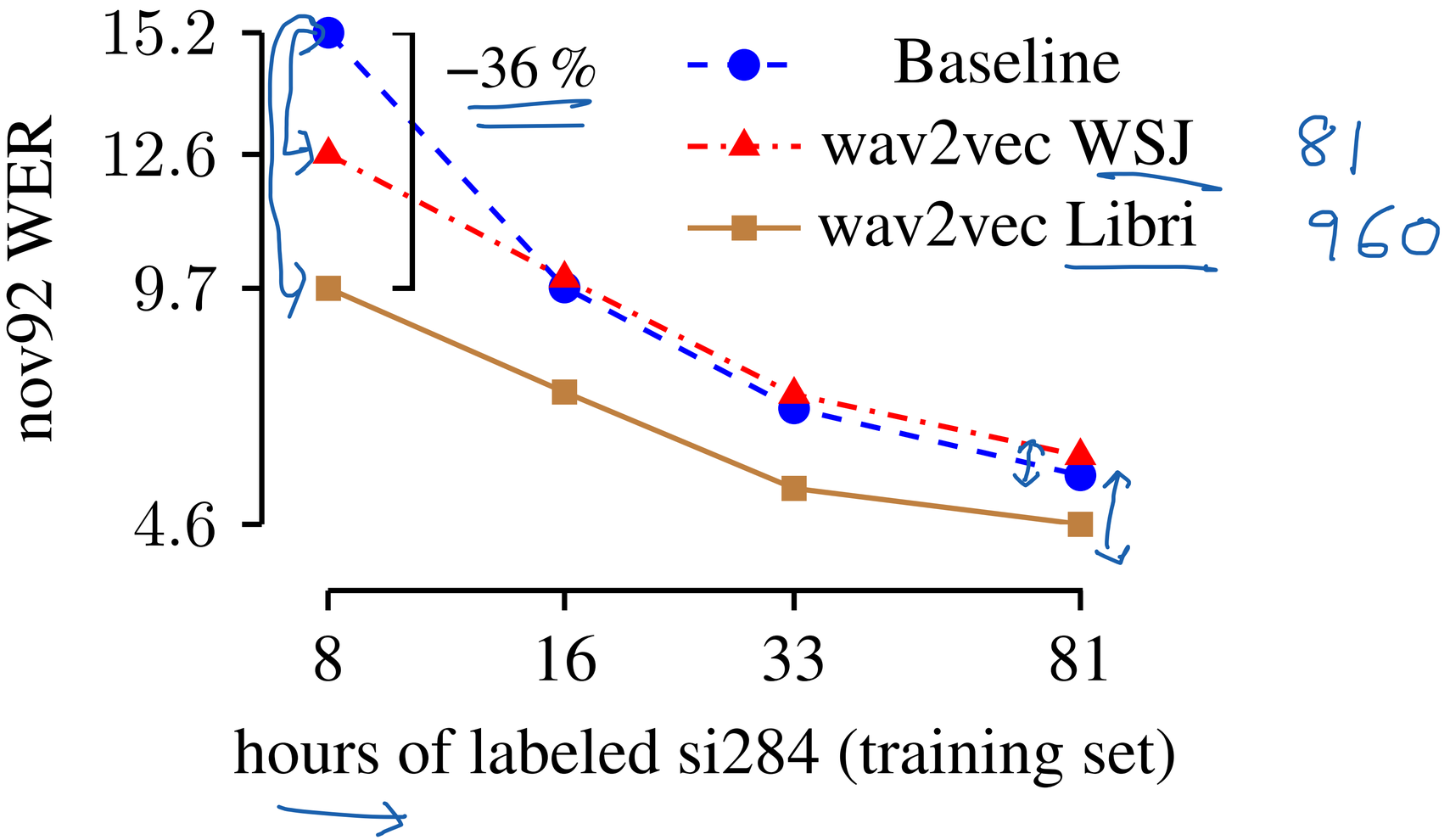


Image from: https://arxiv.org/pdf/1904.05862.pdf

wav2vec 2.0: Learning Speech Representations from Raw Audio

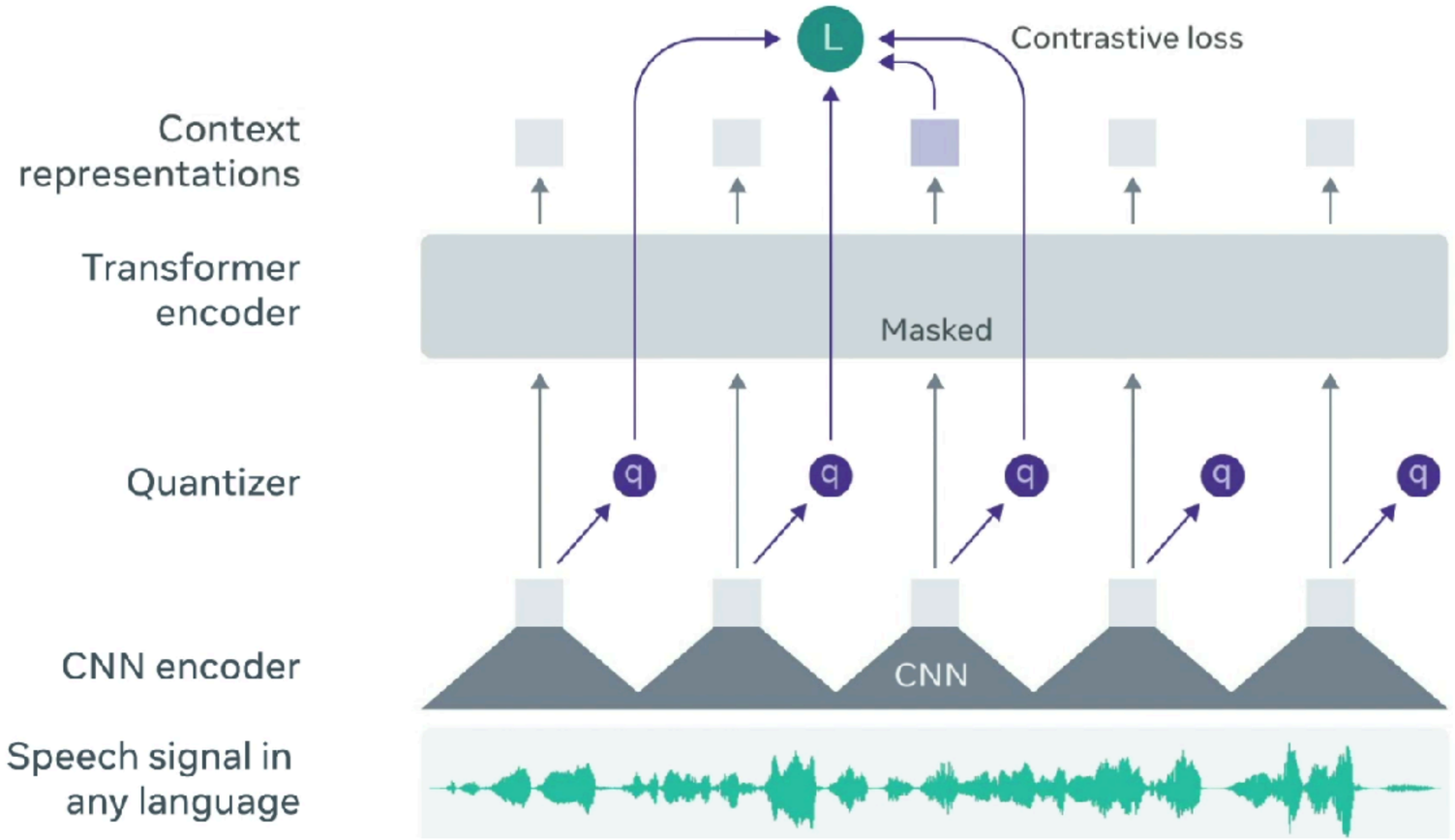


Image from:https://ai.facebook.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/

[1]: https://arxiv.org/abs/1810.04805

wav2vec 2.0: Learning Speech Representations from Raw Audio

- Similar to wav2vec. Outputs from the encoder are further quantized.

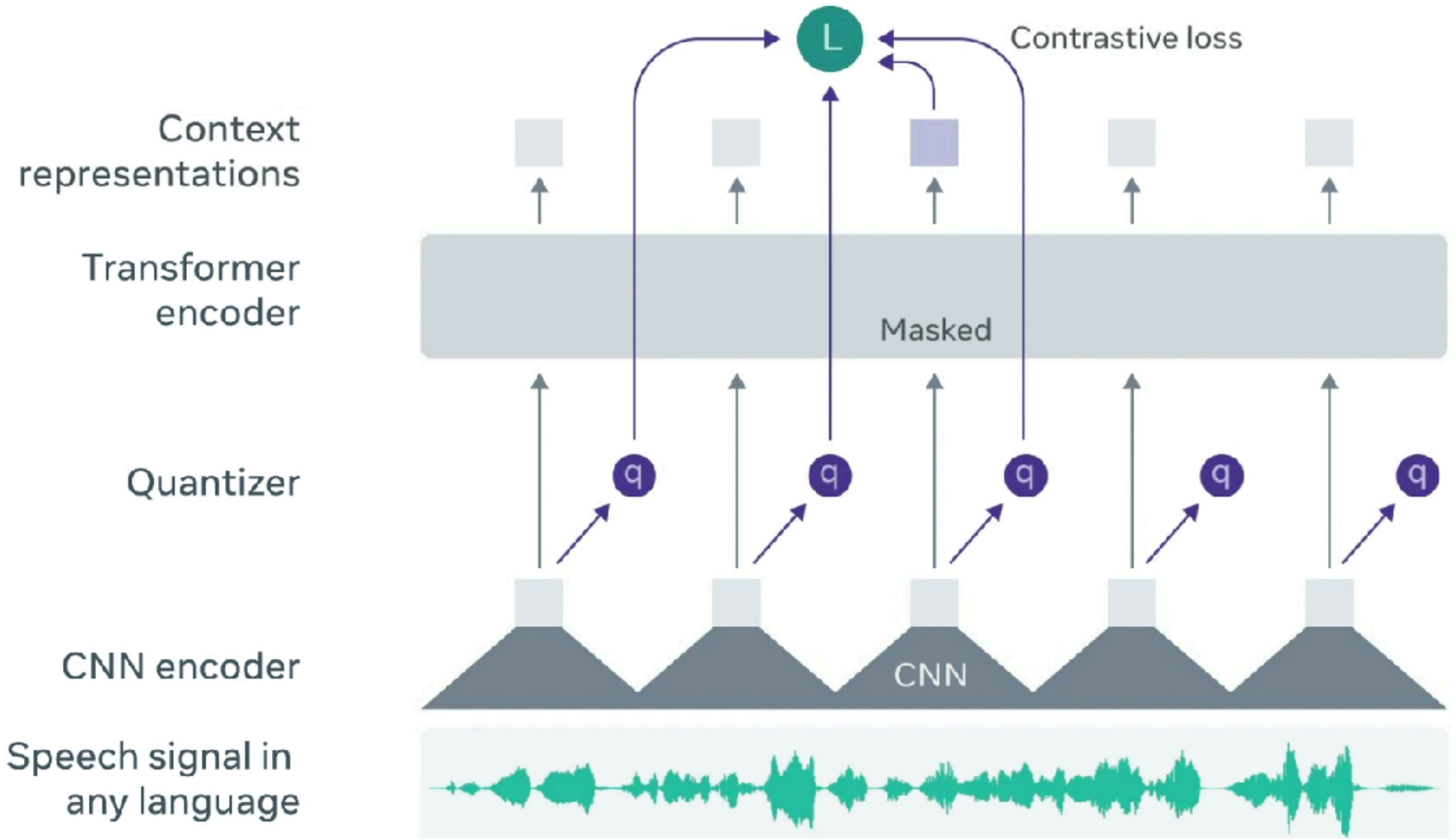


Image from:https://ai.facebook.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/

[1]: https://arxiv.org/abs/1810.04805

wav2vec 2.0: Learning Speech Representations from Raw Audio

- Similar to wav2vec. Outputs from the encoder are further quantized.
- Masks spans of speech representations (as in masked language modelling for BERT [1])

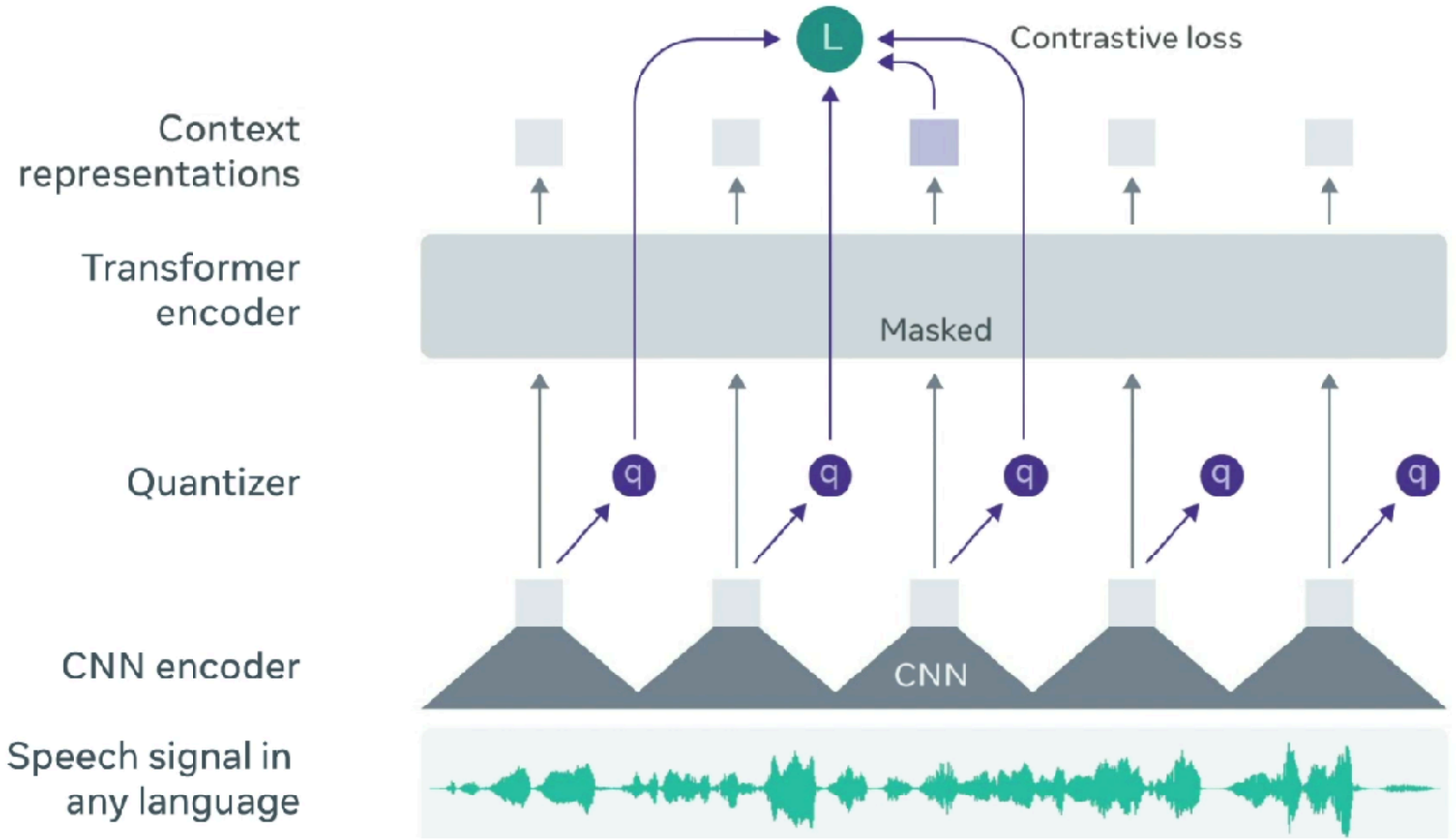


Image from:https://ai.facebook.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/

[1]: https://arxiv.org/abs/1810.04805

wav2vec 2.0: Learning Speech Representations from Raw Audio

vq-wav2vec

- Similar to wav2vec. Outputs from the encoder are further quantized.
- Masks spans of speech representations (as in masked language modelling for BERT [1])
- Training objective is to recover the masked representations among a set of distractors.

