Speech

Multilingual Speech Representations

2025

Bhuvana Ramabhadran

Contributors

Speech / Research teams in Google

Recent Talks and Special Sessions at Interspeech 2024

- Hung-yi Lee, “Development of Spoken Language Models“
- David Harwath, “Visually grounded speech models “
- **Rohit Prabhavalkar, “Novel architectures for ASR “**
- Shinji Watanabe, “Toward speech and audio foundation models”

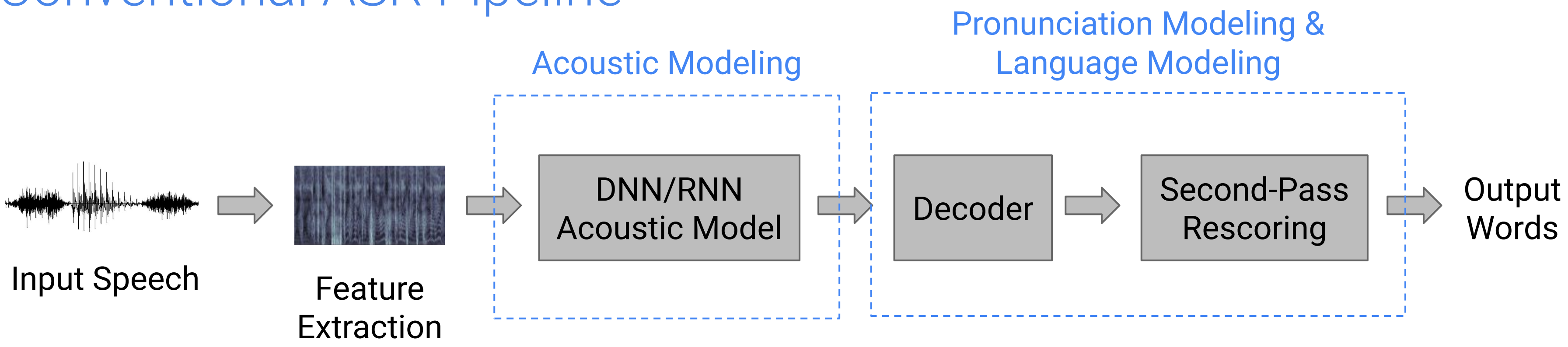
- Speech Processing Using Discrete Speech Units (SS10)
- Satellite workshop: SynData4GenAI 2024 (Aug. 31, 2024)

.... and many more in the conference !

Adapted from Rohit's Interspeech 2024 survey talk

Brief Introduction to ASR Architectures

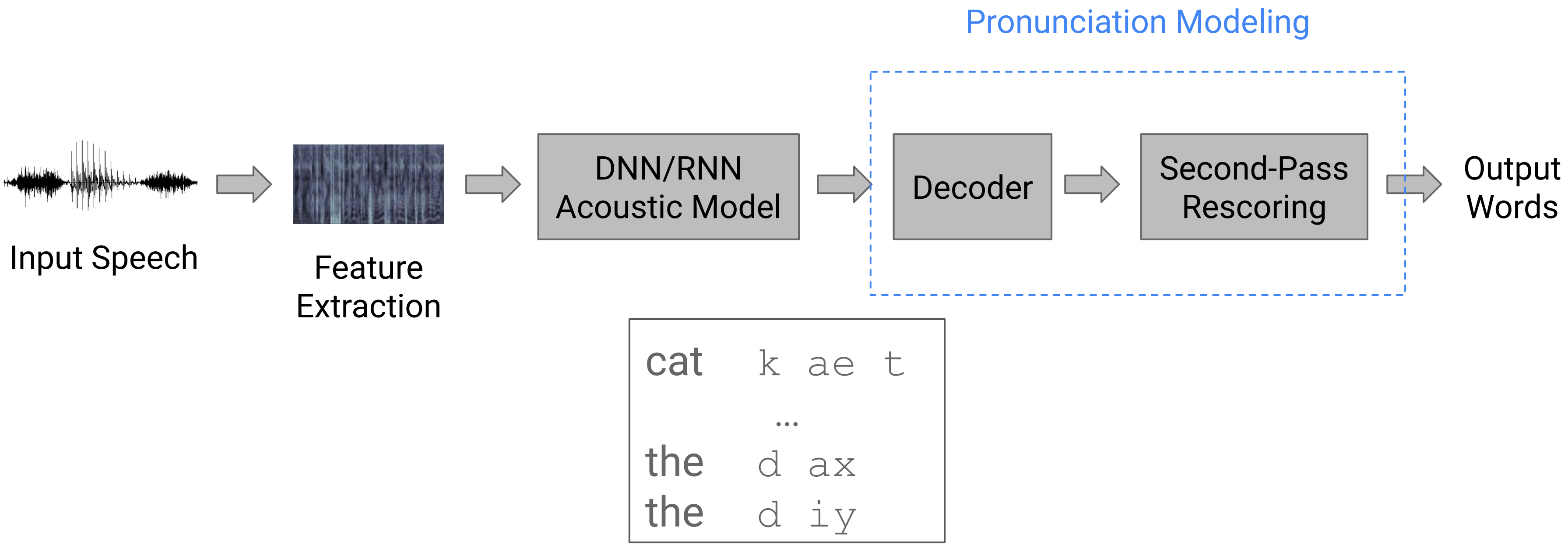
Conventional ASR Pipeline



e.g., [Rabiner,89] [Mohri+,02]

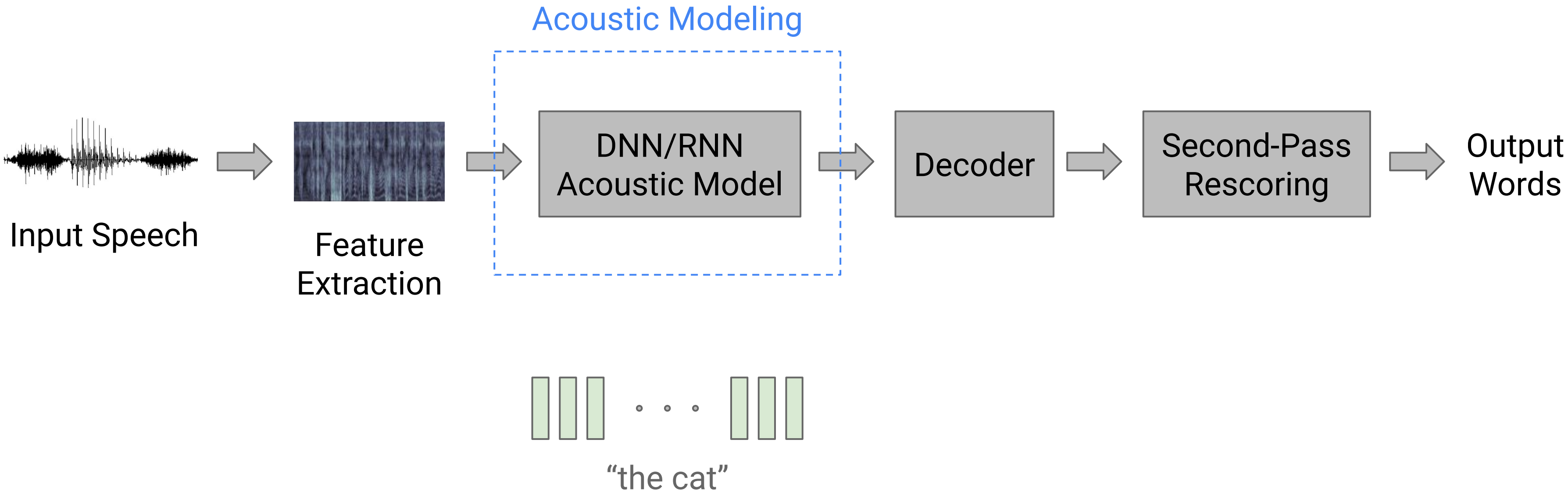
The “hybrid” ASR pipeline breaks the overall problem down into modular tasks: acoustic, pronunciation, and language modeling

Conventional ASR Pipeline



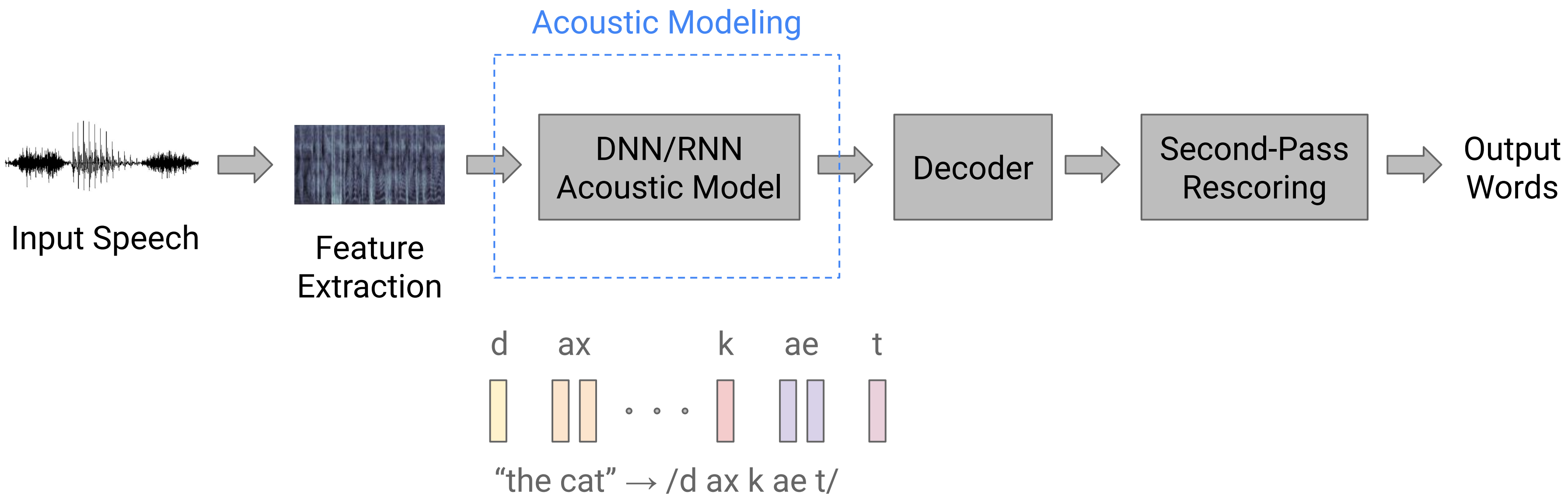
A pronunciation lexicon lists the pronunciation of words in terms of (phonemic) acoustic units; usually done through an expensive manual curation process

Conventional ASR Pipeline



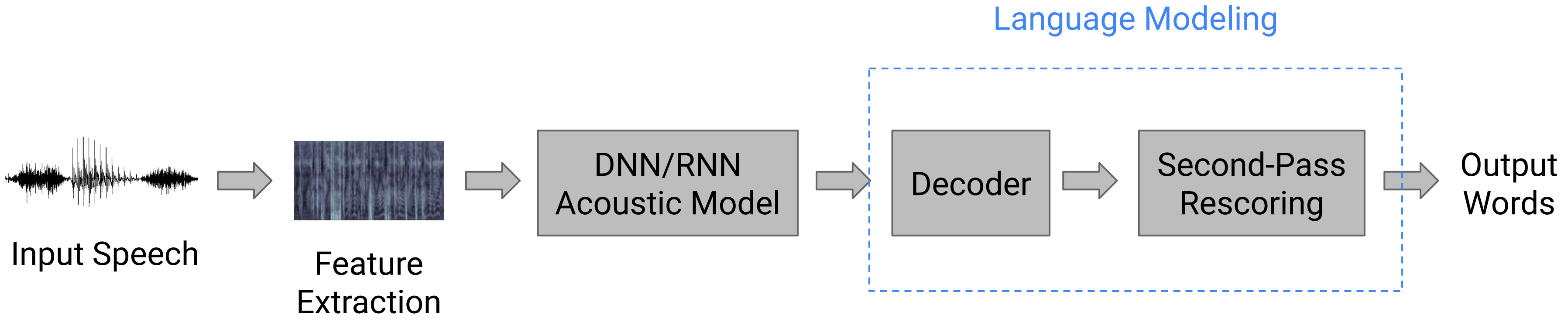
Acoustic modeling learns to associate acoustic feature vectors with the corresponding phonetic acoustic units

Conventional ASR Pipeline



Acoustic modeling learns to associate acoustic feature vectors with the corresponding phonetic acoustic units

Conventional ASR Pipeline

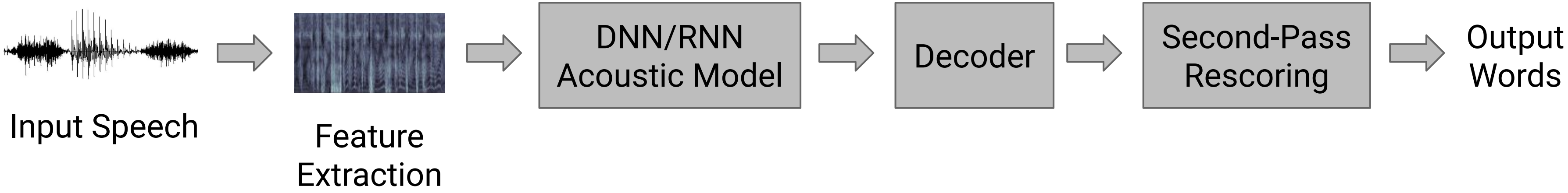


the cat sat ✓

sat cat the ✗

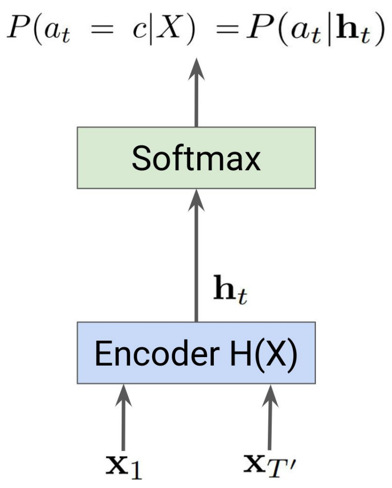
Language modeling assigns probabilities to word sequences, to model prior beliefs of the likelihoods of various sequences

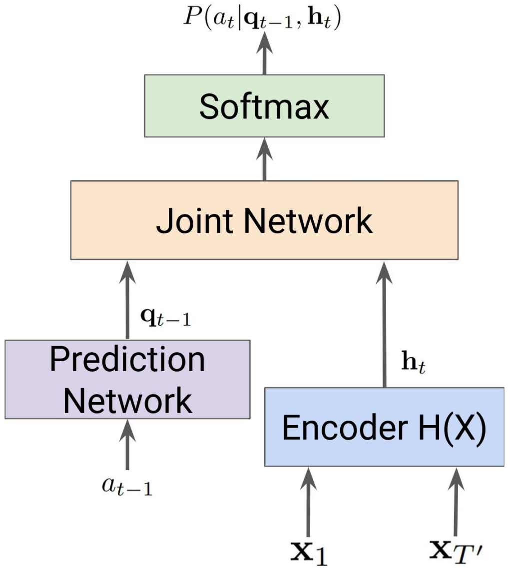
Conventional ASR Pipeline

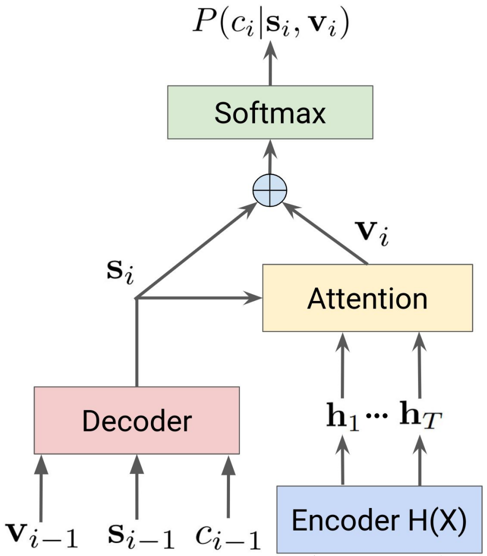


The pronunciation + language model can be represented efficiently in a finite-state transducer, with generic and efficient algorithms to model the search process.
System is modular and is extensible (e.g., adaptation to new domains)

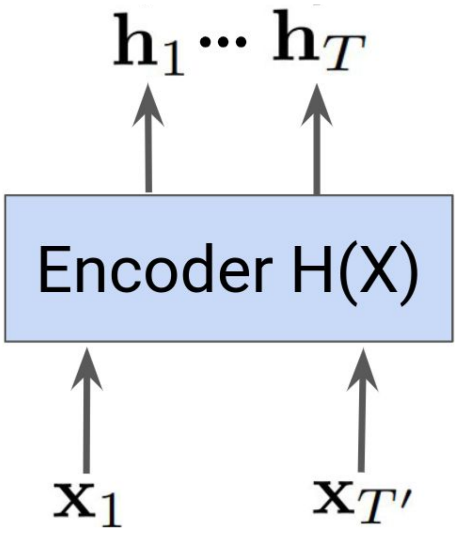
E2E Approaches







Encoders



What is the role of the speech encoder?

Synthesizing context and learning representations

Representation learning

Is there a joint latent representation of multiple modalities that can help multilingual speech and language understanding?

- Speech Understanding (Speech ⟶ Text with speaker/language/style annotations)
  - Scaling to many languages
  - Taking advantage of found data
  - Emotional, Medical diagnosis, atypical speech processing, perception
- Audio Generation (Text ⟶ Speech)
  - Can we share the same ideas from ASR?
  - Understanding shared representations of multiple modalities key?
- Language Understanding
  - Question Answering, Dialog / Conversations, Summarization, etc.

01

Representation Learning

Self-supervised Learning allows for the efficient use of unlabeled data (audio, text, image and video) in models with the promise a single universal model that would benefit a wide variety of tasks and domains.

Historically, in speech and language processing, semi-supervised and unsupervised learning has been used extensively (BABEL program)

Cui, Jia, et al. "Multilingual representations for low resource speech recognition and keyword search." 2015 IEEE workshop on automatic speech recognition and understanding (ASRU). IEEE, 2015.

Mohamed, Abdelrahman, Hung-yi Lee, Lasse Borgholt, Jakob D. Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff et al. "Self-supervised speech representation learning: A review." *IEEE Journal of Selected Topics in Signal Processing* 16, no. 6 (2022): 1179-1210.

Machine Learning has focussed on learning speech representations for a variety of tasks in an unsupervised fashion and combined with supervised training for maximum results

Diversity of the data used for unsupervised learning plays a key role

Self Supervision

Model learns discriminating patterns in the data in contrast to providing the model with annotations and asking the model to be sensitive to those specific annotations.

- Cost function
- Adversarial Methods
- Supervised Fine tuning: tailoring learned representations to downstream tasks
- Architectures
- Robustness and transferability of speech representations with few examples
- Injecting other modalities

Self Supervision: Cost function

- Contrastive Predictive Coding (CPC) [1], Autoregressive Predictive Coding (APC) [2], Triplet Loss based approaches[15], Time Contrastive Networks [16]
  - Encode underlying shared information in the high-dimensional signal
  - Maximize Mutual Information between encoded representations
  - Combine predicting future observations (predictive coding) with a probabilistic contrastive loss (NCE variants)

Self Supervision:Cost function

- Selecting positive and negative samples
  - Sampling from joint versus marginalized distributions
  - Momentum Contrast (MoCo)in Vision[9] and Speech[17]
- Multiple tasks (views) for a more complete representation
  - Examples: Regressive, classification tasks
  - Multi task learning and self-supervision
    - Task Agnostic Speech Embeddings (PASE, PASE+) [7]

Self Supervision: Cost function

- Unsupervised Latent Variable Model based data generation [4, 5]
  - Clustering latent representations, Unsupervised Unit Discovery
    - Codebook learning, VQVAE

- wav2vec/ wav2vec2 : Combine contrastive loss and discrete representations [10]

- HuBERT: Iterative learning of discretized representations (k-means clustering) and representation learning [21]

- W2V-BERT [22]: Combines contrastive loss on continuous signal with masked language modeling (MLM) loss on the discretized representations.

Self Supervision: *Adversarial /Augmentation Methods*

- SimCLR [3]: input augmentation is coupled with a contrastive consistency loss to allow model to learn without any labels.

- Virtual adversarial training(VAT) is a form of model regularization that applies adversarial noise to the model input.

- Masking and Reconstruction
  - Augmentation methods (SpecAugment) to learn invariant representations [6]

Speech-only Self-supervised Pretrain

Wav2vec 2.0 has shown to be an effective encoder pre-training strategy, especially on in-domain data.

State of the art results when combined with Conformer encoders.

Leverages large amounts of untranscribed data without a good teacher model.

Nevertheless, Cannot take advantage of **unspoken text.**

Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." arXiv preprint arXiv:2006.11477 (2020).
Hsu, Wei-Ning, et al. "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training." arXiv preprint arXiv:2104.01027 (2021).
Zhang, Yu, et al. "Pushing the limits of semi-supervised learning for automatic speech recognition." arXiv preprint arXiv:2010.10504 (2020).

Self Supervision in Speech and Audio

Text Injection[24-36]

- Injecting Text In Self-Supervised Speech Pre-Training
  - Combining supervised and unsupervised loss and TTS

- Joint training of speech and text using text, speech and multimodal encoders

Self Supervision: Supervised Fine Tuning

- Joint training of encoder and decoders (predictors) as proposed in PASE and PASE+ [7]
- Strategies to freeze different parts of the network during fine tuning on a small amount of labeled data
- Meta Learning: Self supervision and fine-tuning on few samples [18]
- BERT [8] and its variants used in Natural Language Processing tasks
- Foundation Models [23] to capture representations that can be fine-tuned for tasks such as, object recognition, image captioning, information retrieval, etc.

Self Supervision:Architectures and Transferability

- Encoders: Transformers, VGGs, Conformers, SincNet [19], etc.
- Multilingual representations and bottleneck features from various architectures
- Are the losses used in speech and language transferable to vision or robotics and vice-versa [20] ?

02
Multilinguality

Multilingual work dates back to 90s and earlier.....

- Multilinguality refers to the ability to handle several languages for different tasks (recognition, translation, synthesis, etc.)
- More recently, training multilingual representations [1, 2] and end-to-end models [3, 4] have demonstrated that the best performing models require conditioning on language information

  - [1] B. Ma, C. Guan, H. Li, and C.-H. Lee, “Multilingual speech recognition with language identification,” in ICSLP, 2002.
  - [2] A. Cutler, Y. Zhang, E. Chuangsuwanich, and J.R. Glass, “Language ID-based training of multilingual stacked bottleneck features,” in Interspeech, 2014.
  - [3] S. Watanabe, T. Hori, and J.R. Hershey, “Language independent end-to-end architecture for joint language identification and speech recognition,” in ASRU, 2017.
  - [4] A. Kannan, A. Datta, T.N. Sainath, E. Weinstein, B. Ramabhadran, Y. Wu, A. Bapna, Z. Chen, and S. Lee, “Large-scale multilingual speech recognition with a streaming end-to-end model,” arXiv preprint arXiv:1909.05330, 2019.

Key requirement

- Need to track language switches within an utterance [5, 6], adjust language sampling ratios, or add additional parameters based on the data distribution [4]

    - [5] H. Seki, S. Watanabe, T. Hori, J. Le Roux, and J.R. Hershey, “An end-to-end language-tracking speech recognizer for mixed-language speech,” in ICASSP, 2018.
    - [6] A. Waters, N. Gaur, P. Haghani, P. Moreno, and Z. Qu, “Leveraging language id in multilingual end-to-end speech recognition,” in ASRU, 2019.

What language clusters and why?

For example, many Indic languages can we cover with one model?

- Take advantage of overlap in acoustic and lexical content
  - due to either language family relations or the geographic and cultural proximity of the native speakers.
- However, their writing systems occupy different unicode blocks
- Can we combine languages from multiple languages families efficiently and produce “usable” models for users?
- Can the representations derived by these models help with building models for “unseen” languages?

...what challenge does this pose?

Challenges: Code-Switching

- Code-switching is a commonly occurring phenomenon in many multilingual communities, wherein a speaker switches between languages within a single utterance (Hindi-English, Bengali-English, Arabic-English and Chinese-English, Spanish-English, etc.)
- Can occur at morphological, lexical, syntactic, semantic, pragmatic levels
- A good read on Bilingual Speech from a linguistic perspective:
  - Analysis of many language-pairs
  - Bilingual verbs: the phenomenon of verbal compounds combining elements from two languages
  - Impact of psycholinguistic and social factors : language dominance, duration of contact, bilingual proficiency, speaker type, age-group or generation and language attitudes.

Pieter Muysken, Bilingual speech: A typology of code-mixing. Cambridge: Cambridge University Press, 2000.

Examples of code-switching

- Words with different language indices are inserted into a phrase structure
- Spanish-English
    - Cuando mi novio *tweet*ea pero no contesta (When my boyfriend tweets but doesn’t answer)
    - Agarrar *my Master’s* (Get my Master’s)
- Ambiguities in transcription
    - *डिस्कवरी vs discovery*
    - *होम्योपथी में अर्थराइटिस treatment vs Homeopathy में arthritis treatment*
- These *rendering* errors artificially inflate the **W**ord **E**rror **R**ate (WER)
- Harder to differentiate between ***modeling*** and ***rendering*** errors
    - *fancy साड़ी दिखाइए vs fancy Sadi dikhaiye*

Handling Code-Switching

- Handled the problem of foreign word pronunciation using language dependent phonemes by creating linguistically motivated pairwise mappings for each language involved in code-switching.

  White, Christopher M., Sanjeev Khudanpur, and James K. Baker. "An investigation of acoustic models for multilingual code-switching." *Ninth Annual Conference of the International Speech Communication Association*. 2008.

- In Mandarin-English use of combined subwords from both languages as modeling units along with an additional objective of training with language ID was found to be useful.

  Luo, Ne, et al. "Towards end-to-end code-switching speech recognition." *arXiv preprint arXiv:1810.13091* (2018).

Handling Code-Switching

- Separately train an E2E CTC model and a frame-level language identification (LID) model. Linearly adjust the posteriors of an E2E CTC model using the LID scores (Mandarin-English)

  Li, Ke, et al. "Towards code-switching ASR for end-to-end CTC models." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

- Effectiveness of multilingual models on NLU tasks such as named entity recognition and part-of-speech tagging tasks (Hindi-English, Spanish-English, and Modern Standard Arabic-Egyptian)? Pretrained multilingual models not as effective as hierarchical embeddings to deal with code-switching

  White, Christopher M., Sanjeev Khudanpur, and James K. Baker. "An investigation of acoustic models for multilingual code-switching." Ninth Annual Conference of the International Speech Communication Association. 2008.

Handling Code-Switching

- In Frisian-Dutch merging phones of both languages provides the best recognition performance for code-switched words

  Yılmaz, Emre, Henk van den Heuvel, and David Van Leeuwen. "Investigating bilingual deep neural networks for automatic recognition of code-switching frisian speech." Procedia Computer Science 81 (2016): 159-166.

- Data Augmentation by generating synthetic code-switched data with word translation or word insertion followed by audio splicing using text-to-speech

  Du, Chenpeng, et al. "Data augmentation for end-to-end code-switching speech recognition." *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021.

Handling Code-Switching

Output token embeddings of two monolingual languages are differently distributed; Constrain with Jensen-Shannon divergence to force embeddings of monolingual languages to possess similar distributions

Khassanov, Yerbolat, et al. "Constrained output embeddings for end-to-end code-switching speech recognition with only monolingual data." *arXiv preprint arXiv:1904.03802* (2019).