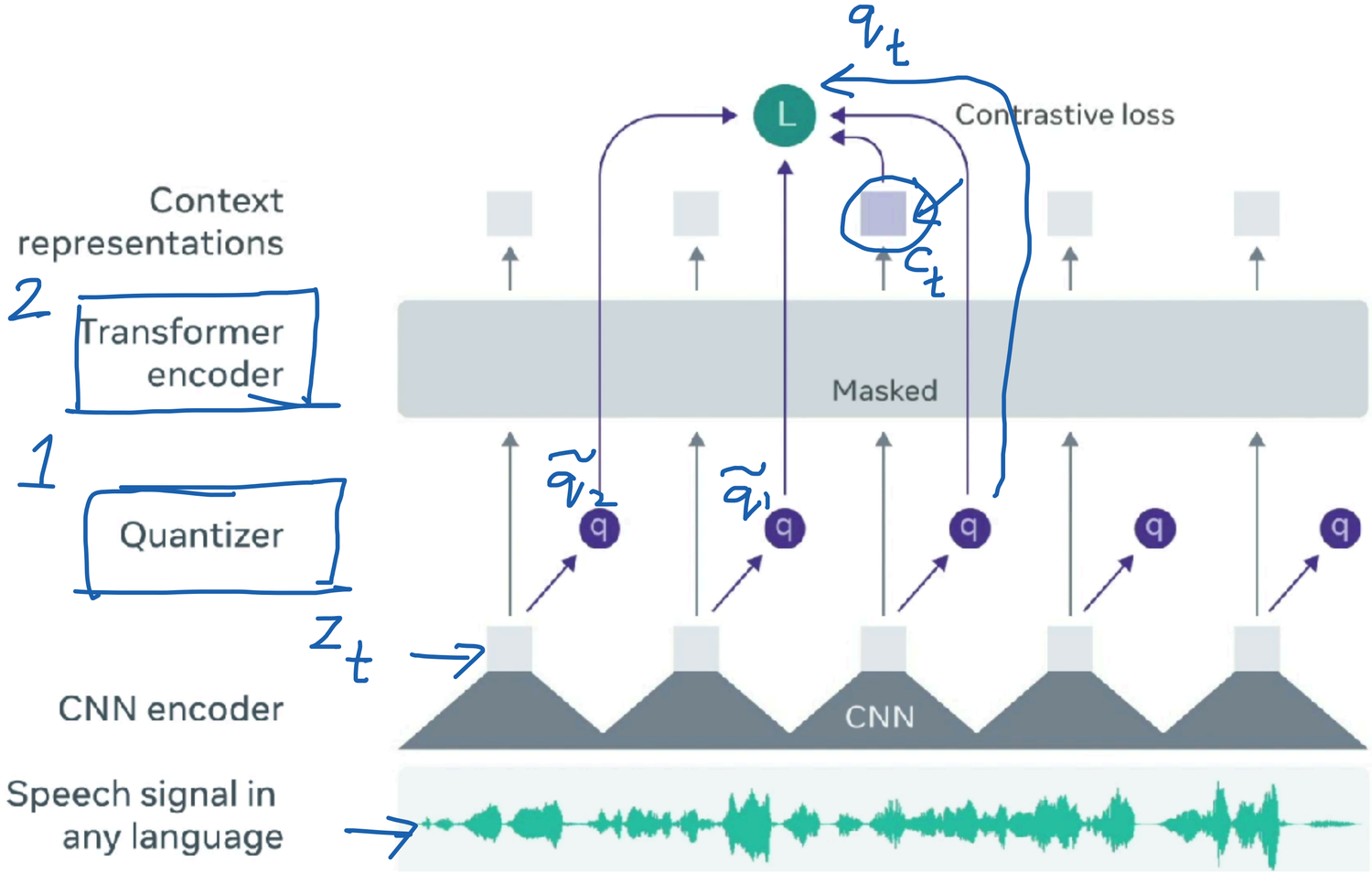
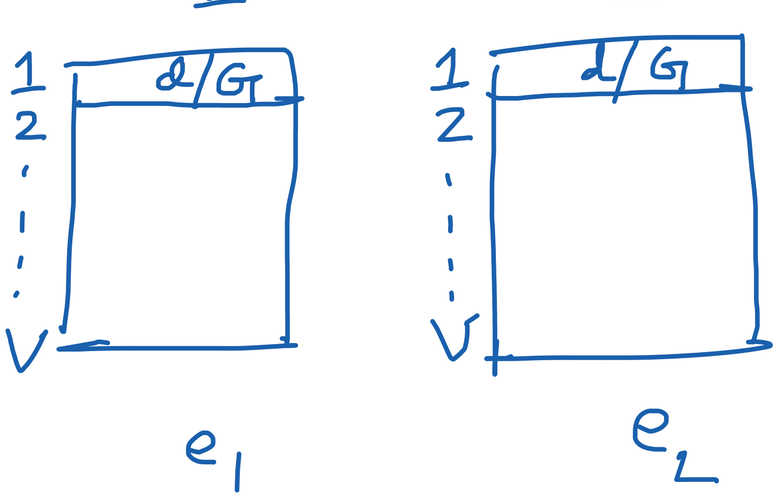
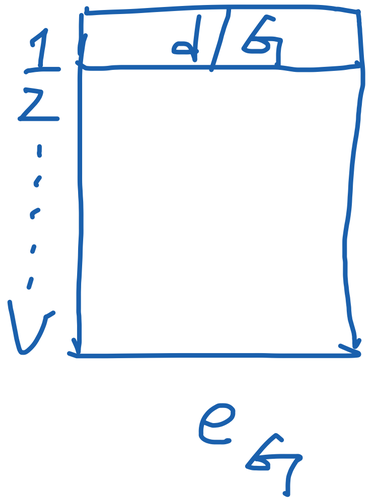


Image from:https://ai.facebook.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/

[1]: https://arxiv.org/abs/1810.04805

(2) mapped using a learnable quantization matrix $\rightarrow$ $\ell \in \mathbb{R}^{G \times V}$





2019

$$P_{g,v} = \frac{\exp(\ell_{g,v} + b_v / \tau)}{\sum_{v'=1}^{V} \exp(\ell_{g,v'} + b_{v'} / \tau)}$$

where $b_v \sim \text{Gumbel}(0,1)$

$$\tilde{Z} = [e_1 | e_2 | \dots | e_G]$$

Gumbel Softmax Trick

$\Rightarrow$ Sample from categorical distribution & make it differentiate

$\pi_i, \sum_i \pi_i = 1$

Gumbel trick: $b = \text{one-hot}(\text{argmax}_i(g_i + \log \pi_i))$

$g_i \sim \text{Gumbel}(0,1)$ $\quad u \sim \text{Unif}(0,1)$, $g = -\log(-\log u)$

wav2vec 2.0: Results on English

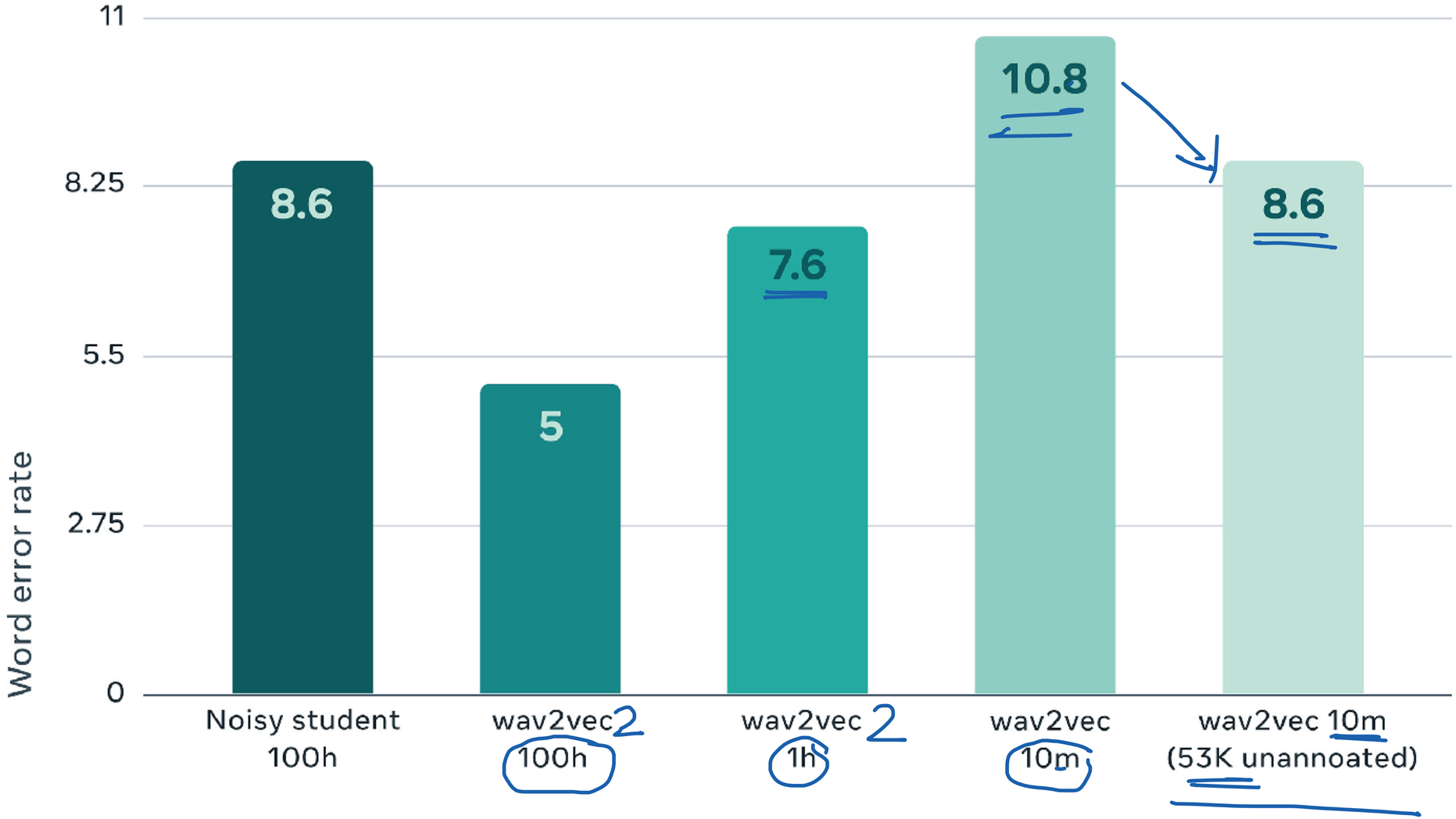


Image from:https://ai.facebook.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/

XLSR: Multilingual Self-supervised Speech Model

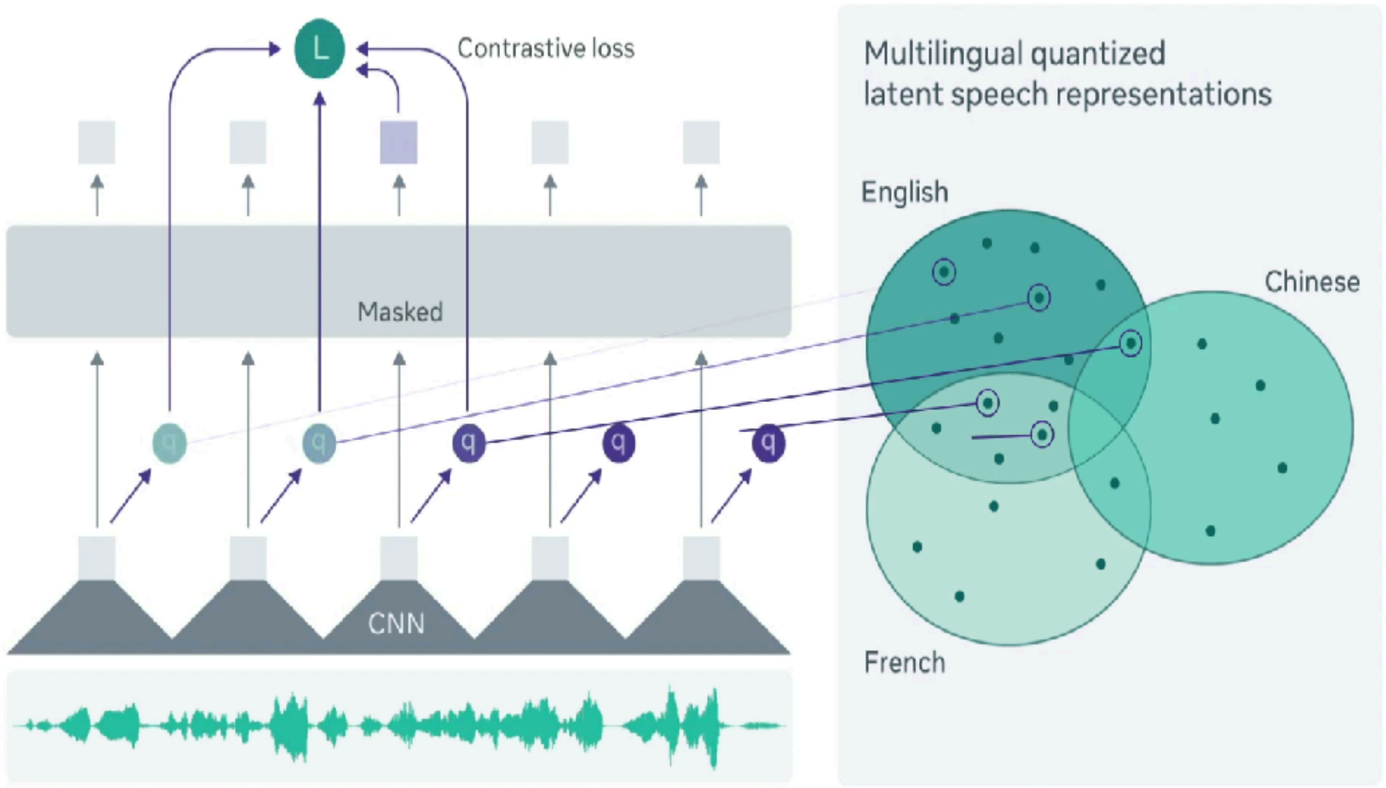


Image from:https://ai.facebook.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/

[1]: https://arxiv.org/abs/1810.04805

XLSR: Multilingual Self-supervised Speech Model

- wav2vec 2.0 model trained on speech in 53 languages

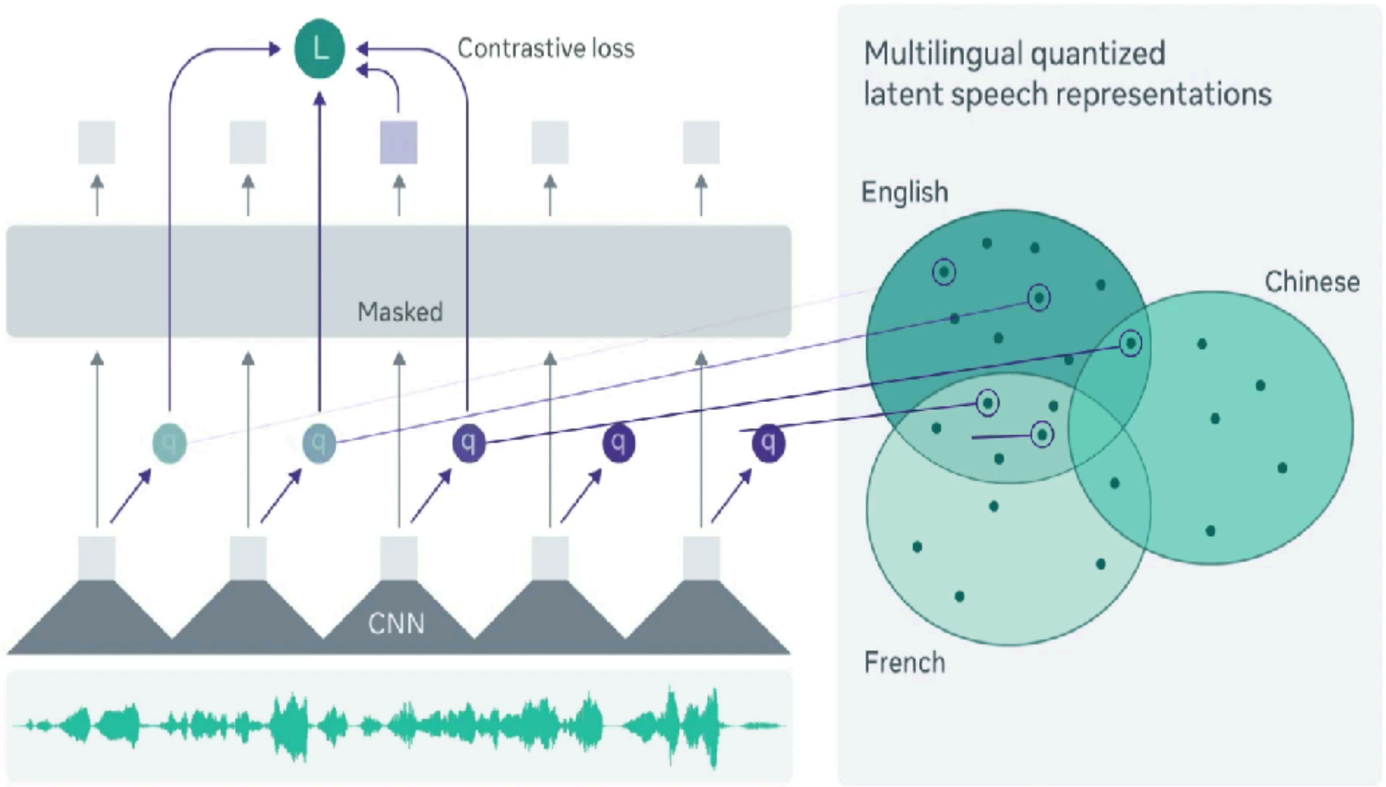


Image from:https://ai.facebook.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/