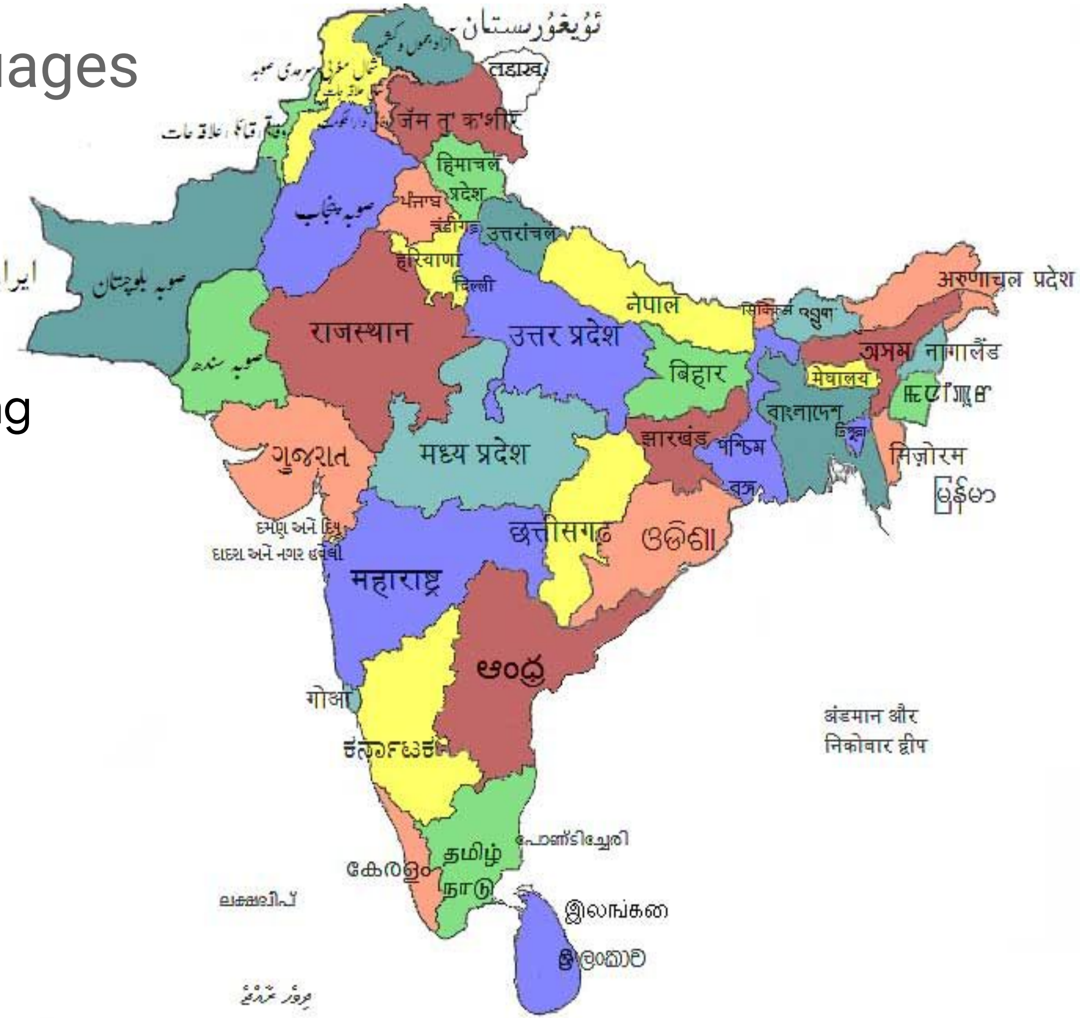
What are these techniques forcing the model to learn?

Joint multilingual representations (embeddings) at multiple levels?

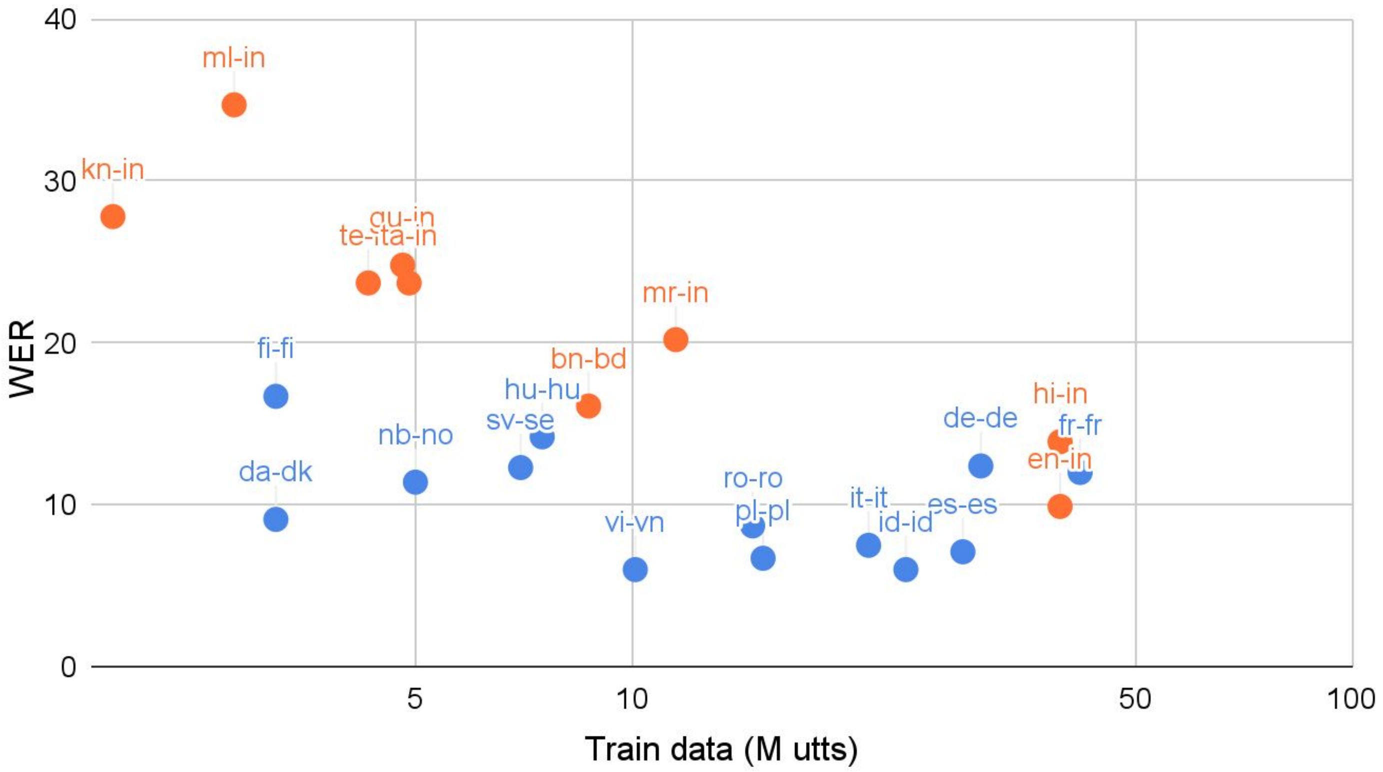
Learn implicit patterns in the data across languages?

South Asia: A land of languages

- **Scripts**: Several writing systems
- **Code-switching:** language mixing
- **Several** languages and dialects!
- But **overlap** due to linguistic similarity, and/or geographic & cultural proximity of the native speakers.



Data Distribution across languages



Pooling Language Resources to learn representations

- State-of-the-art
  - Allow for joint training of data-rich and data-scarce languages in a single model
  - Require the encoding of language information which makes it less flexible

- Challenges in building a language-agnostic multilingual ASR system?
  - Can similar sounding acoustics across languages be mapped to a single, canonical target sequence of graphemes or sub-word units?

Add maestro paper

Data Normalization: Challenge in multilingual transliteration

Attested romanizations of the English word "discovery"

Bengali ডিসকভারি	Hindi डिस्कवरी	Kannada ಡಿಸ್ಕವರಿ	Tamil டிஸ்கவரி
discoveri	discovery	discovary	tiskavari
discovery		discovery	discovery
diskovary		discoveri	
diskovery		discowery	
diskoveri			

Code-Switching Benchmark: For NLP research (https://ritual.uh.edu/lince/)

LinCE is a continuous effort, and we will expand it with more low-resource languages and tasks.

Language Pairs	**LID**	**POS**	**NER**	**SA**	**MT**
Spanish–English	✓	✓	✓	✓	
Hindi–English	✓	✓	✓		
Nepali–English	✓				
Modern Standard Arabic–Egyptian Arabic	✓		✓		
English–Hinglish					✓
Spanglish–English					✓
English–Spanglish					✓
(Modern Standard Arabic–Egyptian Arabic)–English					✓
English–(Modern Standard Arabic–Egyptian Arabic)					✓

Text Representation Benchmark: (https://huggingface.co/spaces/mteb/leaderboard)

Massive Text Embedding Benchmark (MTEB) Leaderboard. To submit, refer to the MTEB GitHub repository 🤗 Refer to the MTEB paper for details on metrics, tasks and models. Also check out MTEB Arena ⚔️

Search Bar (separate multiple queries with `;`)

🔍 Search for a model and press enter...

Model types

- [x] Open
- [x] Proprietary
- [x] Sentence Transformers
- [x] Cross-Encoders
- [x] Bi-Encoders
- [x] Uses Instructions
- [x] No Instructions

Model sizes (in number of parameters)

- [x] <100M
- [x] 100M to 250M
- [x] 250M to 500M
- [x] 500M to 1B
- [x] >1B

Overall | Bitext Mining | Classification | Clustering | Pair Classification | Reranking | Retrieval | STS | Summarization | MultilabelClassification | Retrieval w/Instructions

English | Chinese | French | Polish | Russian

Overall MTEB English leaderboard 🔮

- **Metric:** Various, refer to task tabs
- **Languages:** English

Rank ▲	Model ▲	Model Size (Million Parameters) ▲	Memory Usage (GB, fp32) ▲	Embedding Dimensions ▲	Max Tokens ▲	Average (56 datasets) ▲	Classification Average (12 datasets) ▲	Clustering Average (11 datasets)
1	NV-Embed-v2	7851	29.25	4096	32768	72.31	90.37	58.46
2	bge-en-icl	7111	26.49	4096	32768	71.67	88.95	57.89
3	stella_en_1.5B_v5	1543	5.75	8192	131072	71.19	87.63	57.69
4	SFR-Embedding-2_R	7111	26.49	4096	32768	70.31	89.05	56.17
5	gte-Qwen2-7B-instruct	7613	28.36	3584	131072	70.24	86.58	56.92
6	dunzhang-stella_en_400M_v5	435	1.62	1024	8192	70.11	86.67	56.7
7	stella_en_400M_v5	435	1.62	8192	8192	70.11	86.67	56.7
8	bge-multilingual-gemma2	9242	34.43	3584	8192	69.88	88.08	54.65

Towards the future....

Can we have a similar code-switching only benchmark for speech across hundreds of languages ?

What tasks and associated metrics would help advance state-of-the-art?

Can these multilingual representations now be extended to do several tasks in one model? More than ASR?

03
Multitask and multilingual representations

State-of-the-art performance in ASR and ST tasks

- Efficient Pre-training
- Incorporating Untranscribed Speech, Unspoken Text, Paired Speech-Text
- Modality matching for in the Injection of unspoken text
- Language-ID
- Code-Switching

Bharadwaj, S., Ma, M., Vashishth, S., Bapna, A., Ganapathy, S., Axelrod, V., Dalmia, S., Han, W., Zhang, Y., van Esch, D. and Ritchie, S., 2023. Multimodal Modeling For Spoken Language Identification. arXiv preprint arXiv:2309.10567.

Google Universal Speech Model for 100+ Languages [43]

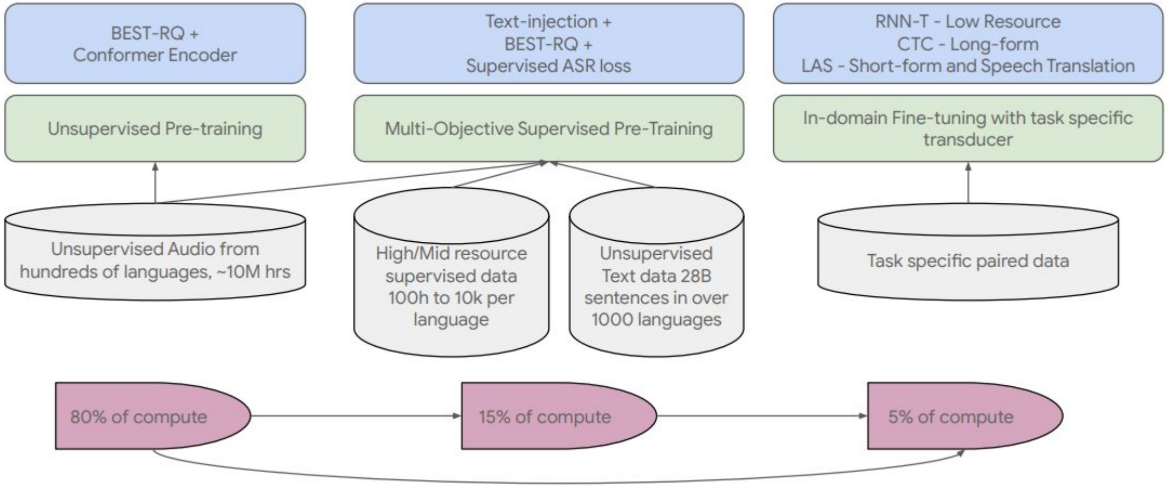


Figure 1: An overview of our approach. Training is split into three stages. (i) The first stage trains a conformer backbone on a large unlabeled speech dataset, optimizing for the BEST-RQ objective. (ii) We continue training this speech representation learning model while optimizing for multiple objectives, the BEST-RQ objective on unlabeled speech, the modality matching, supervised ASR and duration modeling losses on paired speech and transcript data and the text reconstruction objective with an RNN-T decoder on unlabeled text. (iii) The third stage fine-tunes this pre-trained encoder on the ASR or AST tasks.

Representations learnt during pre-training

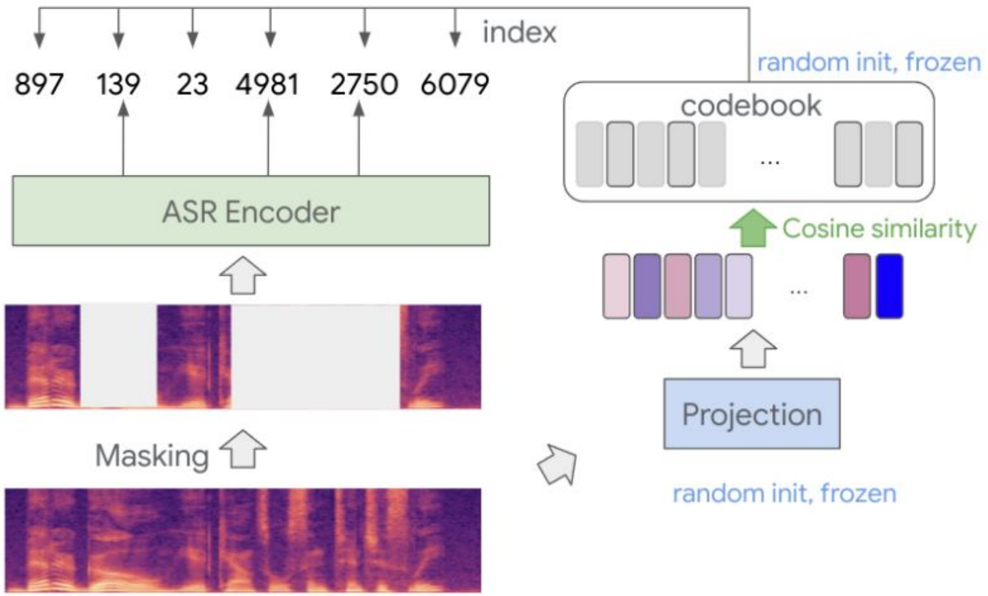


Figure 3: BEST-RQ based pre-training with conformer encoder.

BEST-RQ (BERT-based Speech pre-Training with Random projection Quantizer) is used to pre-train the encoder of the conformer model

Text-Injection and modality matching

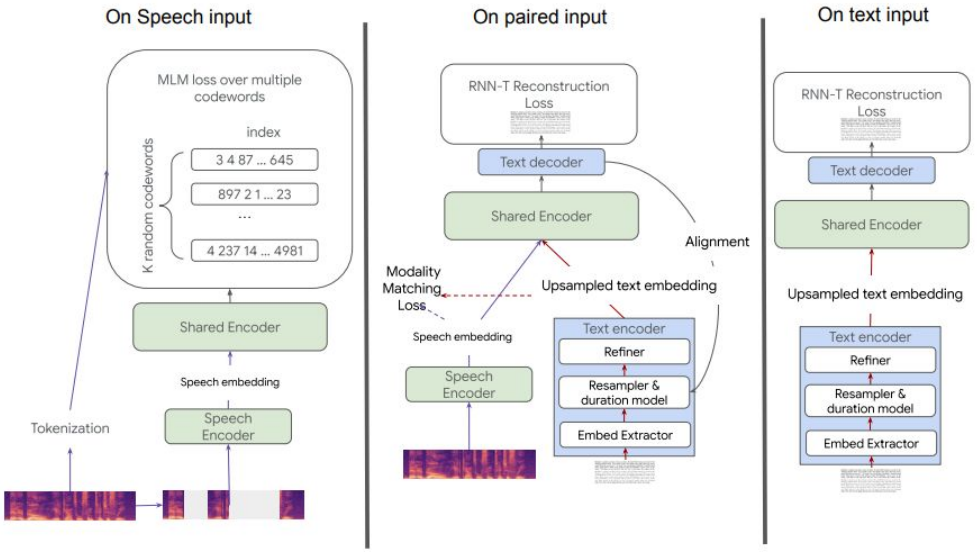


Figure 5: Overview of MOST text injection. The left-most panel depicts MOST training on unlabeled speech input; the center panel depicts training on paired speech and text input; the right-most panel depicts training on unlabeled text data.

Key Findings

- BEST-RQ is a scalable speech representation learner: We find that BEST-RQ pre-training can effectively scale to the very large data regime with a 2B parameter Conformer-based backbone.
- MOST (BEST-RQ + text-injection) is a scalable speech and text representation learner: It is an effective method for utilizing large scale text data for improving quality on downstream speech tasks, as demonstrated by quality gains exhibited for the FLEURS and CoVoST 2 tasks.
- Representations from MOST (BEST-RQ + text-injection) can quickly adapt to new domains with light-weight residual adapters.
- SoTA results for downstream multilingual speech tasks:
    - SpeechStew (mono-lingual ASR)
    - CORAAL (African American Vernacular English (AAVE) ASR)
    - FLEURS (multi-lingual ASR) [16], YT (multilingual long-form ASR)
    - CoVoST (AST from English to multiple languages).

Scalability: Language Expansion Results

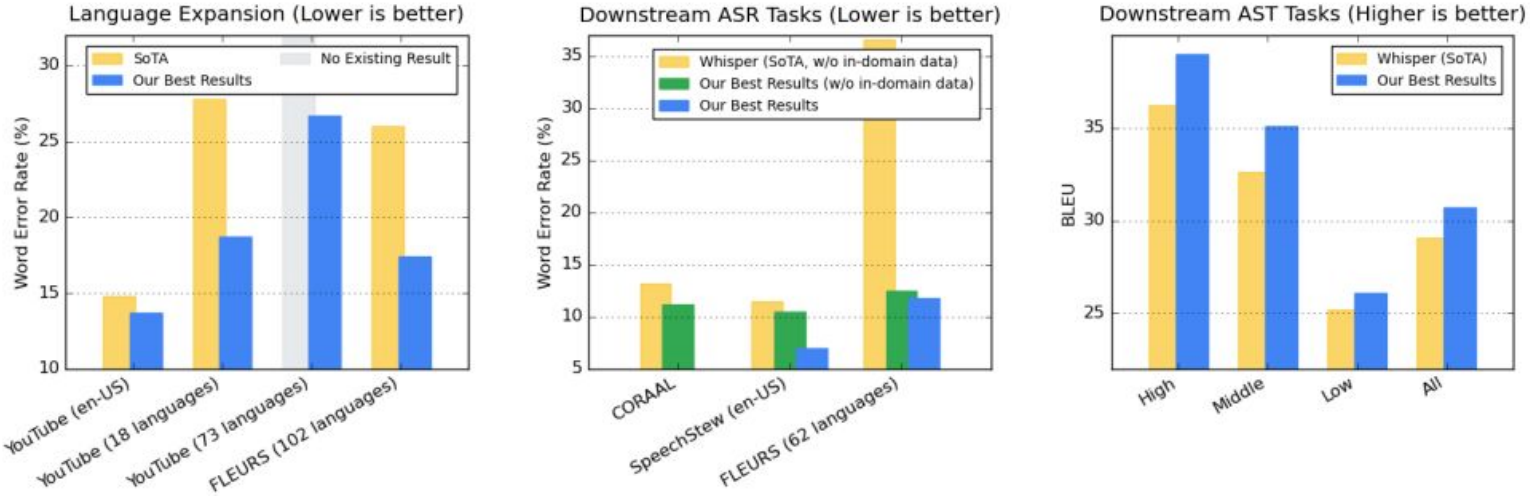


Figure 2: **(Left)**† WERs (%) Our language expansion effort to support more languages on YouTube (73 languages) and extending to 100+ languages on the public dataset (FLEURS). Lower is better. To the best of our knowledge, no published model can successfully decode all 73 languages from our YouTube set, thus we only list our results. **(Middle)**† Our results on ASR benchmarks, with or without in-domain data. Lower is better. **(Right)** SoTA results on public speech translation tasks. Results presented are presented as high/middle/low resources languages defined in [20]. Higher is better.

USM Results across ASR and ST tasks

Task	Multilingual Long-form ASR				Multidomain en-US	Multilingual ASR		AST
Dataset	YouTube			CORAAL	SpeechStew	FLEURS		CoVoST 2
Langauges	en-US	18	73	en-US	en-US	62	102	21
Prior Work (single model)								
Whisper-longform	17.7	27.8	-	23.9	12.8			
Whisper-shortform†	-	-	-	13.2‡	11.5	36.6	-	29.1
Our Work (single model)								
USM-LAS	14.4	19.0	29.8	**11.2**	**10.5**	**12.5**	-	-
USM-CTC	**13.7**	**18.7**	**26.7**	12.1	10.8	15.5	-	-
Prior Work (in-domain fine-tuning)								
BigSSL [3]	14.8	-	-	-	7.5	-	-	-
Maestro [67]					7.2			25.2
Maestro-U [67]							26.0 (8.7)	
Our Work (in-domain fine-tuning)								
USM	13.2	-	-	-	7.4	13.5	19.2 (6.9)	28.7
USM-M	**12.5**	-	-	-	**7.0**	**11.8**	**17.4 (6.5)**	**30.7**
Our Work (frozen encoder)								
USM-M-adapter§	-	-	-	-	7.5	12.4	17.6 (6.7)	29.6

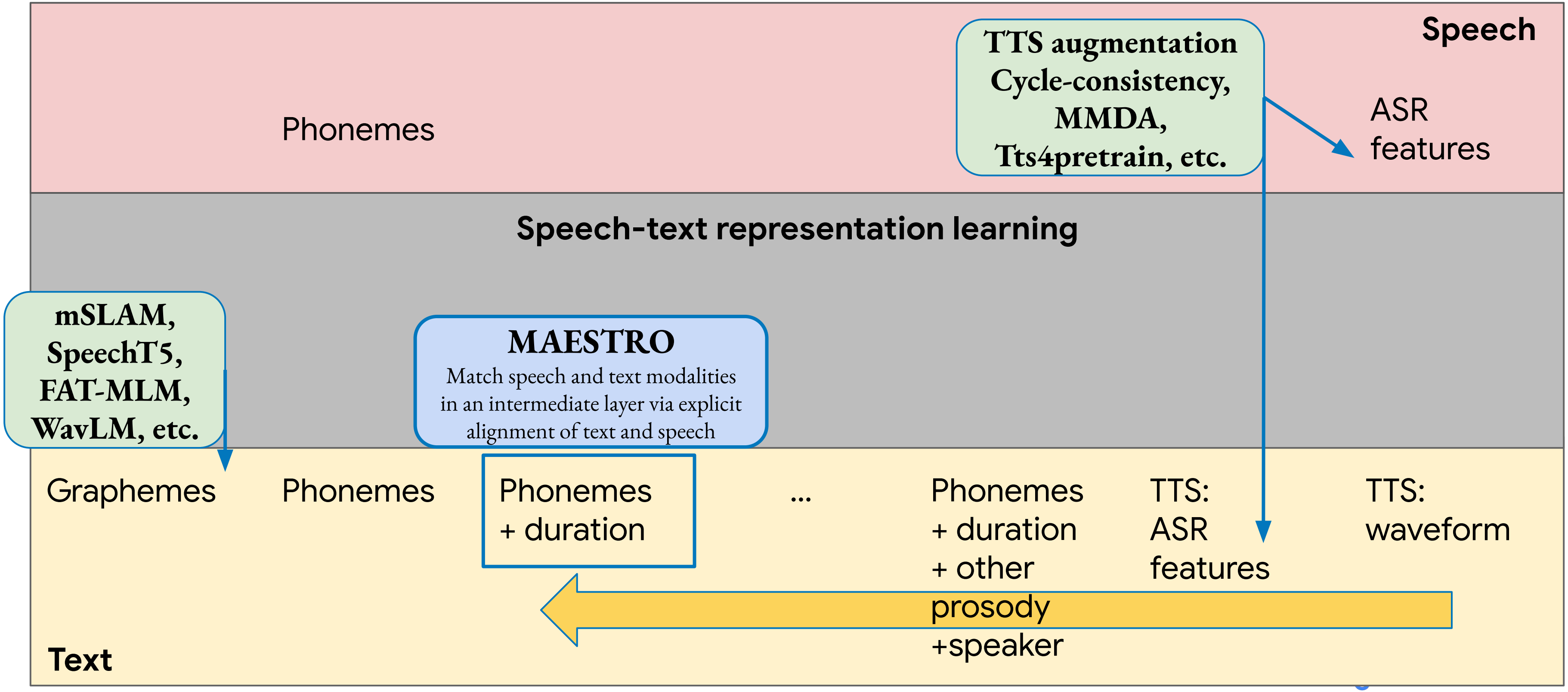
Speech-text representation learning

- **Complementary information** contained in Text and Speech[1]
  - **text**: domain; **speech**: acoustic conditions, speakers, etc.
- **Unify** speech and text representations
  - Simplify learning from both modalities
  - Learn better linguistic context in (conformer) encoders
- **Data minimization** by incorporating **unspoken text**
  - Low-resource speech processing

Cross-modality and Cross-lingual Knowledge Transfer

- **Maestro-U (ASR with zero-transcribed speech)**
  - Modality specific encoders feed a shared encoder.
  - Language specific adapters in the shared encoder.
  - Labeled speech for some languages
  - Only unpaired speech and unpaired text for some languages
  - NO LEXICON or G2P - Unicode Byte inputs support performance even on unseen scripts
- **Virtuoso (TTS with zero-transcribed speech)**
  - Similar approach but applied to TTS
  - Speech decoder (feature to spectrogram) doesn’t see any transcribed audio.
  - NO LEXICON or G2P - graphemic to acoustic form can be learned directly without explicit intermediate phone labels

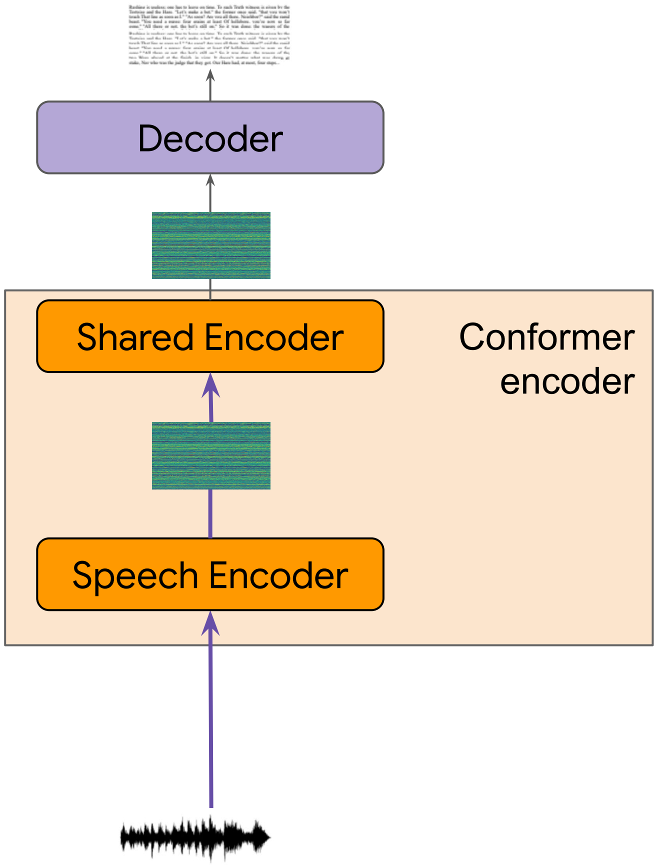
Joint speech+text representations



More related works can be referred to the paper.