[1]: https://arxiv.org/abs/1810.04805

XLSR: Multilingual Self-supervised Speech Model

- wav2vec 2.0 model trained on speech in 53 languages
- Training objective is to recover the masked representations within a set of distractors

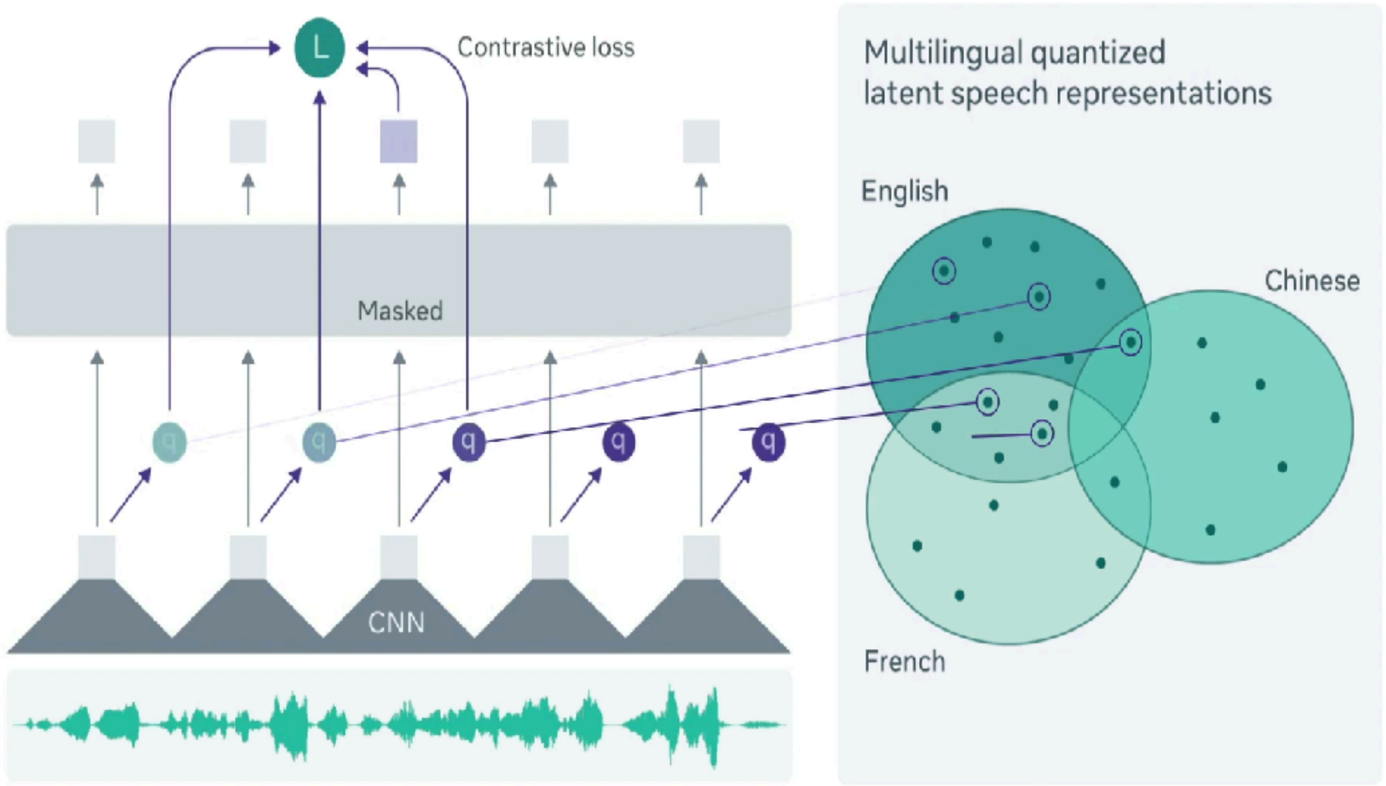


Image from:https://ai.facebook.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/

[1]: https://arxiv.org/abs/1810.04805

XLSR: Multilingual Self-supervised Speech Model

- wav2vec 2.0 model trained on speech in 53 languages
- Training objective is to recover the masked representations within a set of distractors
- Cross-lingual pretraining significantly outperforms monolingual pretraining

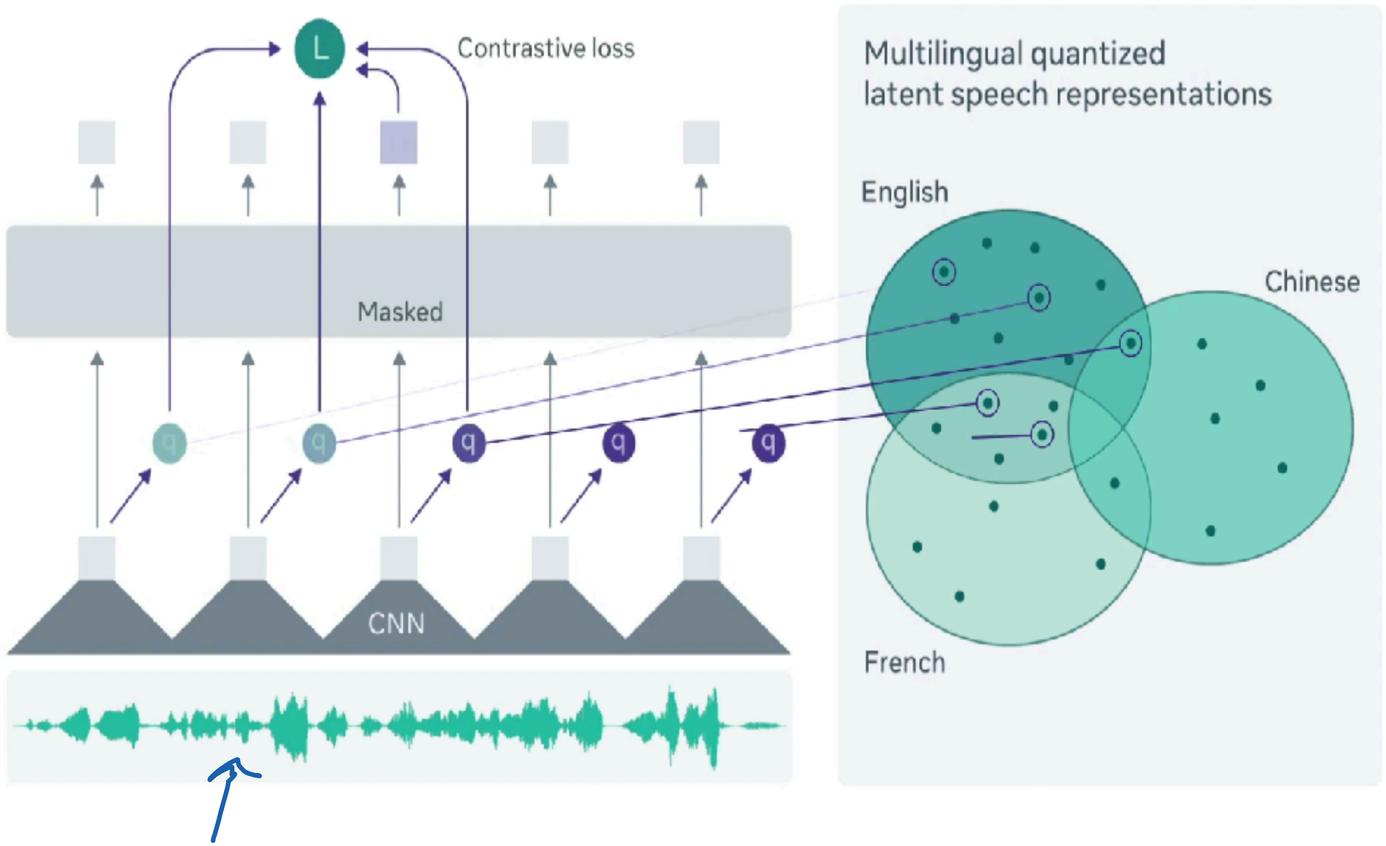


Image from:https://ai.facebook.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/

[1]: https://arxiv.org/abs/1810.04805

XLSR Results

Model	D	#pt	#ft	es	fr	it	ky	nl	ru	sv	tr	tt	zh	Avg
Number of pretraining hours per language				168h	353h	90h	17h	29h	55h	3h	11h	17h	50h	793h
Number of fine-tuning hours per language				1h	1h	1h	1h	1h	1h	1h	1h	1h	1h	10h
Baselines from previous work														
m-CPC† (Rivière et al., 2020)	LS_{100h}	10	1	38.7	49.3	42.1	40.7	44.4	45.2	48.8	49.7	44.0	55.5	45.8
m-CPC† (Rivière et al., 2020)	LS_{360h}	10	1	38.0	47.1	40.5	41.2	42.5	43.7	47.5	47.3	42.0	55.0	44.5
Fer et al.† (Fer et al., 2017)	BBL_{all}	10	1	36.6	48.3	39.0	38.7	47.9	45.2	52.6	43.4	42.5	54.3	44.9
Our monolingual models														
XLSR-English	CV_{en}	1	1	13.7	20.0	19.1	13.2	19.4	18.6	21.1	15.5	11.5	27.1	17.9
XLSR-Monolingual	CV_{mo}	1	1	6.8	10.4	10.9	29.6	37.4	11.6	63.6	44.0	21.4	31.4	26.7
Our multilingual models														
XLSR-10 (unbalanced)	CV_{all}	10	1	9.7	13.6	15.2	11.1	18.1	13.7	21.4	14.2	9.7	25.8	15.3
XLSR-10	CV_{all}	10	1	9.4	14.2	14.1	8.4	16.1	11.0	20.7	11.2	7.6	24.0	13.6
XLSR-10 (separate vocab)	CV_{all}	10	10	10.0	13.8	14.0	8.8	16.5	11.6	21.4	12.0	8.7	24.5	14.1
XLSR-10 (shared vocab)	CV_{all}	10	10	9.4	13.4	13.8	8.6	16.3	11.2	21.0	11.7	8.3	24.5	13.8
Our multilingual models (Large)														
XLSR-10	CV_{all}	10	1	7.9	12.6	11.7	7.0	14.0	9.3	20.6	9.7	7.2	22.8	12.3
XLSR-10 (separate vocab)	CV_{all}	10	10	8.1	12.1	11.9	7.1	13.9	9.8	21.0	10.4	7.6	22.3	12.4
XLSR-10 (shared vocab)	CV_{all}	10	10	7.7	12.2	11.6	7.0	13.8	9.3	20.8	10.1	7.3	22.3	12.2
Our Large XLSR-53 model pretrained on 56k hours														
XLSR-53	D_{53}	53	1	**2.9**	**5.0**	**5.7**	**6.1**	**5.8**	**8.1**	**12.2**	**7.1**	**5.1**	**18.3**	**7.6**

Image from: https://arxiv.org/pdf/2006.13979.pdf

XLSR Results

Model	D	#pt	#ft	es	fr	it	ky	nl	ru	sv	tr	tt	zh	Avg
Number of pretraining hours per language				168h	353h	90h	17h	29h	55h	3h	11h	17h	50h	793h
Number of fine-tuning hours per language				1h	1h	1h	1h	1h	1h	1h	1h	1h	1h	10h
Baselines from previous work														
m-CPC† (Rivière et al., 2020)	LS_{100h}	10	1	38.7	49.3	42.1	40.7	44.4	45.2	48.8	49.7	44.0	55.5	45.8
m-CPC† (Rivière et al., 2020)	LS_{360h}	10	1	38.0	47.1	40.5	41.2	42.5	43.7	47.5	47.3	42.0	55.0	44.5
Fer et al.† (Fer et al., 2017)	BBL_{all}	10	1	36.6	48.3	39.0	38.7	47.9	45.2	52.6	43.4	42.5	54.3	44.9
Our monolingual models														
XLSR-English	CV_{en}	1	1	13.7	20.0	19.1	13.2	19.4	18.6	21.1	15.5	11.5	27.1	17.9
XLSR-Monolingual	CV_{mo}	1	1	6.8	10.4	10.9	29.6	37.4	11.6	63.6	44.0	21.4	31.4	26.7
Our multilingual models														
XLSR-10 (unbalanced)	CV_{all}	10	1	9.7	13.6	15.2	11.1	18.1	13.7	21.4	14.2	9.7	25.8	15.3
XLSR-10	CV_{all}	10	1	9.4	14.2	14.1	8.4	16.1	11.0	20.7	11.2	7.6	24.0	13.6
XLSR-10 (separate vocab)	CV_{all}	10	10	10.0	13.8	14.0	8.8	16.5	11.6	21.4	12.0	8.7	24.5	14.1
XLSR-10 (shared vocab)	CV_{all}	10	10	9.4	13.4	13.8	8.6	16.3	11.2	21.0	11.7	8.3	24.5	13.8
Our multilingual models (Large)														
XLSR-10	CV_{all}	10	1	7.9	12.6	11.7	7.0	14.0	9.3	20.6	9.7	7.2	22.8	12.3
XLSR-10 (separate vocab)	CV_{all}	10	10	8.1	12.1	11.9	7.1	13.9	9.8	21.0	10.4	7.6	22.3	12.4
XLSR-10 (shared vocab)	CV_{all}	10	10	7.7	12.2	11.6	7.0	13.8	9.3	20.8	10.1	7.3	22.3	12.2
Our Large XLSR-53 model pretrained on 56k hours														
XLSR-53	D_{53}	53	1	**2.9**	**5.0**	**5.7**	**6.1**	**5.8**	**8.1**	**12.2**	**7.1**	**5.1**	**18.3**	**7.6**

Image from: https://arxiv.org/pdf/2006.13979.pdf

XLSR Results

Model	D	#pt	#ft	es	fr	it	ky	nl	ru	sv	tr	tt	zh	Avg
Number of pretraining hours per language				168h	353h	90h	17h	29h	55h	3h	11h	17h	50h	793h
Number of fine-tuning hours per language				1h	1h	1h	1h	1h	1h	1h	1h	1h	1h	10h
Baselines from previous work														
m-CPC† (Rivière et al., 2020)	LS_{100h}	10	1	38.7	49.3	42.1	40.7	44.4	45.2	48.8	49.7	44.0	55.5	45.8
m-CPC† (Rivière et al., 2020)	LS_{360h}	10	1	38.0	47.1	40.5	41.2	42.5	43.7	47.5	47.3	42.0	55.0	44.5
Fer et al.† (Fer et al., 2017)	BBL_{all}	10	1	36.6	48.3	39.0	38.7	47.9	45.2	52.6	43.4	42.5	54.3	44.9
Our monolingual models														
XLSR-English	CV_{en}	1	1	13.7	20.0	19.1	13.2	19.4	18.6	21.1	15.5	11.5	27.1	17.9
XLSR-Monolingual	CV_{mo}	1	1	6.8	10.4	10.9	29.6	37.4	11.6	63.6	44.0	21.4	31.4	26.7
Our multilingual models														
XLSR-10 (unbalanced)	CV_{all}	10	1	9.7	13.6	15.2	11.1	18.1	13.7	21.4	14.2	9.7	25.8	15.3
XLSR-10	CV_{all}	10	1	9.4	14.2	14.1	8.4	16.1	11.0	20.7	11.2	7.6	24.0	13.6
XLSR-10 (separate vocab)	CV_{all}	10	10	10.0	13.8	14.0	8.8	16.5	11.6	21.4	12.0	8.7	24.5	14.1
XLSR-10 (shared vocab)	CV_{all}	10	10	9.4	13.4	13.8	8.6	16.3	11.2	21.0	11.7	8.3	24.5	13.8
Our multilingual models (Large)														
XLSR-10	CV_{all}	10	1	7.9	12.6	11.7	7.0	14.0	9.3	20.6	9.7	7.2	22.8	12.3
XLSR-10 (separate vocab)	CV_{all}	10	10	8.1	12.1	11.9	7.1	13.9	9.8	21.0	10.4	7.6	22.3	12.4
XLSR-10 (shared vocab)	CV_{all}	10	10	7.7	12.2	11.6	7.0	13.8	9.3	20.8	10.1	7.3	22.3	12.2
Our Large XLSR-53 model pretrained on 56k hours														
XLSR-53	D_{53}	53	1	**2.9**	**5.0**	**5.7**	**6.1**	**5.8**	**8.1**	**12.2**	**7.1**	**5.1**	**18.3**	**7.6**