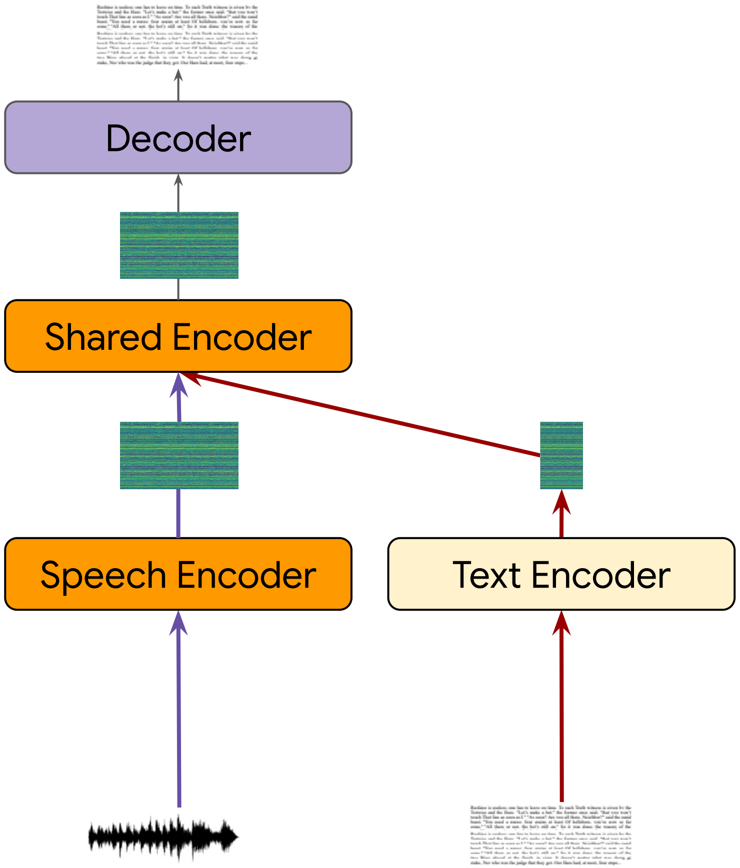
Architecture

Split original Encoder into two



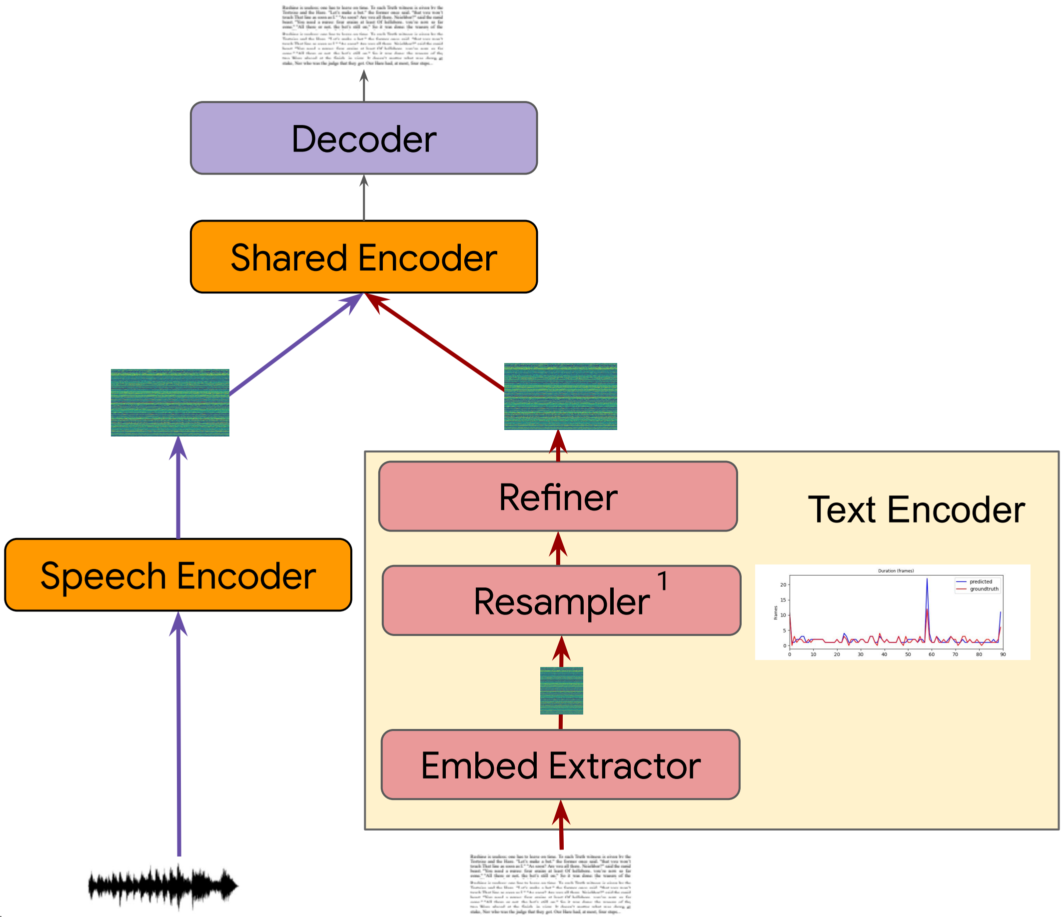
Architecture

Inject text representations in the middle



Architecture

How to match the two modalities?

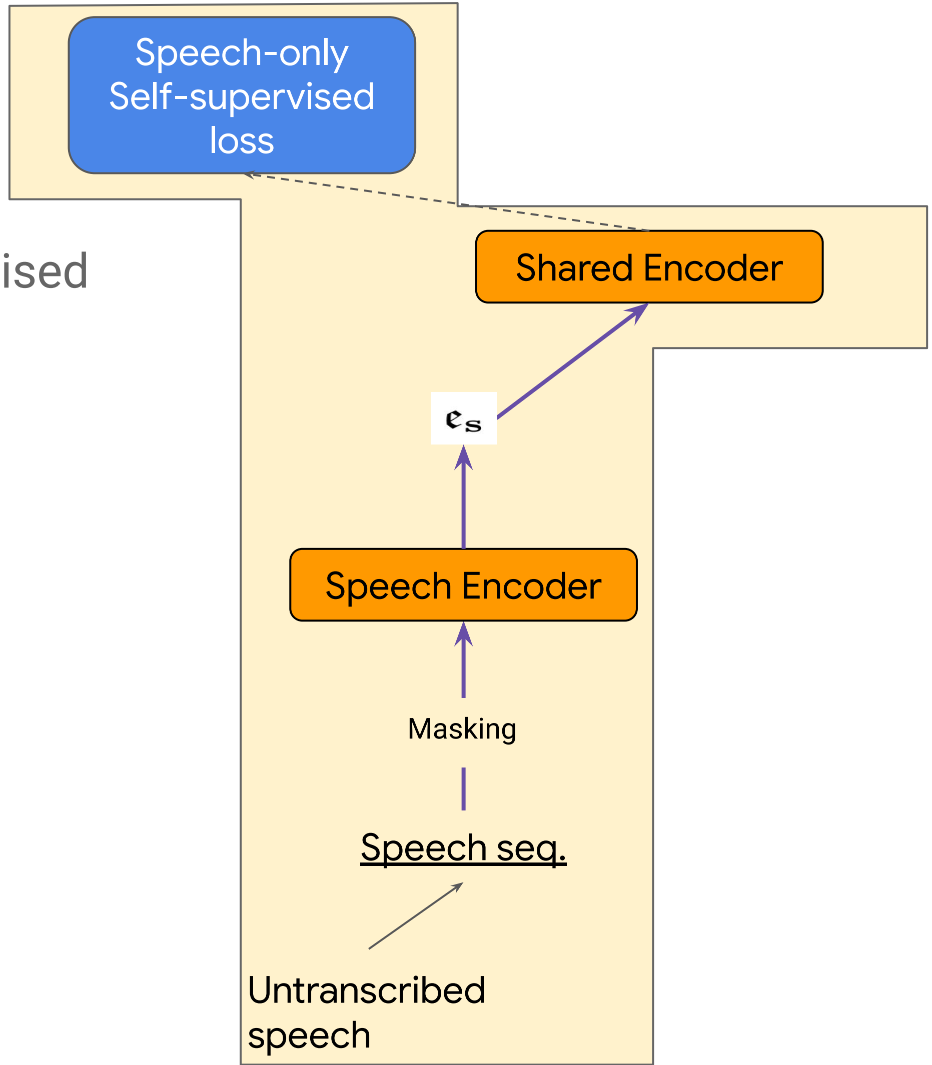


[1] Elias, Isaac, et al. "Parallel tacotron: Non-autoregressive and controllable tts." 2021.

Loss breakdown: **Speech-only**

Reuse any self-supervised pretraining objective

- W2v-BERT
- Best-RQ
- w2v1



Loss breakdown: Paired Speech

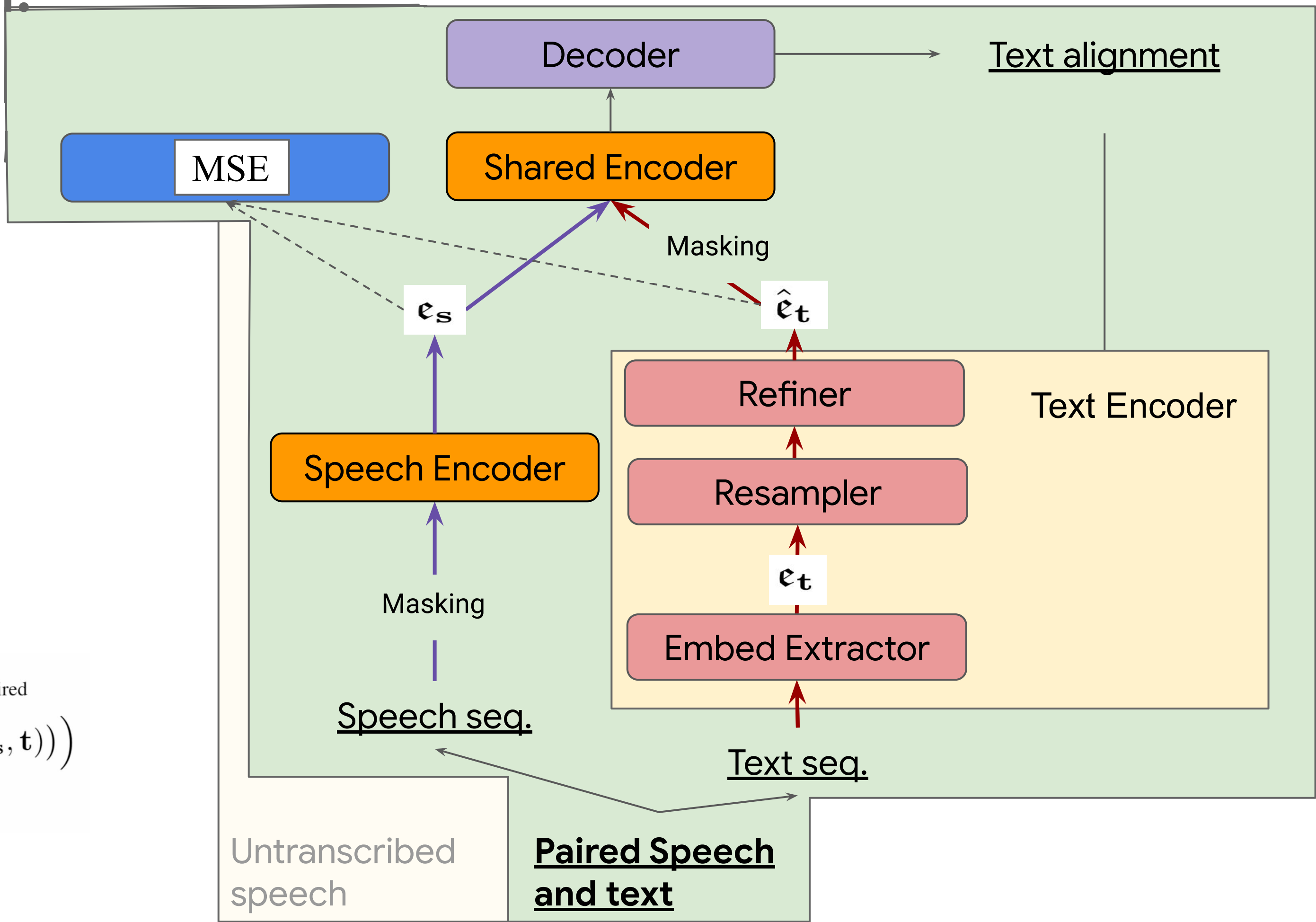
Train with $\mathcal{L}_{\text{MM}}$:

1. Align
2. Resample
3. Refine

$$\mathfrak{e}_\mathbf{s} = \theta_s(\mathbf{s}),\ \mathfrak{e}_\mathbf{t} = \theta_t(\mathbf{t}), \quad (\mathbf{t}, \mathbf{s}) \in \mathcal{X}_{\text{paired}}$$

$$\hat{\mathfrak{e}}_\mathbf{t} = \theta_{\text{Refiner}}\Big(\text{Resample}\big(\mathfrak{e}_\mathbf{t}, \text{Align}_{\text{Rnnt}}(\mathfrak{e}_\mathbf{s}, \mathbf{t})\big)\Big)$$

$$\mathcal{L}_{\text{MM}} = \text{MSE}(\mathfrak{e}_\mathbf{s}, \hat{\mathfrak{e}}_\mathbf{t}) + \mathcal{L}_{\text{Rnnt}}(\mathbf{t} \mid \mathfrak{e}_\mathbf{s})$$



Loss breakdown: Text-only

Inference using **Text Encoder**:

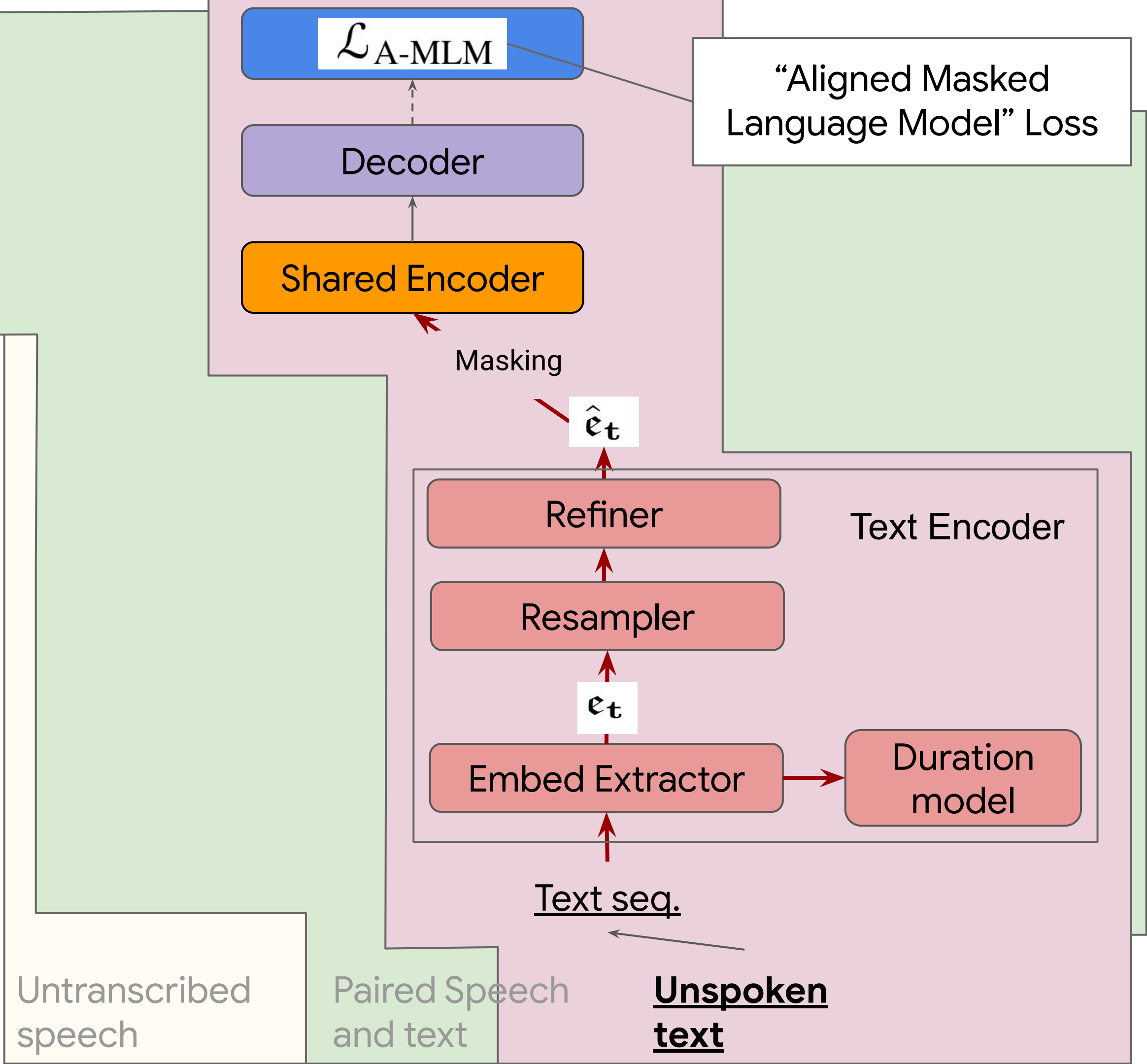
1. **Predict Duration**
2. Resample
3. Refine

Text learning with $\mathcal{L}_{\text{A-MLM}}$

1. Mask
2. Decoder loss (e.g. RNN-T)

$$\mathfrak{e}_{\mathbf{t}} = \theta_t(\mathbf{t}), \hat{\mathfrak{e}}_{\mathbf{t}} = \theta_{\text{Refiner}}\Big(\text{Resample}\big(\mathfrak{e}_{\mathbf{t}}, \theta_{\texttt{Duration}}(\mathfrak{e}_{\mathbf{t}})\big)\Big)$$

$$\mathcal{L}_{\text{A-MLM}} = \mathcal{L}_{\text{Rnnt}}\Big(\mathbf{t} \mid \text{Mask}(\hat{\mathfrak{e}}_{\mathbf{t}})\Big), \quad \mathbf{t} \in \mathcal{X}_{\text{text}}$$



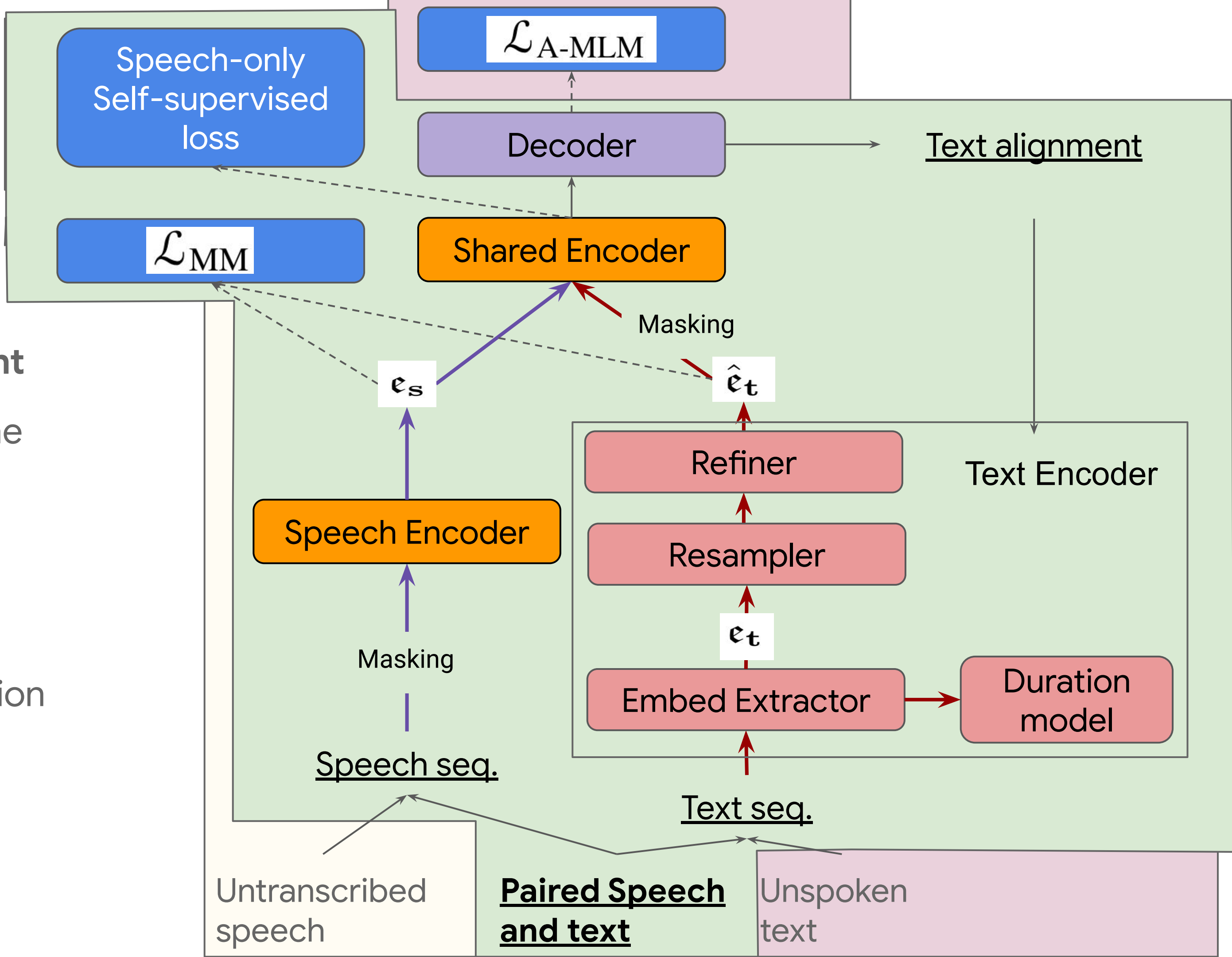
Overview

Sequence **self-alignment**

Modality matching in the intermediate layer

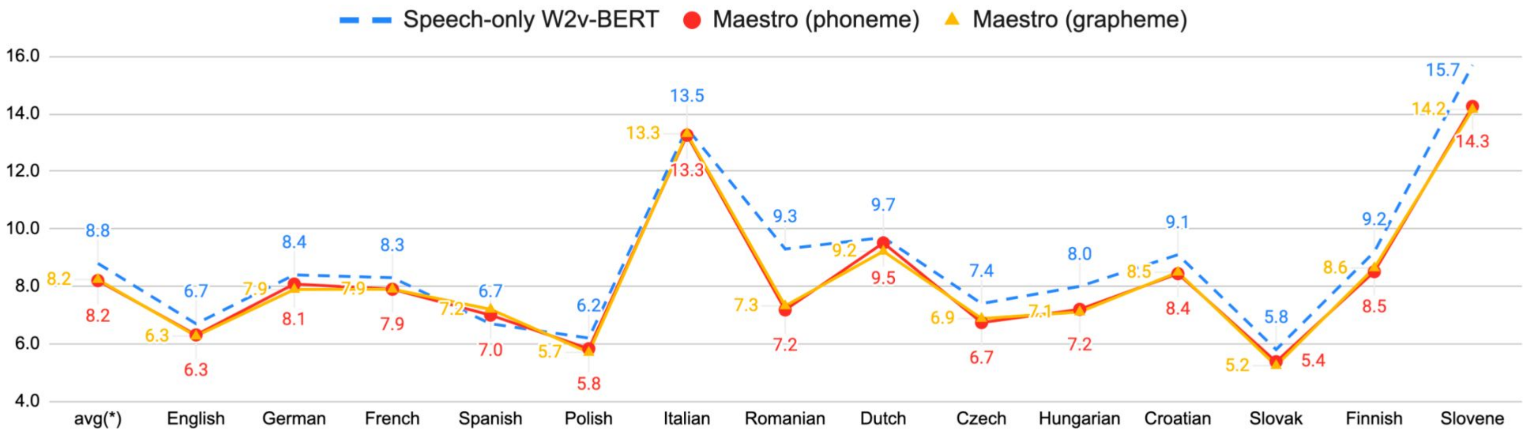
Reuse **duration** part of Parallel Tacotron

Unified framework for text-speech representation learning



Multilingual ASR: Voxpopuli (14 languages)

Breakdown: Languages are **sorted by the amount of paired data**



Generalize to different amount of paired data
No substantial difference from Phonemic and Graphemic modeling

Does this joint representation learning work on other tasks?

Speech-to-text Translation (ST, 21 languages->en)

Method	Model size	Pretraining Data					Avg BLEU
		Speech	Text	ASR	ST	MT	
Finetune: ST-only; mBART decoder init							
XLS-R	1B	437k	-	-	✗	✗	19.3
XLS-R	2B	437k	-	-	✗	✗	22.1
Finetune: ST and Machine translation (MT) jointly							
w2v-bert	0.6B	429k	-	-	✗	✗	21.0
mSLAM	0.6B	429k	mC4	2.4k	✗	✗	22.4
mSLAM	2B	429k	mC4	2.4k	✗	✗	24.8
Maestro	0.6B	429k	VP-T + mC4	2.4k	✗	✗	24.3
Maestro	**0.6B**	**429k**	**VP-T + mC4**	**2.4k**	✓	✓	**25.2**

Strong performance across ASR and Translation tasks

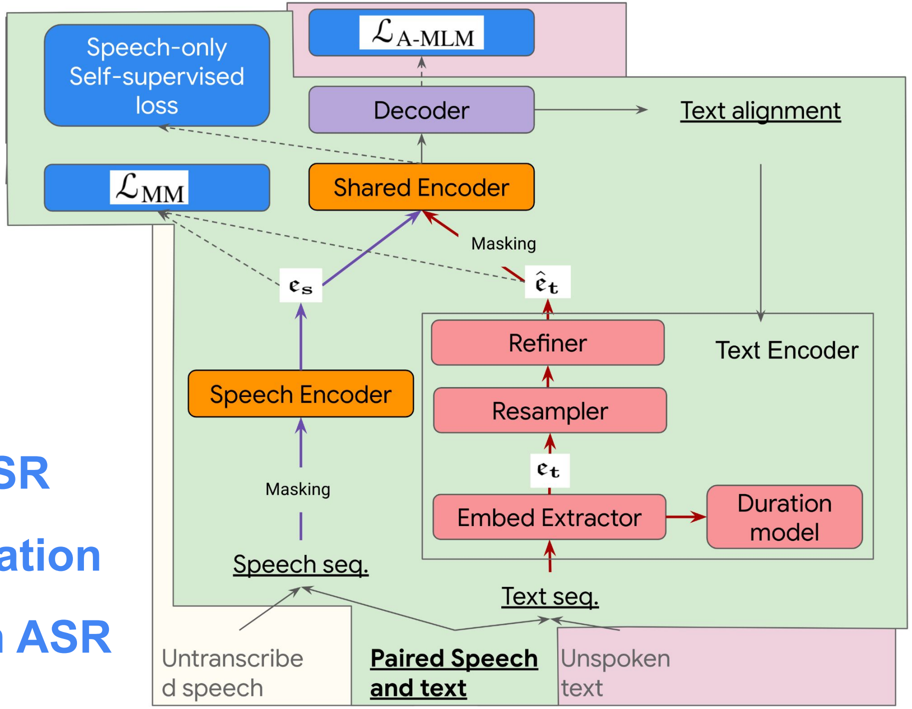
Numbers other than Maestro from "mSLAM: Massively multilingual joint pre-training for speech and text." link.