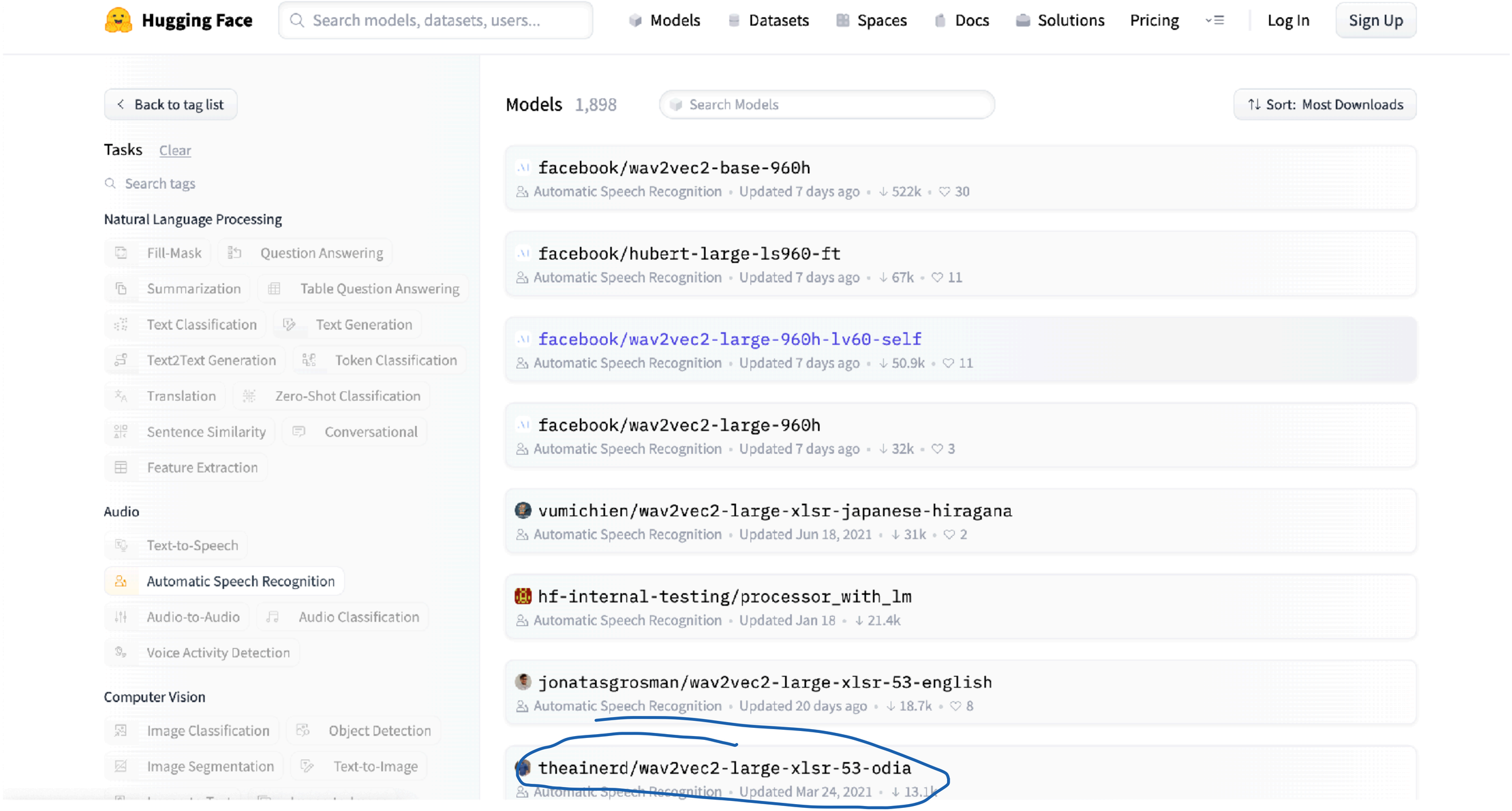
Image from: https://arxiv.org/pdf/2006.13979.pdf

XLSR Results

Model	D	#pt	#ft	es	fr	it	ky	nl	ru	sv	tr	tt	zh	Avg
Number of pretraining hours per language				168h	353h	90h	17h	29h	55h	3h	11h	17h	50h	793h
Number of fine-tuning hours per language				1h	1h	1h	1h	1h	1h	1h	1h	1h	1h	10h
Baselines from previous work														
m-CPC† (Rivière et al., 2020)	LS_{100h}	10	1	38.7	49.3	42.1	40.7	44.4	45.2	48.8	49.7	44.0	55.5	45.8
m-CPC† (Rivière et al., 2020)	LS_{360h}	10	1	38.0	47.1	40.5	41.2	42.5	43.7	47.5	47.3	42.0	55.0	44.5
Fer et al.† (Fer et al., 2017)	BBL_{all}	10	1	36.6	48.3	39.0	38.7	47.9	45.2	52.6	43.4	42.5	54.3	44.9
Our monolingual models														
XLSR-English	CV_{en}	1	1	13.7	20.0	19.1	13.2	19.4	18.6	21.1	15.5	11.5	27.1	17.9
XLSR-Monolingual	CV_{mo}	1	1	6.8	10.4	10.9	29.6	37.4	11.6	63.6	44.0	21.4	31.4	26.7
Our multilingual models														
XLSR-10 (unbalanced)	CV_{all}	10	1	9.7	13.6	15.2	11.1	18.1	13.7	21.4	14.2	9.7	25.8	15.3
XLSR-10	CV_{all}	10	1	9.4	14.2	14.1	8.4	16.1	11.0	20.7	11.2	7.6	24.0	13.6
XLSR-10 (separate vocab)	CV_{all}	10	10	10.0	13.8	14.0	8.8	16.5	11.6	21.4	12.0	8.7	24.5	14.1
XLSR-10 (shared vocab)	CV_{all}	10	10	9.4	13.4	13.8	8.6	16.3	11.2	21.0	11.7	8.3	24.5	13.8
Our multilingual models (Large)														
XLSR-10	CV_{all}	10	1	7.9	12.6	11.7	7.0	14.0	9.3	20.6	9.7	7.2	22.8	12.3
XLSR-10 (separate vocab)	CV_{all}	10	10	8.1	12.1	11.9	7.1	13.9	9.8	21.0	10.4	7.6	22.3	12.4
XLSR-10 (shared vocab)	CV_{all}	10	10	7.7	12.2	11.6	7.0	13.8	9.3	20.8	10.1	7.3	22.3	12.2
Our Large XLSR-53 model pretrained on 56k hours														
XLSR-53	D_{53}	53	1	**2.9**	**5.0**	**5.7**	**6.1**	**5.8**	**8.1**	**12.2**	**7.1**	**5.1**	**18.3**	**7.6**

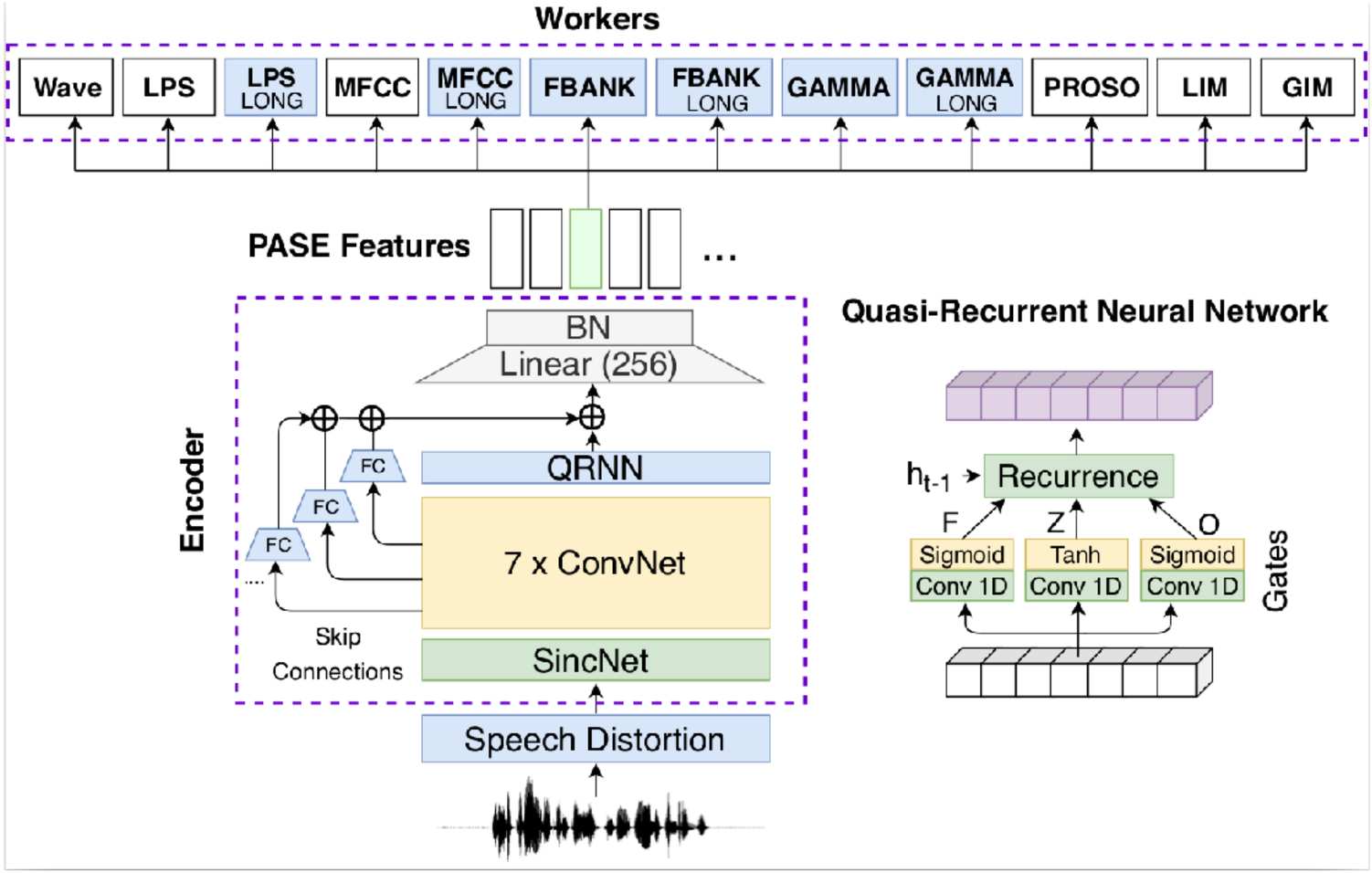
Image from: https://arxiv.org/pdf/2006.13979.pdf

HuggingFace wav2vec Models



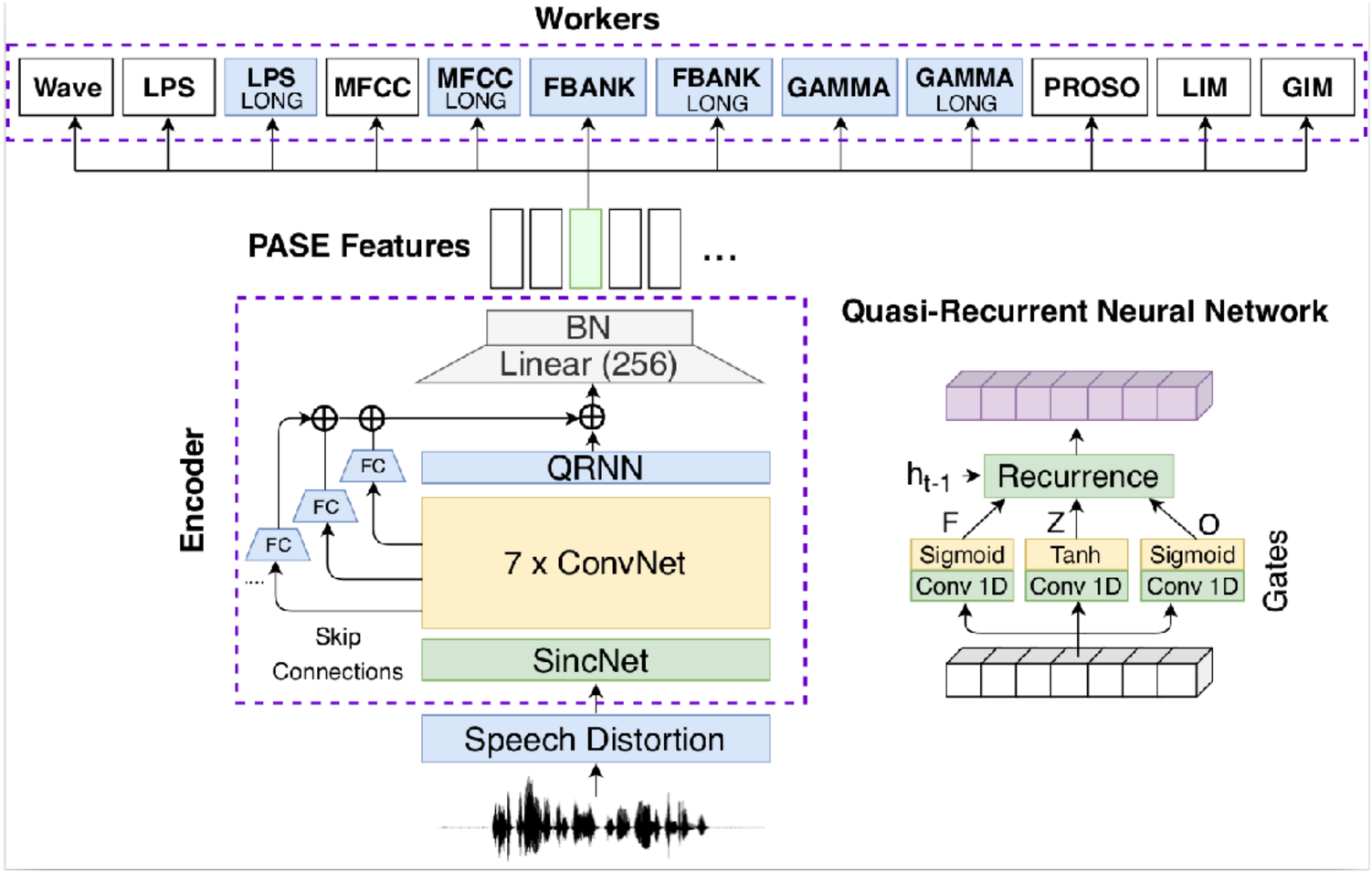
https://huggingface.co/models?pipeline_tag=automatic-speech-recognition&sort=downloads

Multi-task: PASE/PASE+



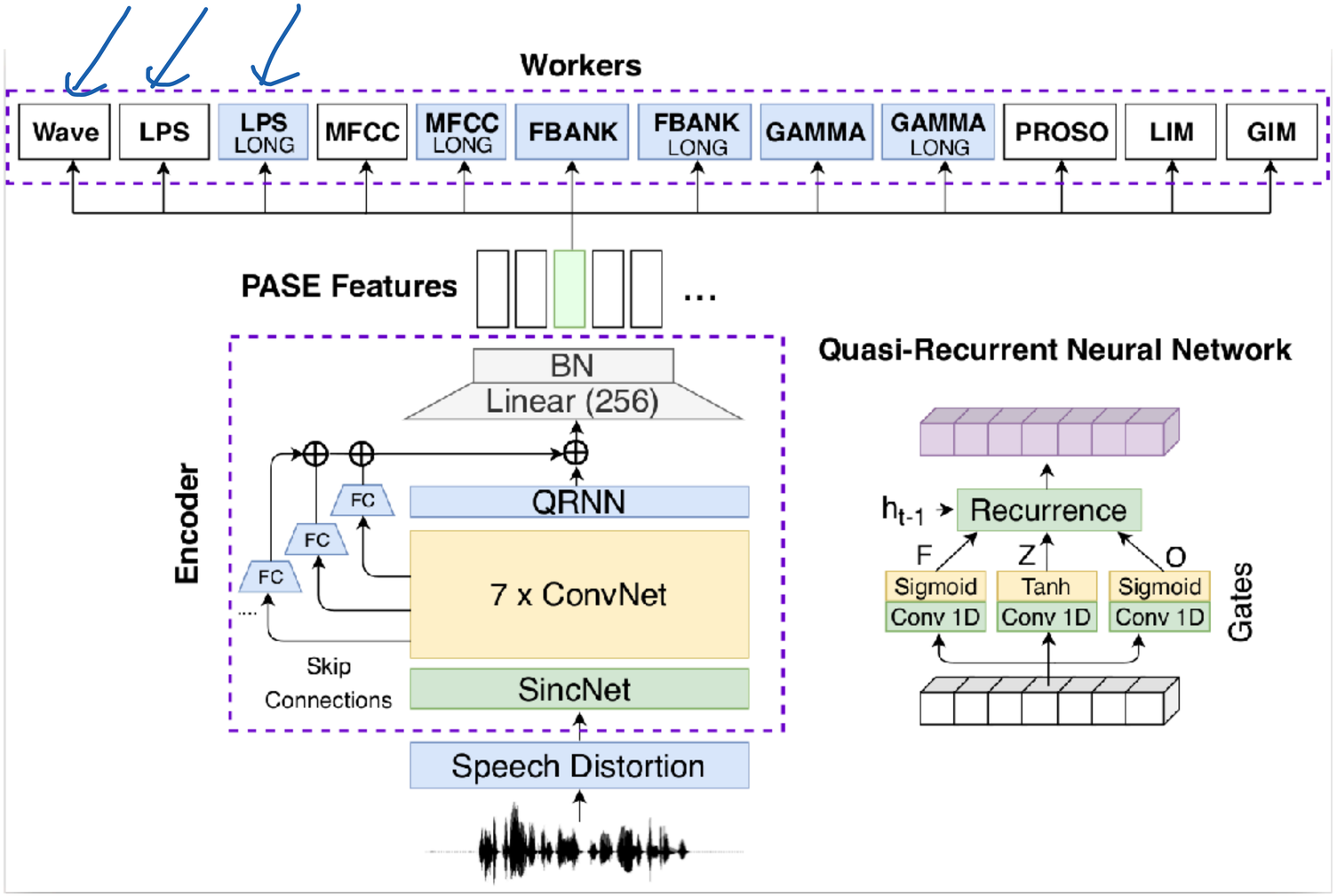
Multi-task: PASE/PASE+

- PASE consists of a speech encoder, followed by an army of workers with self-supervised objectives trained in a multi-task setup.

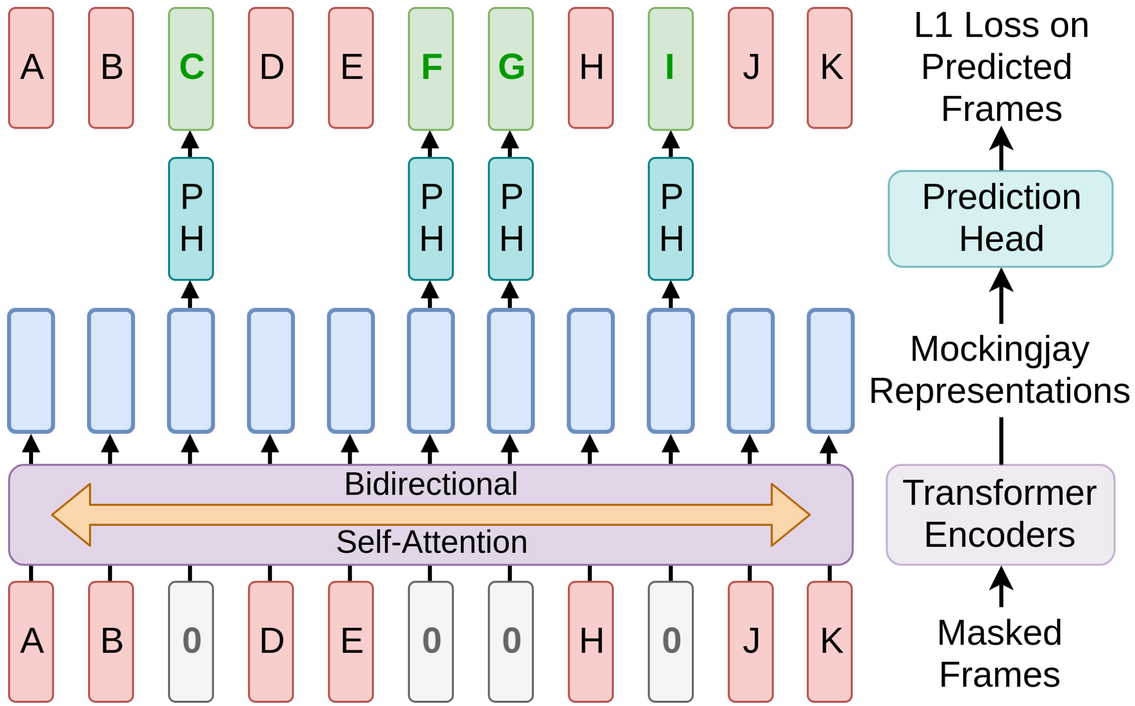


Multi-task: PASE/PASE+

- PASE consists of a speech encoder, followed by an army of workers with self-supervised objectives trained in a multi-task setup.
- Self-supervised tasks are either regression tasks (MFCC, LPS, etc.) or classification tasks (LIM, GIM, etc.).



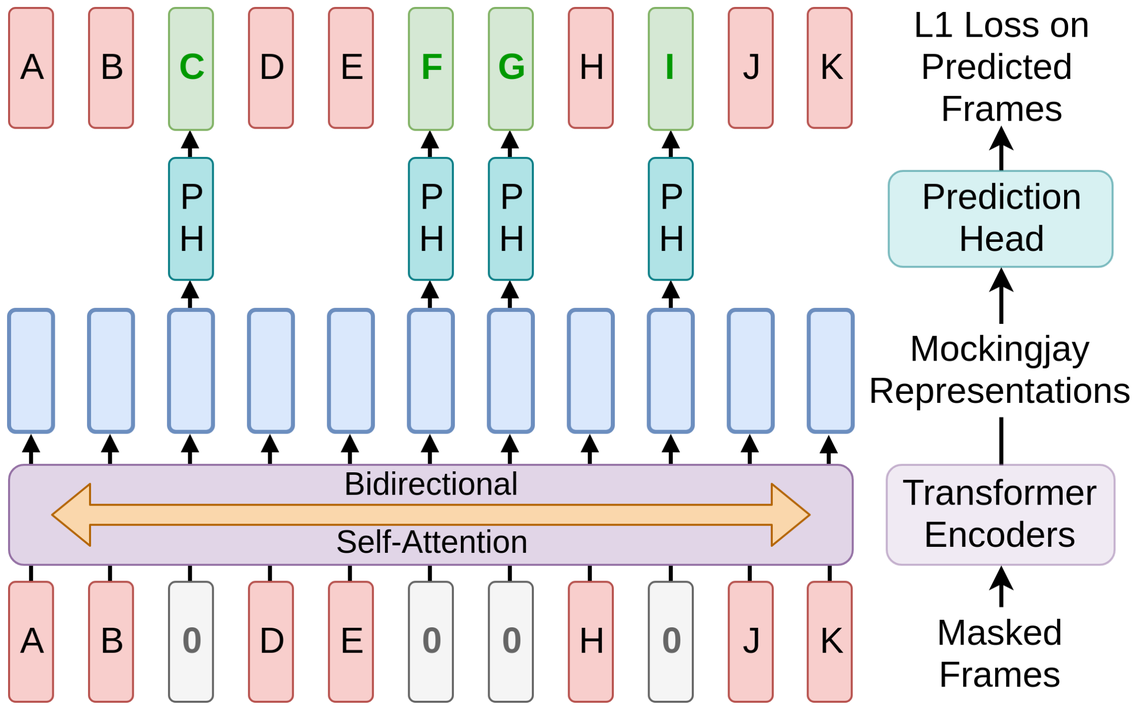
Masked Reconstruction: Mockingjay



Image/table from: https://arxiv.org/pdf/1910.12638.pdf (Liu et al., 2020)

Masked Reconstruction: Mockingjay

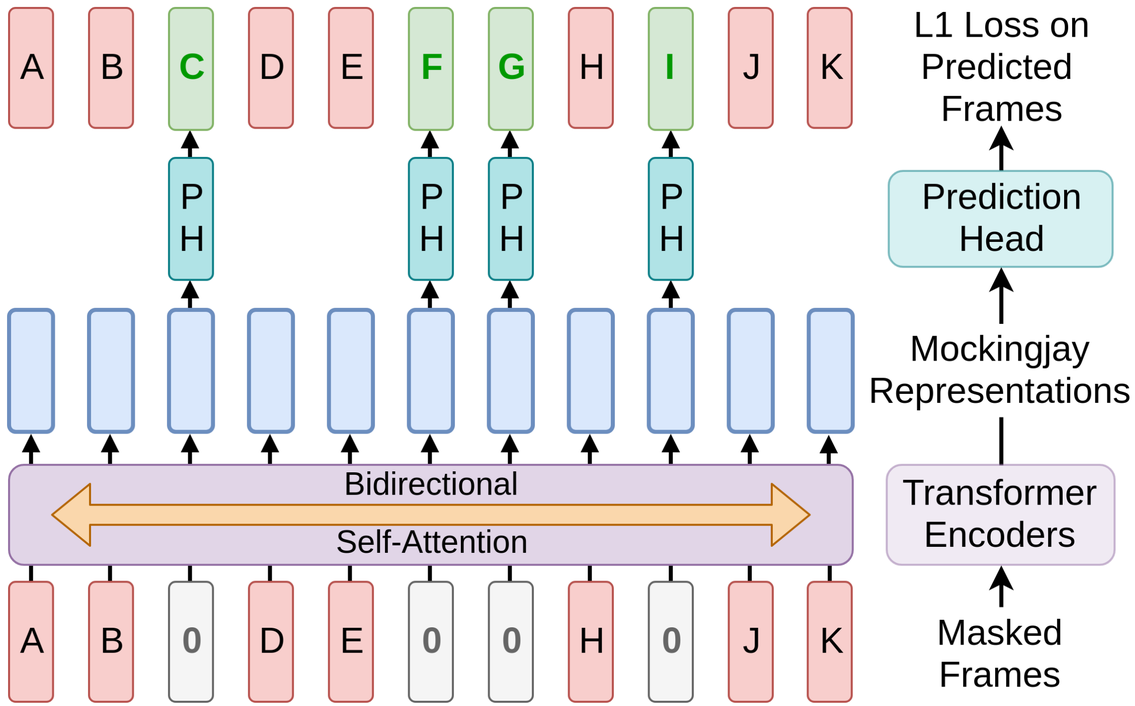
- Predict the current frame by jointly conditioning on both past and future contexts



Image/table from: https://arxiv.org/pdf/1910.12638.pdf (Liu et al., 2020)

Masked Reconstruction: Mockingjay

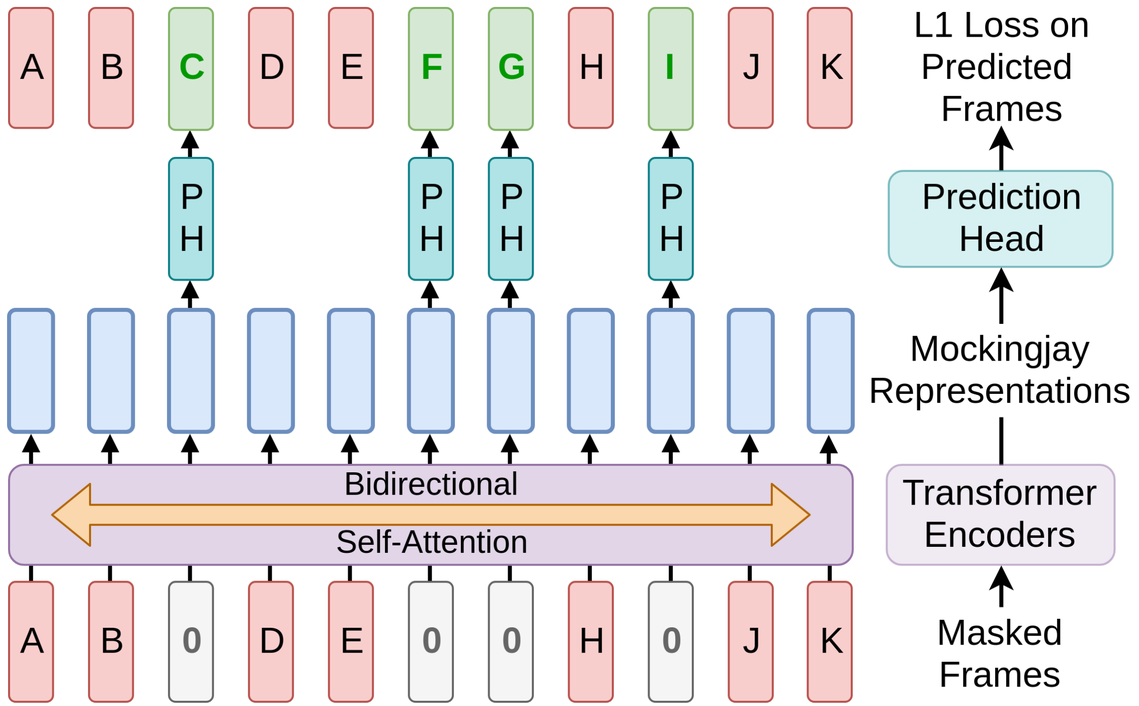
- Predict the current frame by jointly conditioning on both past and future contexts
- Minimize reconstruction error between predictions and ground-truth on the masked frames



Image/table from: https://arxiv.org/pdf/1910.12638.pdf (Liu et al., 2020)

Masked Reconstruction: Mockingjay

- Predict the current frame by jointly conditioning on both past and future contexts
- Minimize reconstruction error between predictions and ground-truth on the masked frames
- Final representations are the Transformer encoder's hidden states



Image/table from: https://arxiv.org/pdf/1910.12638.pdf (Liu et al., 2020)

S3PRL Toolkit

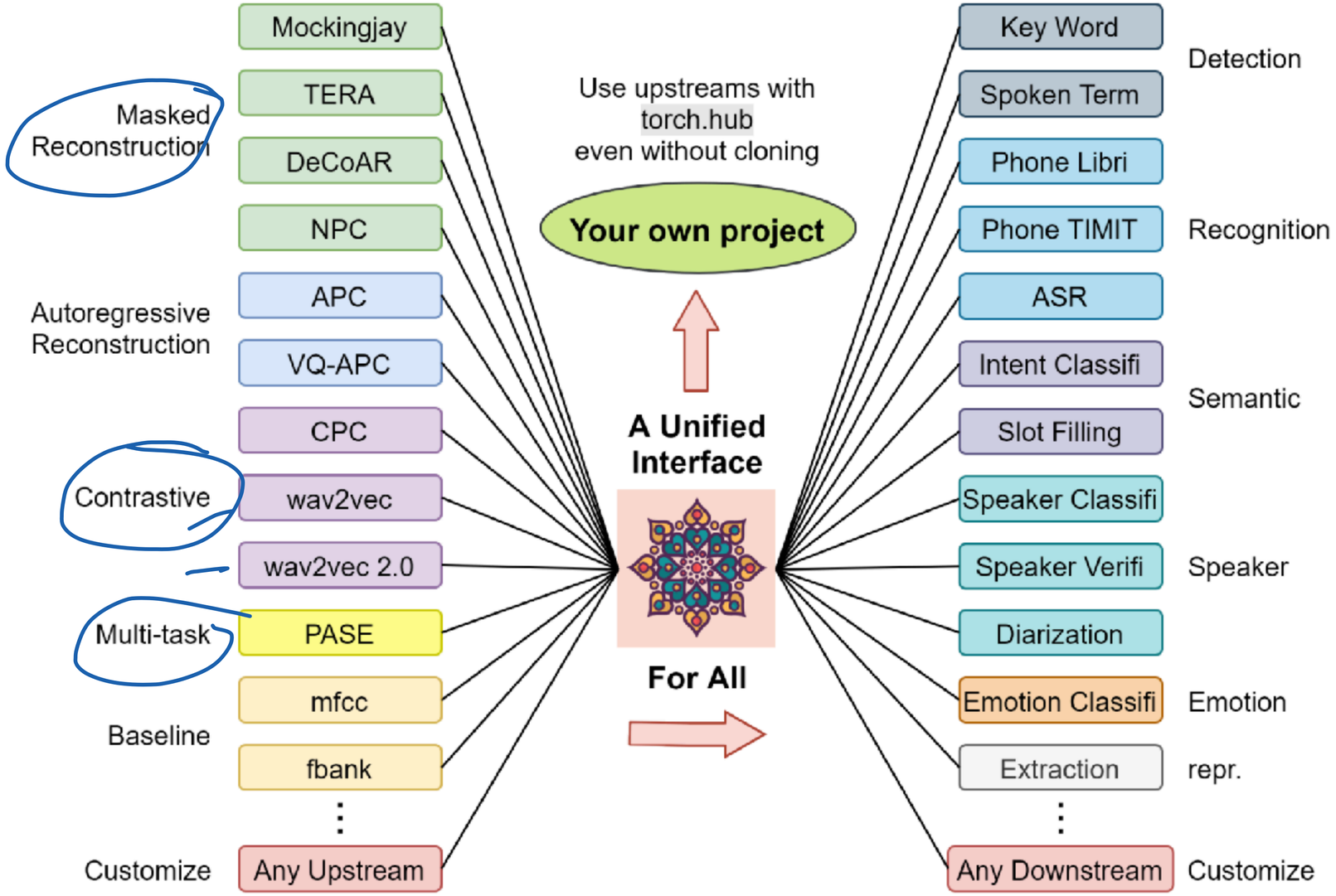
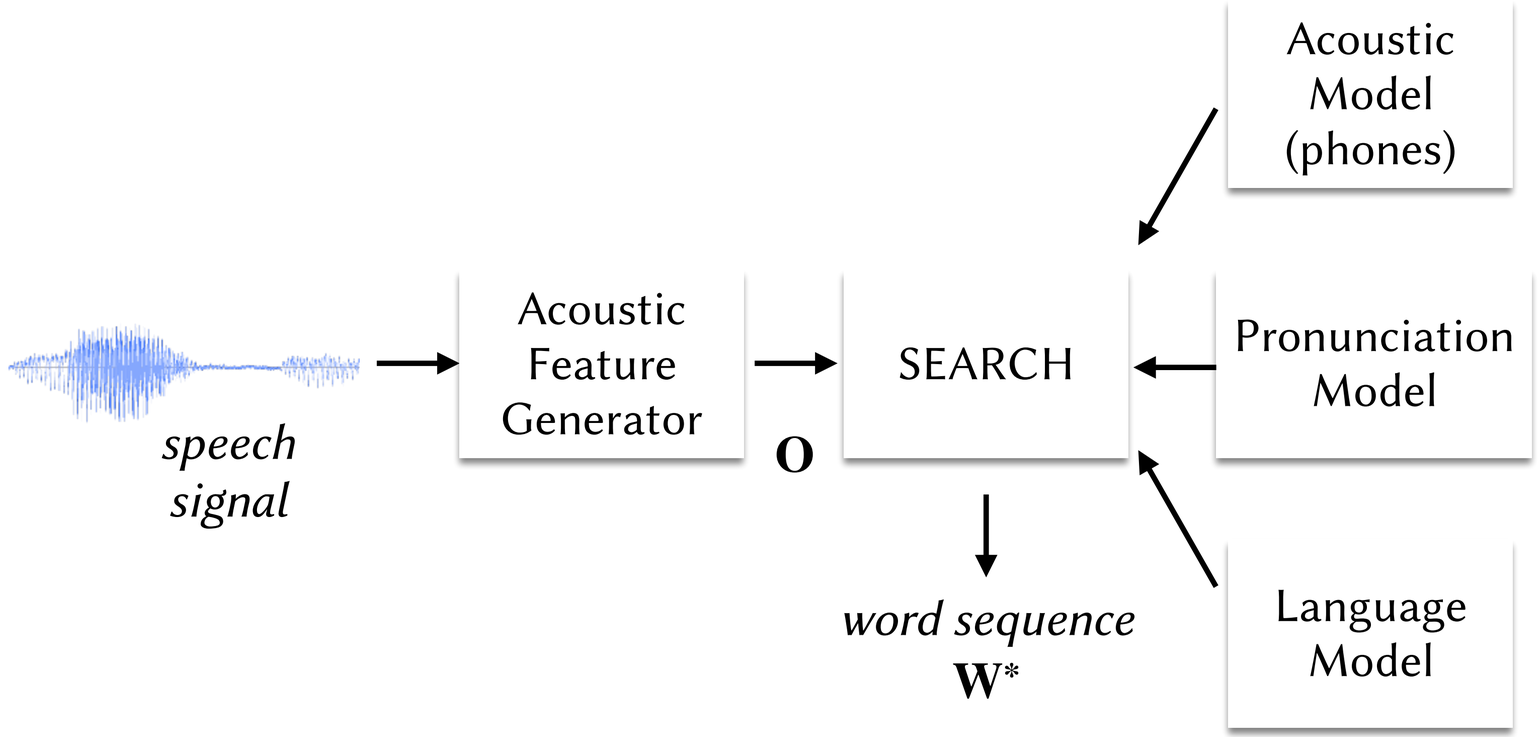