Key Finding:

Learn unified **speech-text** representations simultaneously that can transfer to diverse tasks

Solution: Maestro

- **Match speech and text modalities** in an intermediate layer via **explicit alignment of text and speech**
    - Sequence alignment
    - Matching modality embeddings
    - Duration prediction
    - Aligned masked-language model loss



Result: create new SOTAs

8% WER reduction on VoxPopuli **multilingual ASR**

2.8 BLEU improve on CoVoST 2 **Speech Translation**

4% WER reduction on SpeechStew **multidomain ASR**

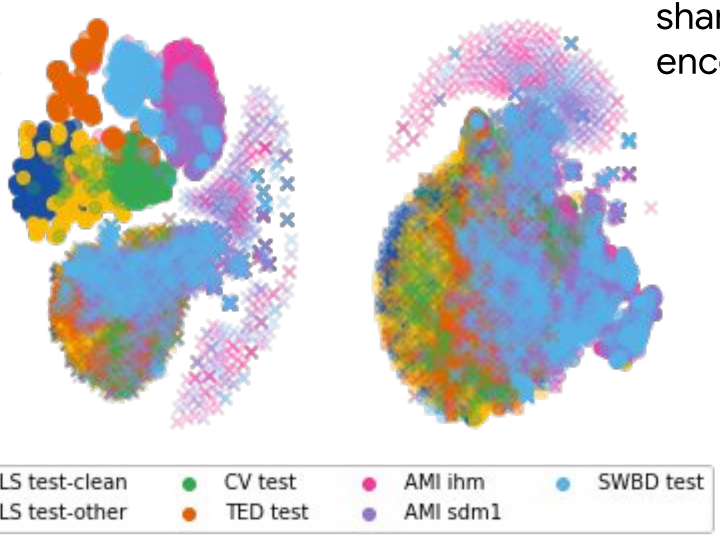
Retrieval to measure Shared Representation (ICASSP 2023)

Task: Given a speech sample, find the matching text sample or vice versa

unimodal encoders

Librispeech retrieval performance
test-clean: 20.5%
test-other: 19.3%

CV retrieval performance: 7.4%



shared encoder

Librispeech retrieval performance
test-clean: 83.5%
test-other: 68.8%

CV retrieval performance: 28.8%

Chance: 0.1% Other models at ~1-2%.
LibriSpeech trained encoders

Inspired by https://arxiv.org/abs/2209.15430 && https://arxiv.org/abs/2210.01738

Goal:

Train ASR **without transcribed speech** and **G2P**

Enable **multilingual transfer** even with unseen writing systems

Solution: Maestro-U

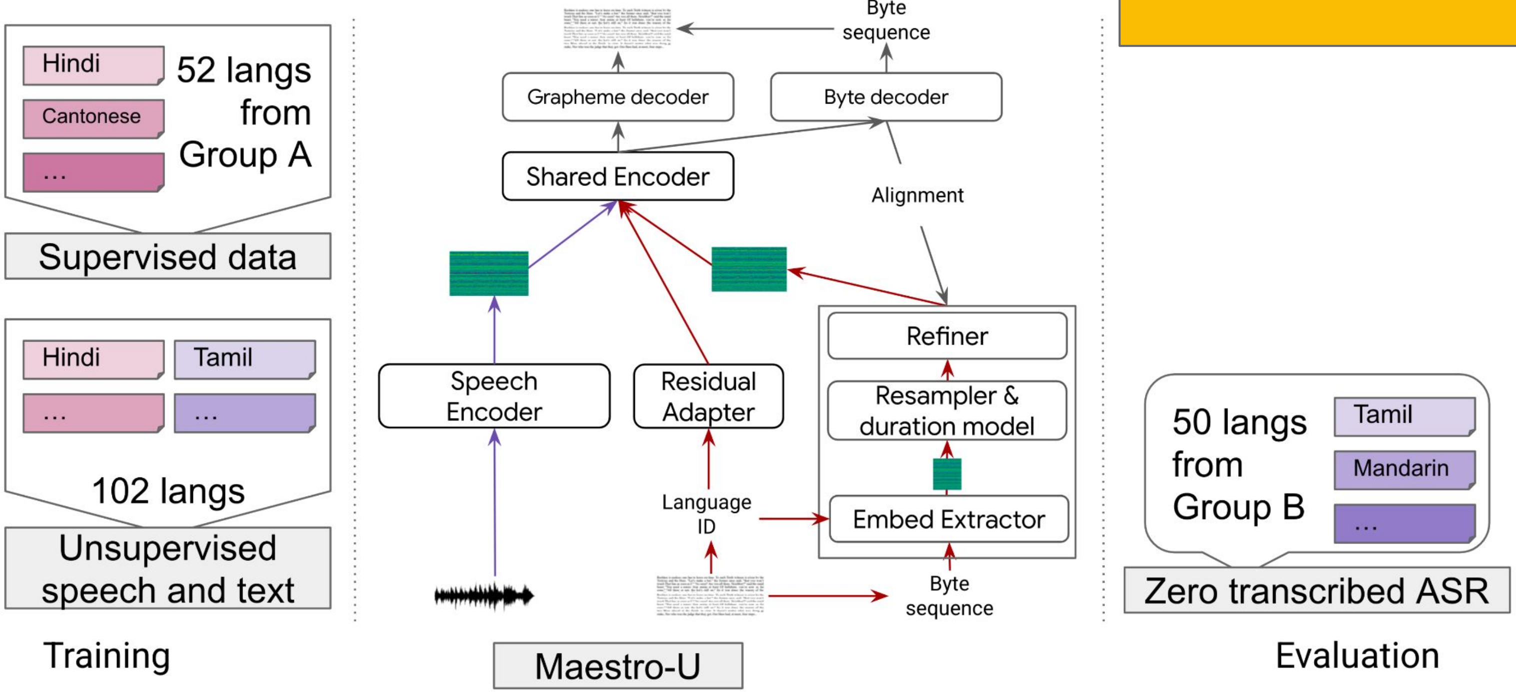
- Unsupervised speech and text learning with Maestro
- Promote multilingual knowledge transfer by Language ID and Residual Adapters
- Handling unseen writing systems by UTF-8 Bytes as text representation units

Result:

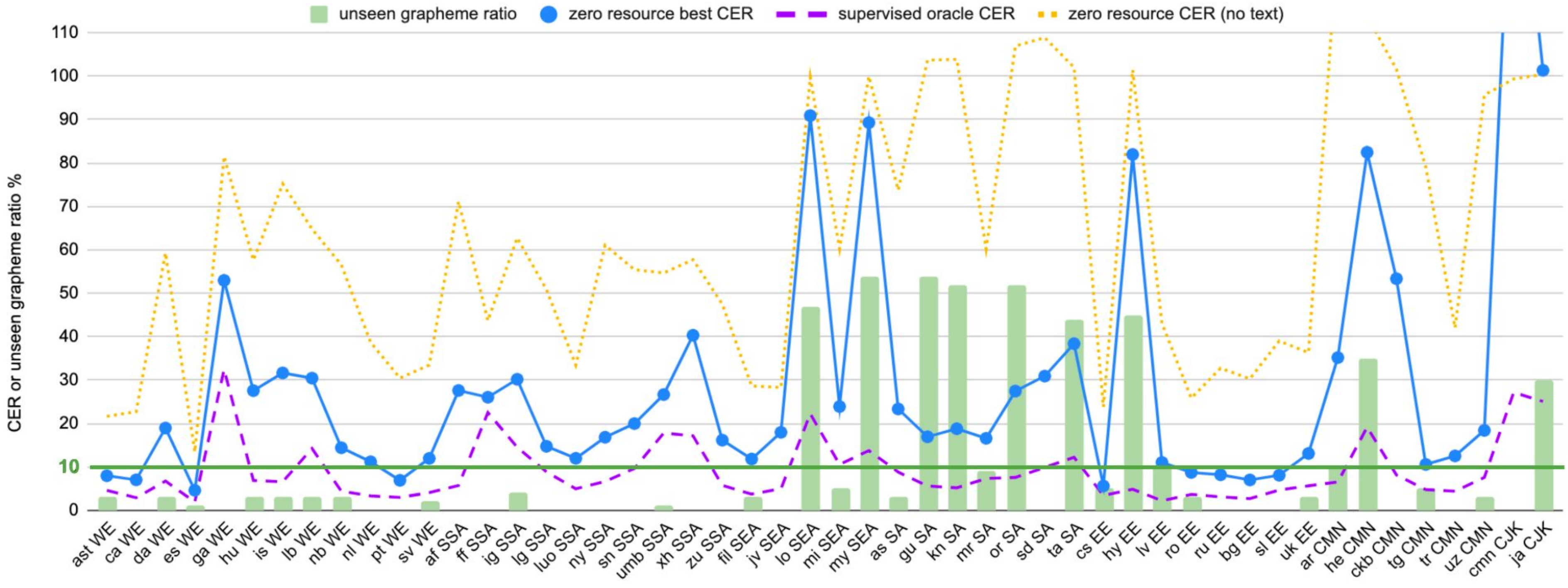
- Train ASR models without transcribed speech on 50 unseen FLEURS languages.
- Reduce the CER on languages with no supervised speech from 64.8% to 30.8%.
- Close the gap to oracle performance by 68.5% relative and reduces the CER of 19 languages below 15%.

Massive multilingual ASR language expansion with zero supervised speech

Text encoder training: learn to predict speech-like text representations on 52 supervised languages
Text encoder inference: unspoken text learning on 102 languages



Results on 50 unseen languages (FLEURS)



Reduce the CER on languages with no supervised speech from 64.8% to 30.8%.
Even on the langs with very different writing systems, e.g. South Asian langs

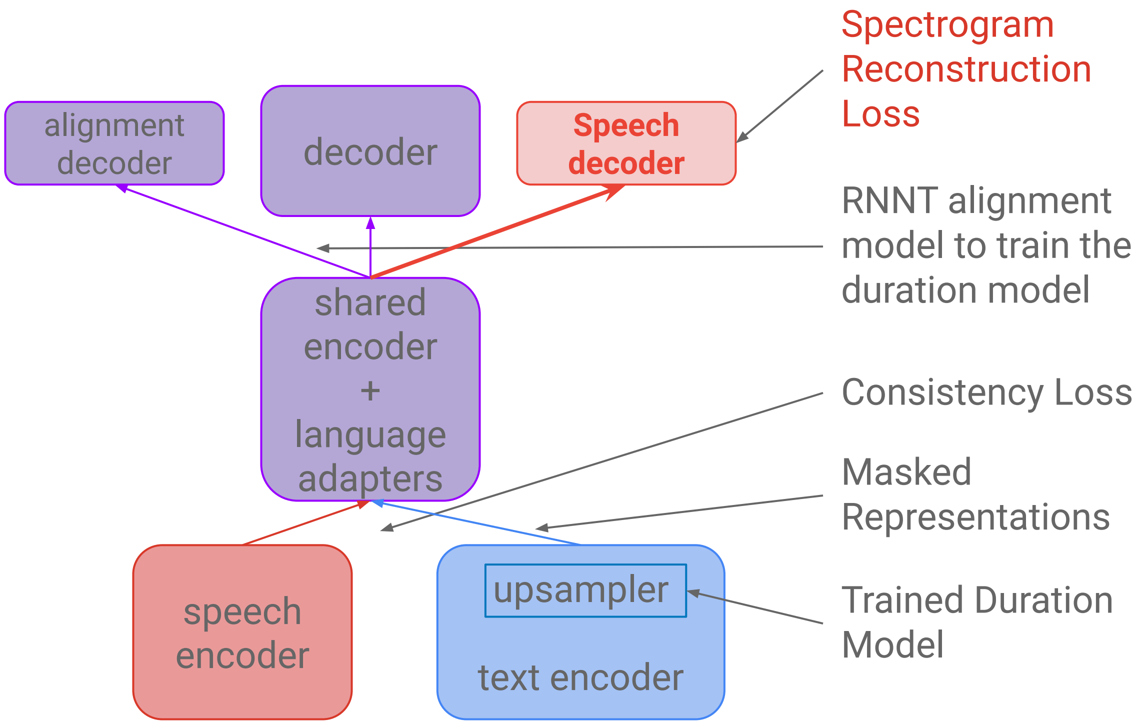
Multilingual Text to Speech (TTS)

Current TTS covers around 100+ languages.

Around 7000 languages exist in the world.

→ **Can these representations do multilingual TTS**??