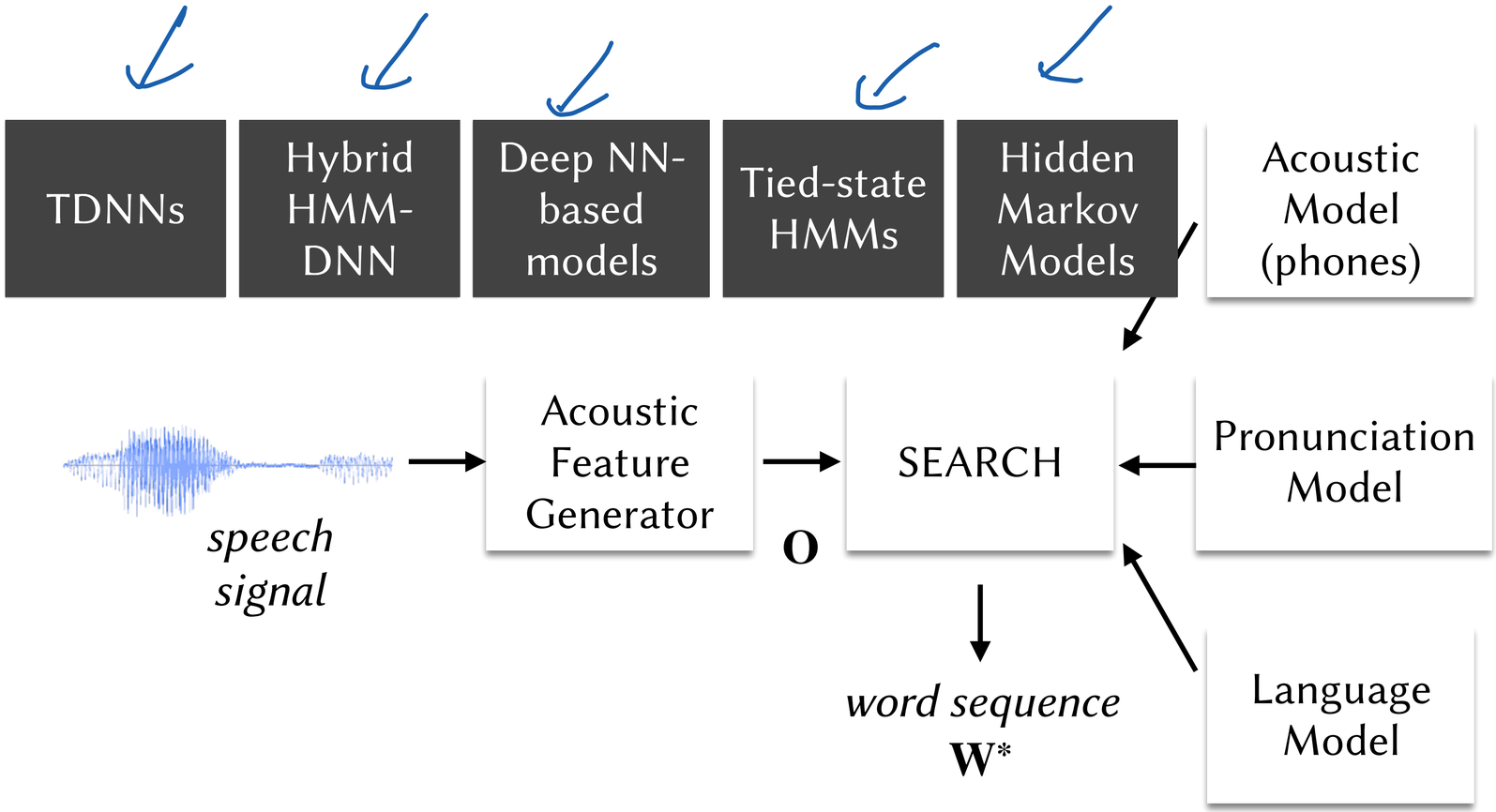
Image from: https://github.com/s3prl/s3prl

CS753 Concluding Remarks

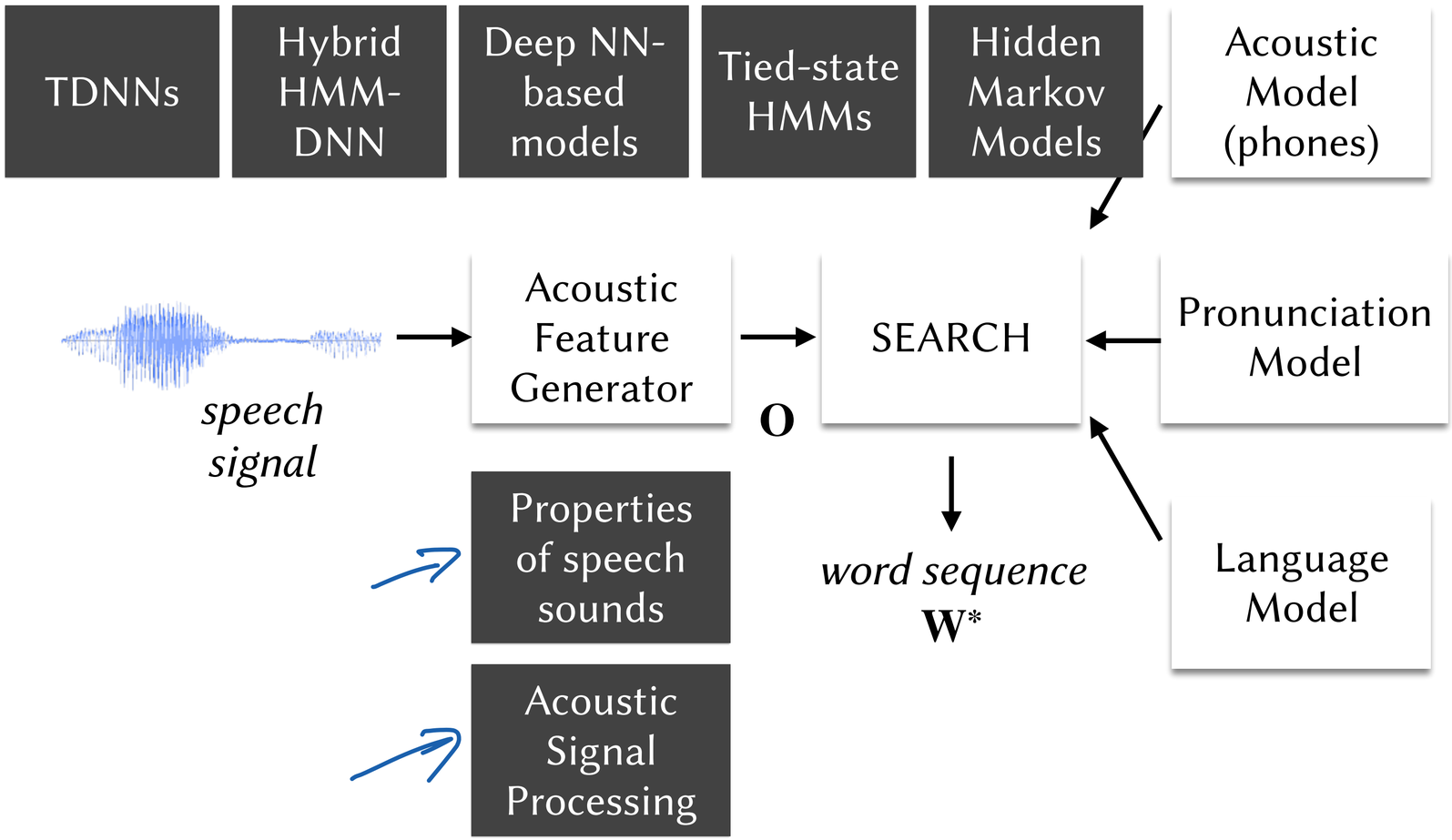
Topics covered



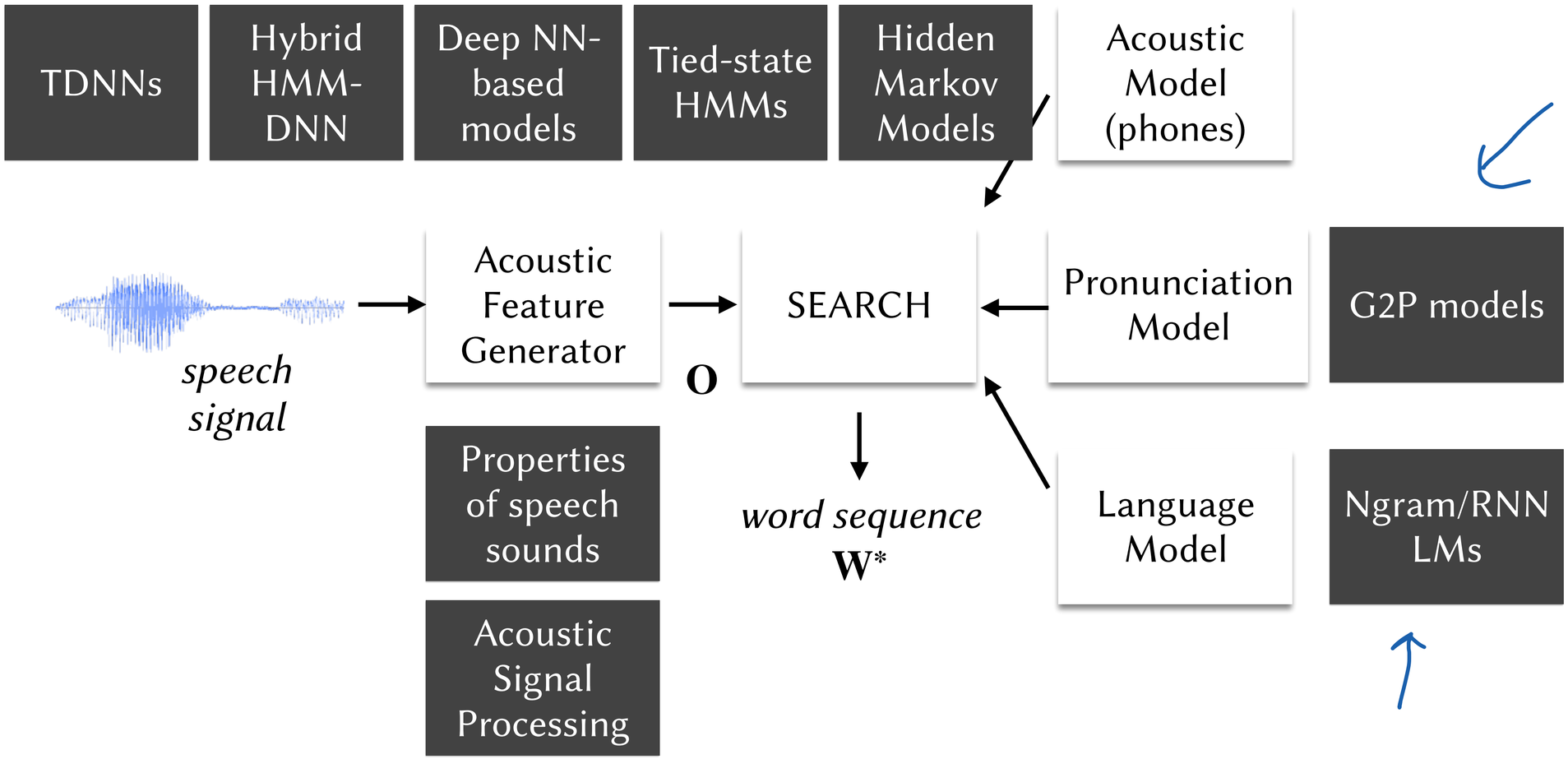
Topics covered



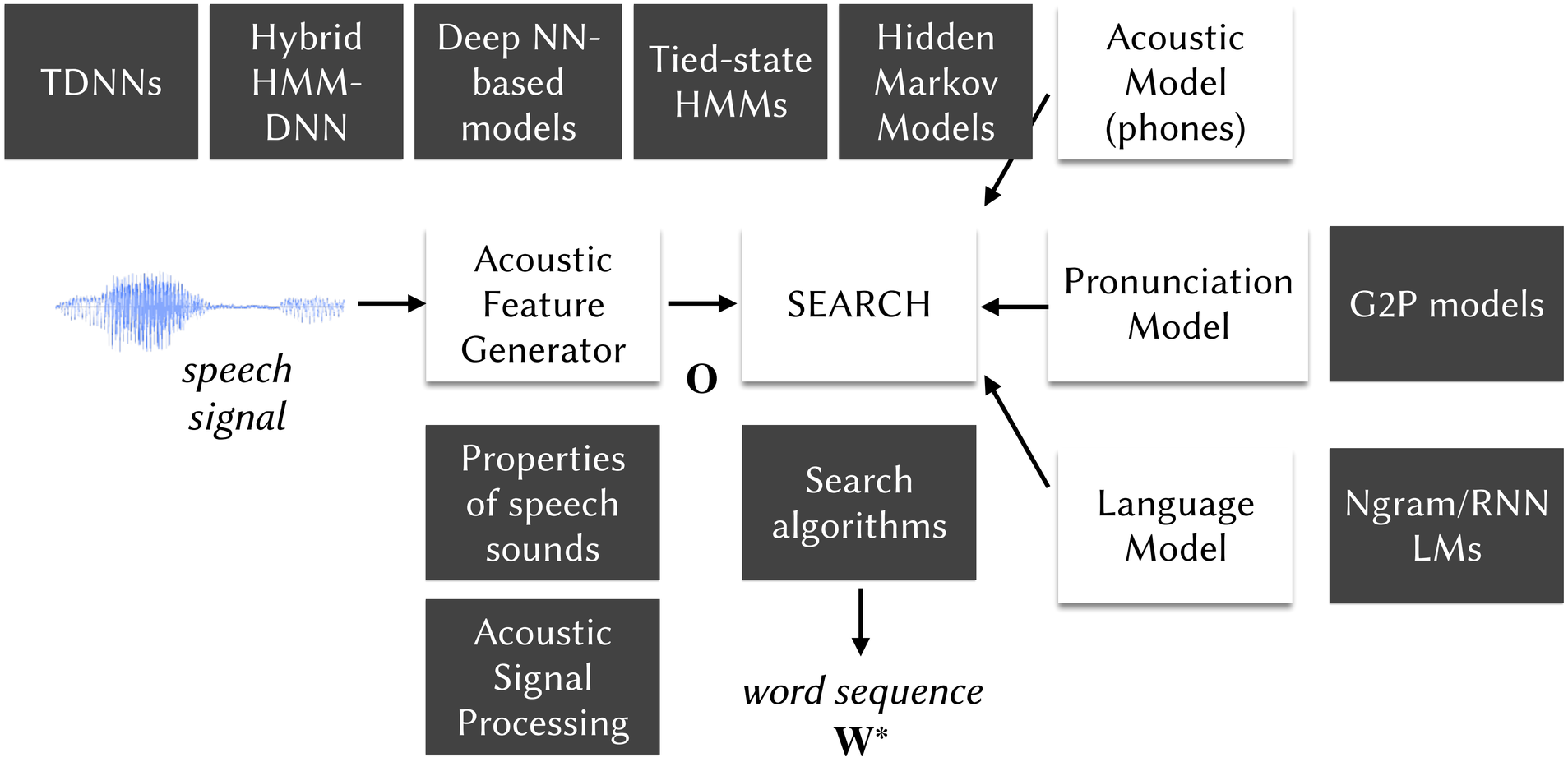
Topics covered



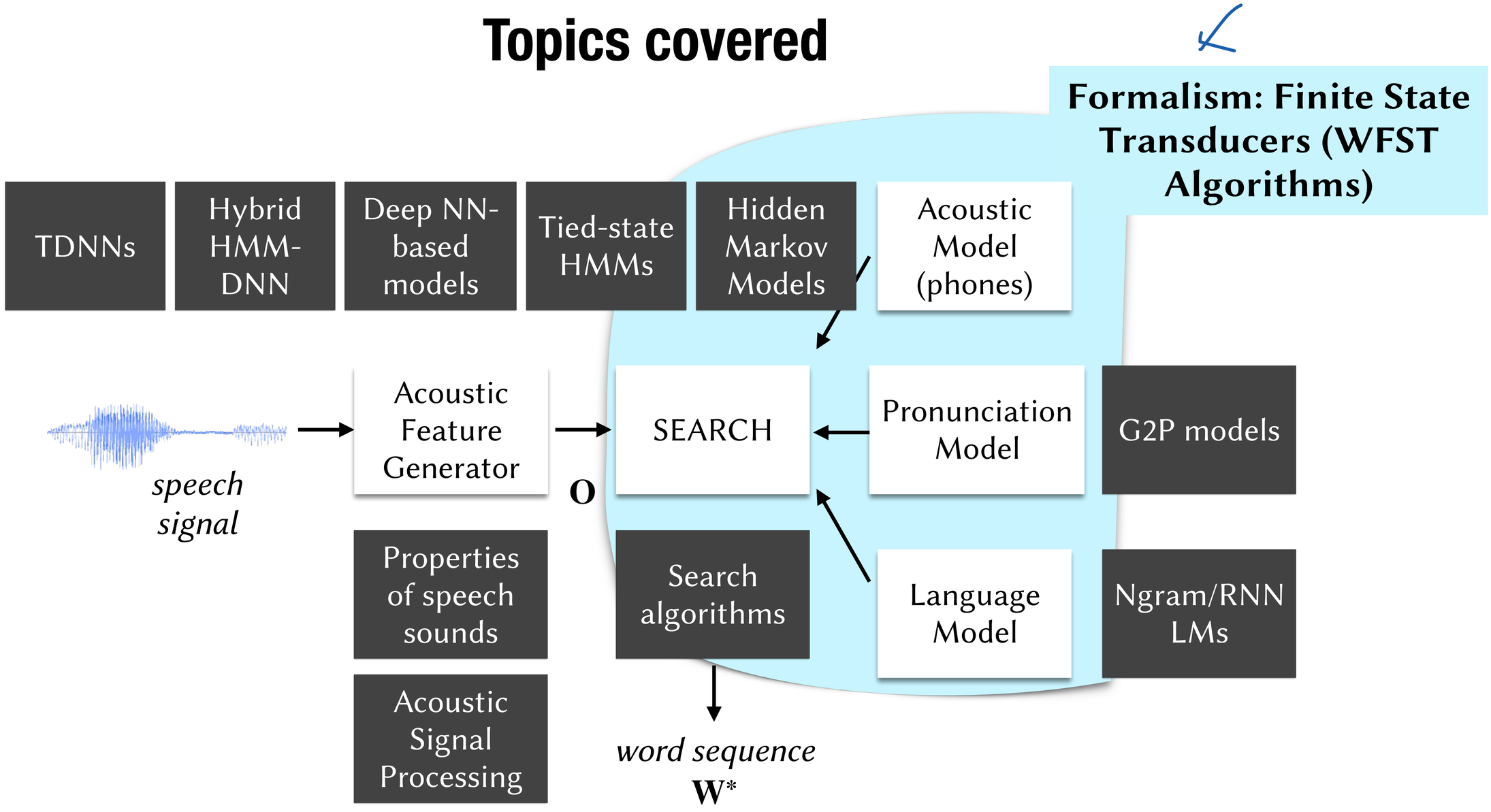
Topics covered



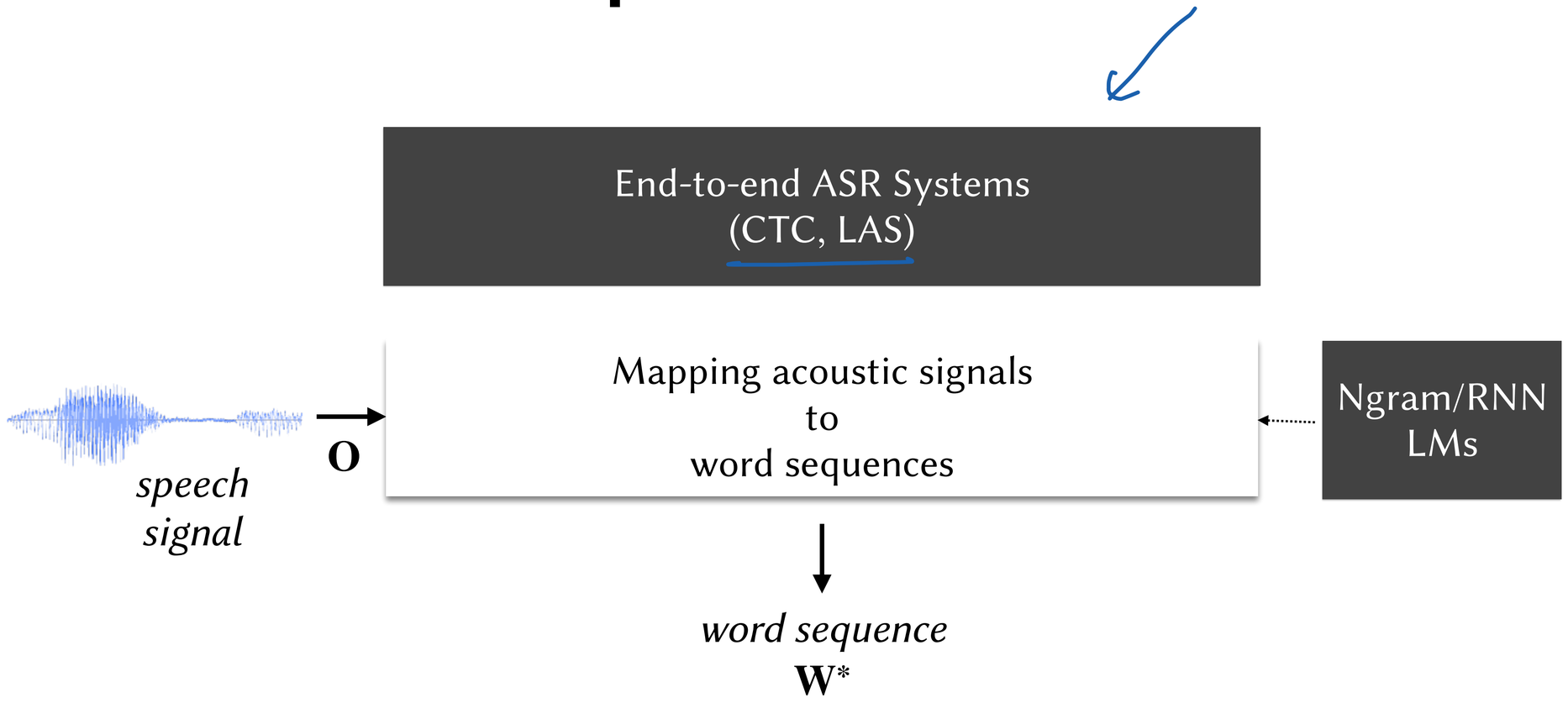
Topics covered



Topics covered



Topics covered

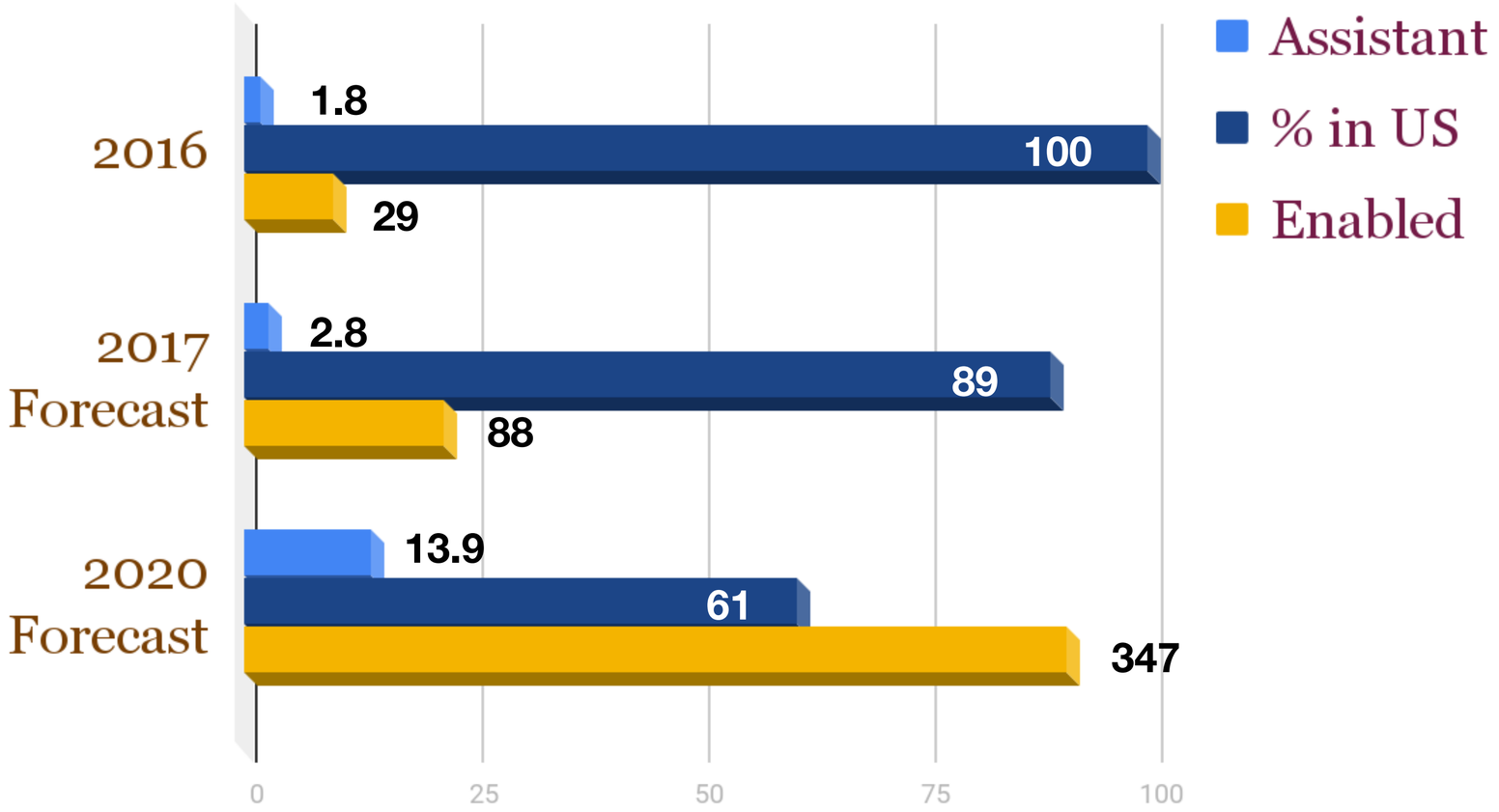


Seminar Topics

	EMFORMER	Blocksync Beam	Voice Conv	Simul Speech	Stream Trans ASR 1,2	
Auto VC	AV Dereverb	Relthink ASR	MaskCTC	Audio Adversary	Quaternion	VoiceSep
Rhythm Transfer	wav2vec U	Right2Talk	NeuralSynth	BERT ASR	AV Robustness	SpeakQL
SoundSep Video	Diarization	Visual Voice	Large Margin SV	Efficient Vocalize	Multiling ASR	Style Tokens
ImpercepAdv ASR	Look2Speech	CluedAVVP	GE2E	TriBERT	Speech2SQL	LatticeInputs
Spoken Translation	Move2Hear	RNNT Decoding	AVID	Word-level E2E ASR	Deepfakes	Metric GAN

Exciting time to do speech research

Market for Voice



Exciting time to do speech research

Coqui, Freeing Speech

Coqui, a startup providing open speech tech for everyone 🐸
Sign up with your email address to receive the Coqui newsletter.

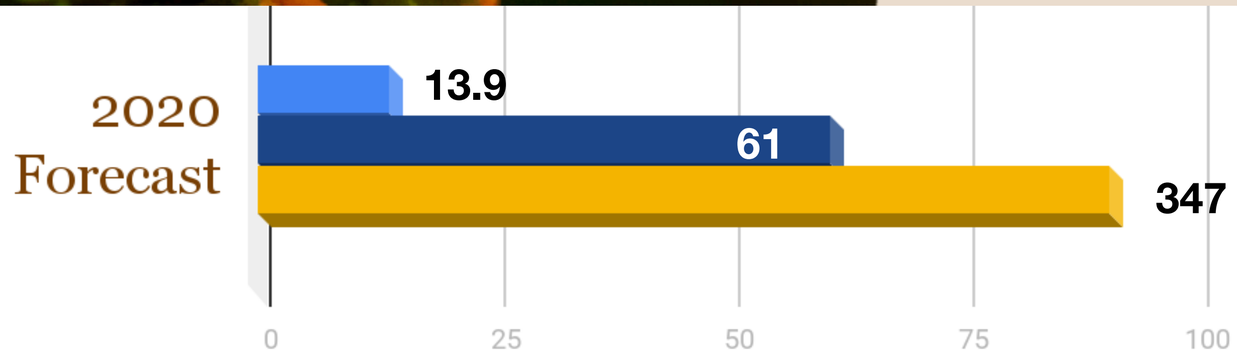
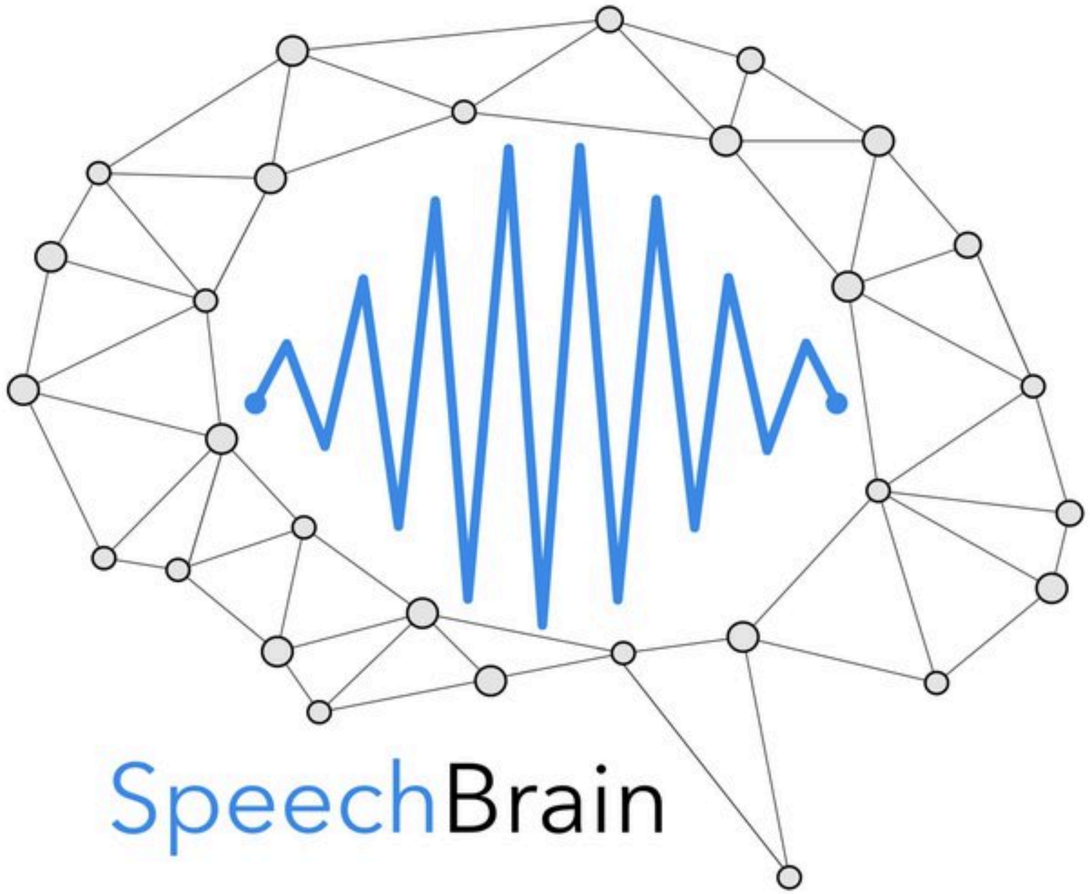


Image from: https://coqui.ai/

Exciting time to do speech research

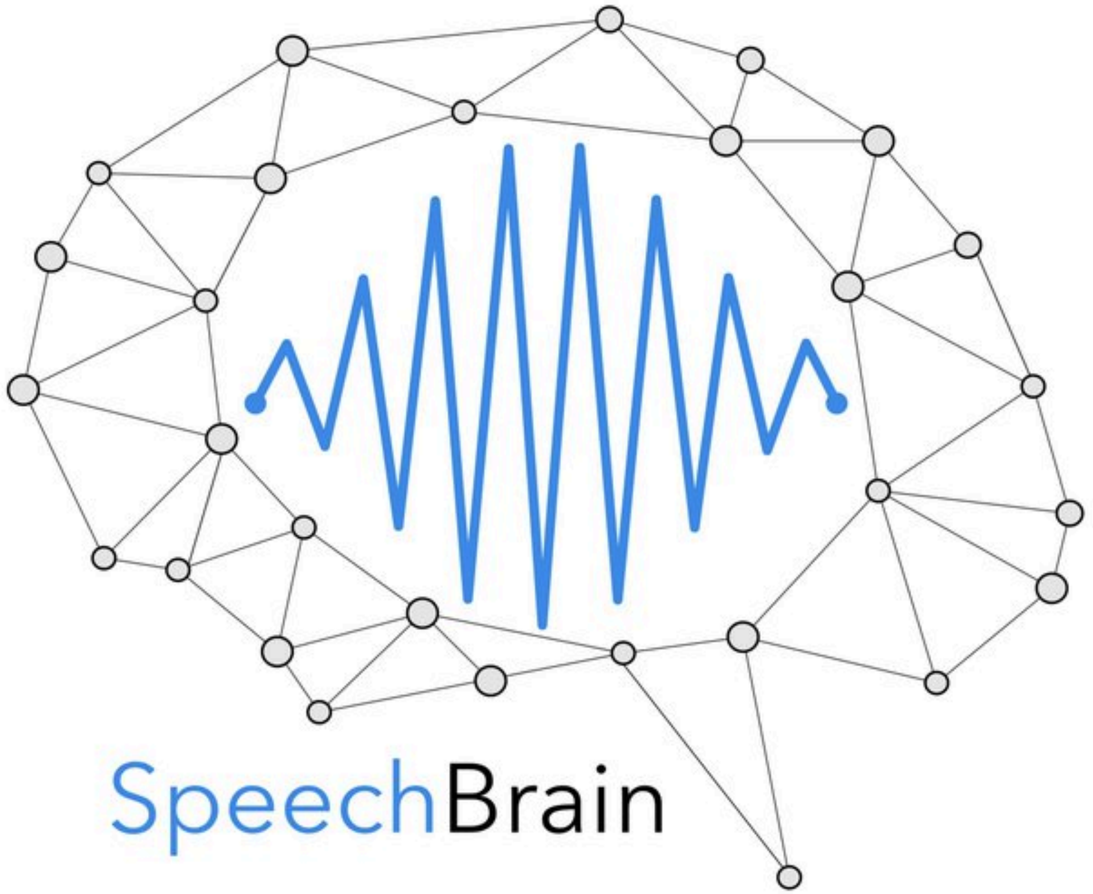


Coqui, Freeing Speech

Coqui, a startup providing open speech tech for everyone 🐸
Sign up with your email address to receive the Coqui newsletter.

Image from: https://coqui.ai/
https://speechbrain.github.io/

Exciting time to do speech research



Coqui, Freeing Speech

Coqui, a startup providing open speech tech for everyone 🐸
Sign up with your email address to receive the Coqui newsletter.

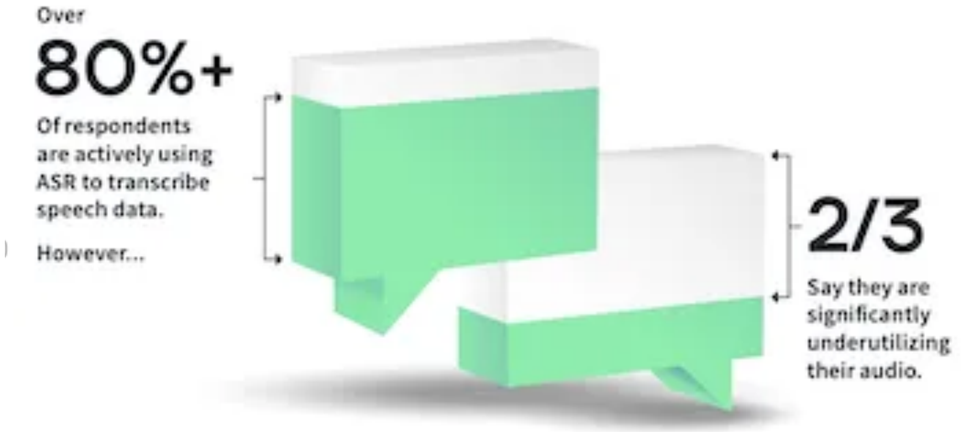


Image from: https://coqui.ai/

https://speechbrain.github.io/

https://opusresearch.net/wordpress/2021/03/03/opus-research-report-2021-state-of-automatic-speech-recognition/

What's next?

Need to do more…

- Robust to variations in age, accent and ability
- Handling noisy real-life settings with many speakers (e.g., meetings, parties)
- Handling pronunciation variability
- Handling new languages/ dialects

Accented Speech Recognition

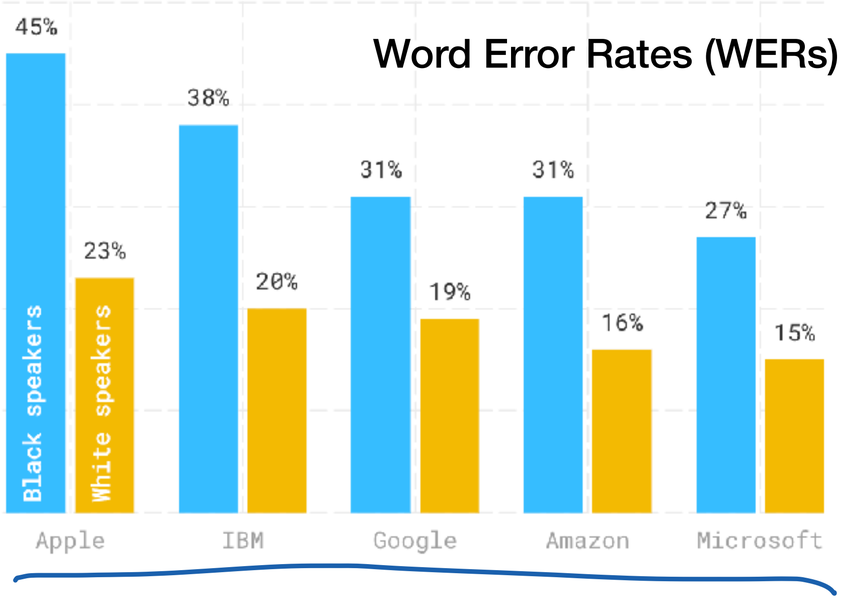
- “Voice is the next big platform, unless you have an accent” Wired’17

Wired’17 https://www.wired.com/2017/03/voice-is-the-next-big-platform-unless-you-have-an-accent/

Accented Speech Recognition

- “Voice is the next big platform, unless you have an accent” Wired’17

  Non-native accents still pose a significant challenge to state-of-the-art ASR systems



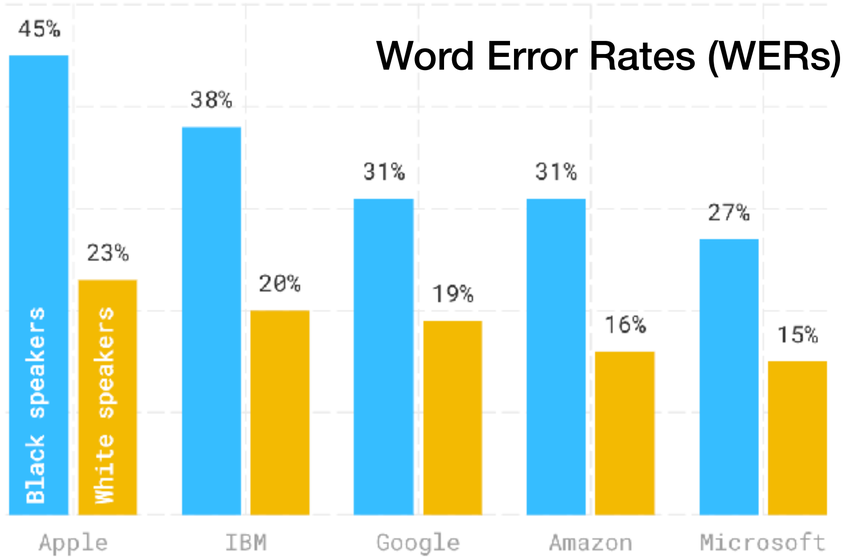
Wired’17 https://www.wired.com/2017/03/voice-is-the-next-big-platform-unless-you-have-an-accent/

Image from https://fairspeech.stanford.edu/, 2020

Accented Speech Recognition

- "Voice is the next big platform, unless you have an accent" Wired'17

  Non-native accents still pose a significant challenge to state-of-the-art ASR systems



- WERs range from 11% to 53% across different Indian accented English speech test sets, using a state-of-the-art ASR system AKSJ'21

Wired'17 https://www.wired.com/2017/03/voice-is-the-next-big-platform-unless-you-have-an-accent/

Image from https://fairspeech.stanford.edu/, 2020

AKSJ'21 Awasthi et al., Error-driven Fixed-Budget ASR Personalization for Accented Speakers, ICASSP 2021

What's next?

Need to do more…

- Robust to variations in age, accent and ability
- Handling noisy real-life settings with many speakers (e.g., meetings, parties)
- Handling pronunciation variability
- Handling new languages/ dialects

… with less

- Fast (real-time) decoding using limited computational power/ memory
- Faster training algorithms
- Reduce duplicated effort across domains/languages
- Reduce dependence on language-specific resources
- Train with less labeled data

Participation Points

Quiz	Points	# of responses
1	5	193
2	7	154
3	6	156
4	5	143

Participation Points

- Four Moodle-quizzes

Quiz	Points	# of responses
1	5	193
2	7	154
3	6	156
4	5	143

Participation Points

- Four Moodle-quizzes
- ≥50% points across 4 quizzes gets 5 participation points

Quiz	Points	# of responses
1	5	193
2	7	154
3	6	156
4	5	143

Participation Points

- Four Moodle-quizzes
- ≥50% points across 4 quizzes gets 5 participation points
- [40-50) — 4
  [30-40) — 3
  [20-30) — 2
  [10-20) — 1
  < 10 — 0

Quiz	Points	# of responses
1	5	193
2	7	154
3	6	156
4	5	143

Participation Points

- Four Moodle-quizzes

- ≥50% points across 4 quizzes gets 5 participation points

- [40-50) — 4
  [30-40) — 3
  [20-30) — 2
  [10-20) — 1
  < 10 — 0

Quiz	Points	# of responses
1	5	193
2	7	154
3	6	156
4	5	143

- Subtract x from your participation points if you do not satisfy the seminar attendance requirement
  - x = 1 if you attended 3 or 4
  - x = 2 if you attended 1 or 2
  - x = 3 if you attended none

Final Exam (35 points)

1. WFST algorithms
2. WFSTs used in ASR
3. HMM algorithms/EM/Tied-state Triphone models
4. DNN-based acoustic models/TDNNs
5. N-gram LMs/Smoothing/RNN LMs
6. End-to-end ASR (CTC, LAS)
7. MFCC feature extraction
8. Search & Decoding
9. Speech Representations

Final Exam (35 points)

1. WFST algorithms
2. WFSTs used in ASR
3. HMM algorithms/EM/Tied-state Triphone models
4. DNN-based acoustic models/TDNNs
5. N-gram LMs/Smoothing/RNN LMs
6. End-to-end ASR (CTC, LAS)
7. MFCC feature extraction
8. Search & Decoding
9. Speech Representations

Questions can be asked on any of the 9 topics listed above. This will be an open-notes exam.

Final Exam (35 points)

Date and Time: April 28th, 17:30 pm - 20:30 pm

1. WFST algorithms
2. WFSTs used in ASR
3. HMM algorithms/EM/Tied-state Triphone models
4. DNN-based acoustic models/TDNNs
5. N-gram LMs/Smoothing/RNN LMs
6. End-to-end ASR (CTC, LAS)
7. MFCC feature extraction
8. Search & Decoding
9. Speech Representations

Questions can be asked on any of the 9 topics listed above. This will be an open-notes exam.

Seminar: Presentation + Review + Scientific Article

Seminar: Presentation + Review + Scientific Article

- Presentation scores (out of 8) will be released this week

Seminar: Presentation + Review + Scientific Article

- Presentation scores (out of 8) will be released this week

- Best presentation: **PhoeniTiX (Efficient Vocalize): T Pavan Kalyan, Pranamya Prashant Kulkarni, Arif Ashfaque Ahmad**

Seminar: Presentation + Review + Scientific Article

- Presentation scores (out of 8) will be released this week

- Best presentation: **PhoeniTiX (Efficient Vocalize): T Pavan Kalyan, Pranamya Prashant Kulkarni, Arif Ashfaque Ahmad**

- Honorable Mentions (in no particular order):
  AudioAdversary (Unmuters) /Visual Voice (DeepSpeech) /Mask CTC (angry_nerds)/ AVID (Tri-Vachan) /SpeakQL (235161) /Large margin SV (Speech Hackers)

Seminar: Presentation + Review + Scientific Article

- Presentation scores (out of 8) will be released this week
- Best presentation: **PhoeniTiX (Efficient Vocalize): T Pavan Kalyan, Pranamya Prashant Kulkarni, Arif Ashfaque Ahmad**
- Honorable Mentions (in no particular order):
  AudioAdversary (Unmuters) /Visual Voice (DeepSpeech) /Mask CTC (angry_nerds)/ AVID (Tri-Vachan) /SpeakQL (235161) /Large margin SV (Speech Hackers)
- Review will be **due on Moodle on or before midnight of April 15th**.

Seminar: Presentation + Review + Scientific Article

- Presentation scores (out of 8) will be released this week
- Best presentation: **PhoeniTiX (Efficient Vocalize): T Pavan Kalyan, Pranamya Prashant Kulkarni, Arif Ashfaque Ahmad**
- Honorable Mentions (in no particular order):
  AudioAdversary (Unmuters) /Visual Voice (DeepSpeech) /Mask CTC (angry_nerds)/ AVID (Tri-Vachan) /SpeakQL (235161) /Large margin SV (Speech Hackers)
- Review will be **due on Moodle on or before midnight of April 15th**.
- Scientific article will be due first week of May (as per poll). Submission portal will be open this week for those who want to submit by April 17th.

Seminar: Poster Presentation + Hacker

Seminar: Poster Presentation + Hacker

- Poster presentation will be held online on May 6, 7, 8. Specifics will be conveyed via Moodle.

- Hacker presentations will also be held on the same 3 dates (May 6, 7, 8). For the Hacker role:

  - ✓ Share a link to your GitHub repository.

  - ✓ Each team describes what they have implemented and how it relates to the paper they were assigned.

  - ✓ Short viva-voce.

  - ✓ Note there is no report or write-up associated with the hacker role.