Virtuoso = Maestro + speech decoder ‼



Unpaired data ⇒ Self-supervised learning

- Sp enc → Sp dec ⇒ Masked AE
- Txt enc → Txt dec ⇒ Masked LM

Paired data ⇒ Supervised learning

- Text enc → Speech dec ⇒ TTS
- Speech enc → Text dec ⇒ ASR

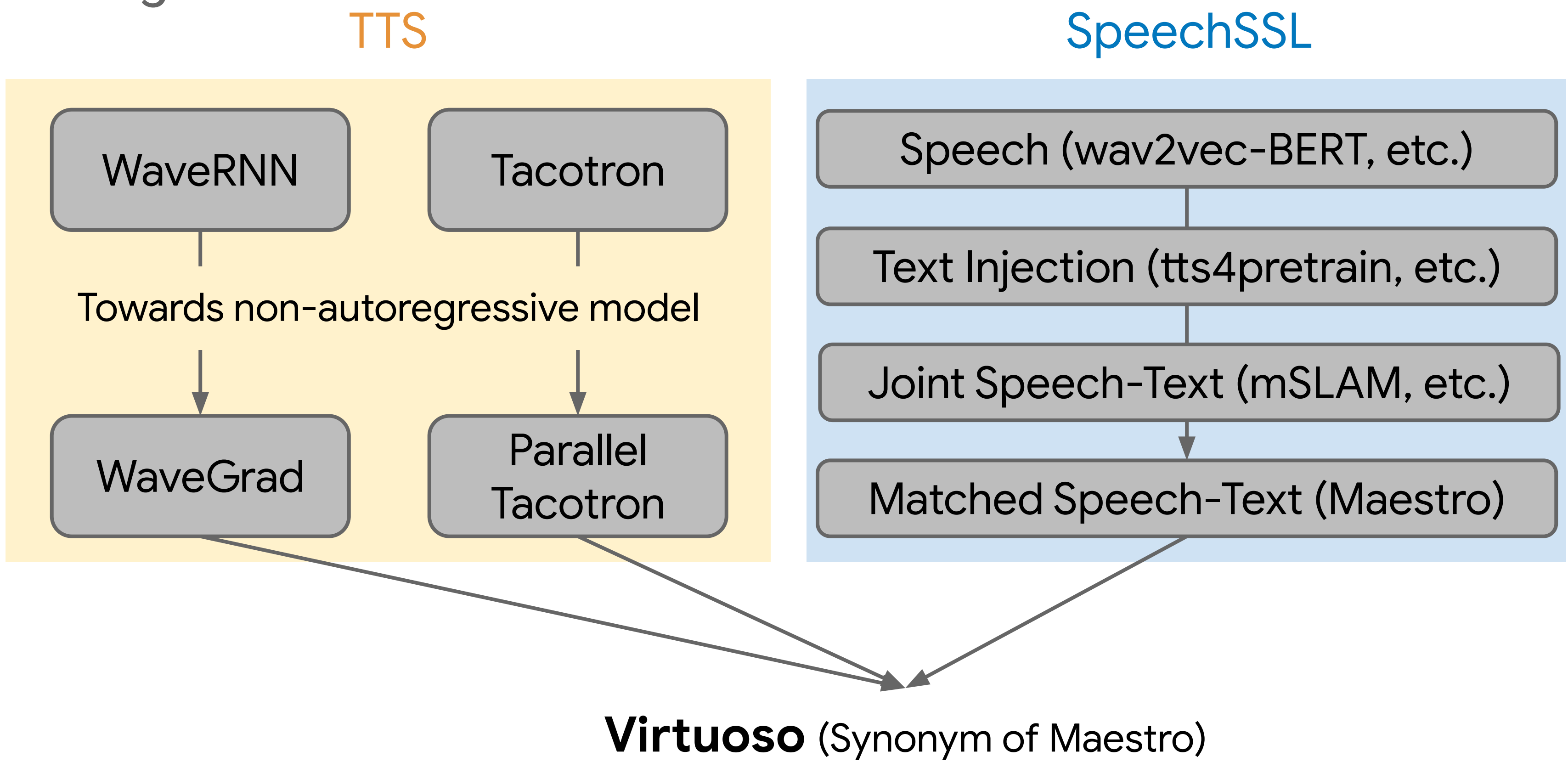
Data

- Untranscribed speech
- Unspoken text
- Paired ASR data (in-the-wild)
- Paired TTS data (in-house)

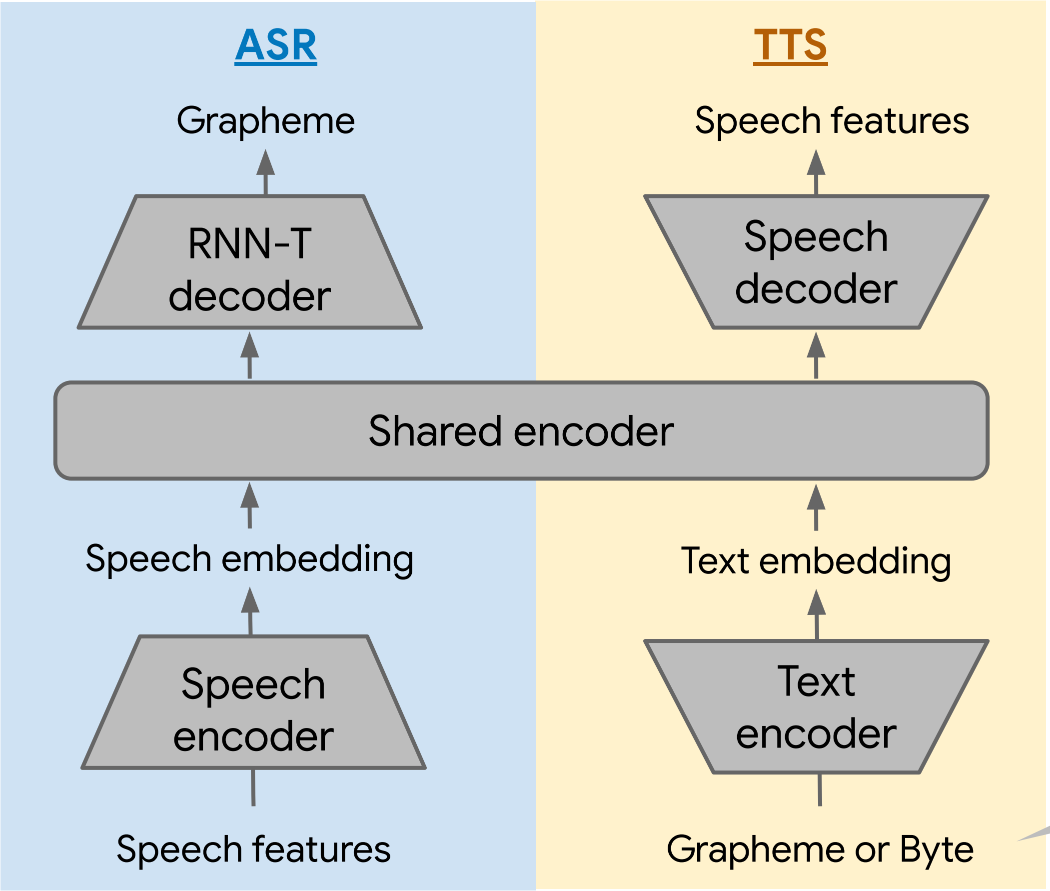
Text representation

- Phonemes; Graphemes; Bytes

Meeting of Conventional TTS and Newer methods in ASR



Consisting of ASR part and TTS part

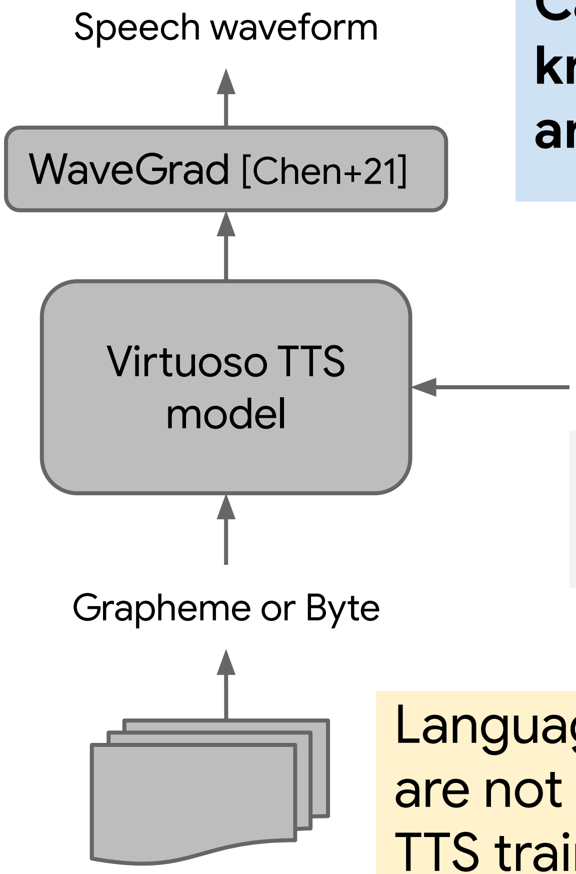


We can obtain full TTS model without fine-tuning

Grapheme or Byte-based TTS without any G2P modules

Zero-Resource TTS

Can massive multilingual knowledge obtained with ASR and SSL be transferred to TTS?



Speaker embedding

Sampled from similar locales included in TTS data

Languages which are not included in TTS training data

Multilingual TTS possible with the same ASR technology

- Virtuoso improved performance for **both major and low-resource locales.**
- Virtuoso performed well in **zero-resource settings.**
- **Byte-based model** achieved the highest linguistic accuracy.
- **Only using paired ASR+TTS** data was better in terms of naturalness.
- **Using unpaired data** was effective for zero-resource settings.

Takaaki Saeki et al., , EXTENDING MULTILINGUAL SPEECH SYNTHESIS TO 100+ LANGUAGES WITHOUT TRANSCRIBED DATA, ICASSP 2024

Representation Learning: Summary

Learning Within Modality

Audio: Full-sum/Sampling based Distillation, Sampling Guided-masking, Diffusion-based masking, Use of ephemeral sources (eg.Radio/Podacsts), SoundStream + AudioLM

Text: Large LMs integrated into e2e model

Learning Across Modalities

Encourage unified representations

Share language adapters within language families

Acoustic Prompting

Additional modalities/signals (image, video, tonal language, etc.)

Intermediate representations help other downstream tasks (phone recognition, NLP?)

Weak Supervision

Conditional adapters (on topic, contextual keywords)

Grounding around other information seen in the same context (text/audio/image/audio)

04
Evaluating representations

Evaluating Representations

- Benchmarks:
  - Pooneh Mousavi et al., “DASB - Discrete Audio and Speech Benchmark”: Discrete audio tokens across a wide range of discriminative tasks, including speech recognition, speaker identification and verification, emotion recognition, keyword spotting, and intent classification, as well as generative tasks such as speech enhancement, separation, and text-to-speech http://arXiv:2406.14294
  - Shikhar Vashishth et al., STAB: Speech Tokenizer Assessment Benchmark : Measurements across in variance, robustness, compressibility, coverage http://arxiv.org/abs/2409.02384

- Characterizing neural representations of context-dependent and dynamic patterns of neural activity: speech perception approach [46]

Google

Evaluating Representations

- Dialog tasks that capture speaking styles in their responses
  - StyleTalk (https://github.com/DanielLin94144/StyleTalk)

- Automatically detect domain shifts over time and adapt for best performance
  - *Vision*: stochastic model restoring (Wang et al., 2022), sample-efficiency entropy minimization (Niu et al., 2022a), sharpness-aware reliable entropy minimization (Niu et al., 2022b), and fixed frequency model reset (Press et al., 2024) are examples to address domain shifts.

  - *Speech*: Recent ACL paper, uses the loss function to detect domain shifts (Lin et al., EMNLP 2024)

- MTEB (Massive Text Embedding Benchmark) spans 8 embedding tasks covering a total of 58 datasets and 112 languages. Could we have a similar benchmark for joint representations?

Google

Promising lines of research

- Can we guide the masking during representation learning?
  - To prevent learning from erroneous samples in the vast amount of audio
  - Need to reduce amount of in-domain data for adapting pre-trained models trained on out-of-domain data [26,27]
  - Can we stabilize performance fluctuations that arise from changes in masking ratio?
  - Can we use a teacher to guide the samples to mask?

- Can we introduce more supervision into the pre-training process?
  - Curriculum training schedule
  - Consistency regularization

- What is a good tokenizer? How do you define ‘good’?

Concluding Remarks

- Code-switching is by no means a solved problem for ASR or other ST/TTS tasks
- Well-represented Data Resources are scarce- how can we grow these?
- Language Identification and Domain shifts are crucial and still remains a difficult problem for several code-switched languages - how do we make models robust to this?
- Joint speech-text representation learning is useful for ASR, ST, TTS. What about other tasks such as recognizing and responding to emphasis in speeches and dialogues? Spoken content retrieval?

Machine Learning continues to produce large models that can scale and be prompted to solve these tasks. These fundamental challenges remain and more research in these areas will pave the way for usable, scalable, multilingual models.

References

1. Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." *arXiv preprint arXiv:1807.03748* (2018).
2. Chung, Yu-An, and James Glass. "Generative pre-training for speech with autoregressive predictive coding." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.
3. Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *arXiv preprint arXiv:2002.05709* (2020).
4. Pascual, Santiago, et al. "Learning problem-agnostic speech representations from multiple self-supervised tasks." *arXiv preprint arXiv:1904.03416* (2019).
5. Chorowski, Jan, et al. "Unsupervised Neural Segmentation and Clustering for Unit Discovery in Sequential Data." *NeurIPS 2019 workshop-Perception as generative reasoning-Structure, Causality, Probability*. 2019.
6. Park, Daniel S., et al. "Specaugment: A simple data augmentation method for automatic speech recognition." *arXiv preprint arXiv:1904.08779* (2019).
7. Pascual, Santiago, et al. "Learning problem-agnostic speech representations from multiple self-supervised tasks." *arXiv preprint arXiv:1904.03416* (2019).
8. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
9. He, Kaiming, et al. "Momentum contrast for unsupervised visual representation learning." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
10. Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." *Advances in Neural Information Processing Systems* 33 (2020).

References

11. M. Ravanelli and Y. Bengio, “Learning Speaker Representations with Mutual Information”, Interspeech 2019.
12. M. Ravanelli, et al. “Multi-Task Self Supervised Learning for Robust Speech Recognition”, ICASSP 2020.
13. Luyu Wang, Aaron van den Oord, “A simple framework for contrastive learning of visual representations." *arXiv preprint arXiv:2002.05709* (2020).
14. Pascual, Santiago, et al. "Multi-Format Contrastive Learning of Audio Representations." *arXiv preprint arXiv:1904.03416* (2019). SAS workshop, NeurIPS, 2020.
15. Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, pages 539–546. IEEE, 2005.
16. Pierre Sermanet, Corey Lynch, Jasmine Hsu, and Sergey Levine. Time-contrastive networks: Self-supervised learning from multi-view observation. arXiv preprint arXiv:1704.06888, 2017.
17. Dong Yu, “ Weak Supervision”, SAS workshop, NeurIPS, 2020.
18. Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." *arXiv preprint arXiv:1703.03400* (2017).
19. Ravanelli, Mirco, and Yoshua Bengio. "Speaker recognition from raw waveform with sincnet." *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018.
20. Trapit Bansal, Rishikesh Jha, Tsendsuren Munkhdalai, Andrew McCallum, “Self-Supervised Meta-Learning for Few-Shot Natural Language Classification Tasks”, arXiv:2009.08445v2 (2020).
21. Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” arXiv preprint arXiv:2106.07447, 2021
22. Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, Yonghui Wu,”W2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training”, arXiv:2108.06209, ASRU 2021.
23. Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021)
24. Chen, Z., Zhang, Y., Rosenberg, A., Ramabhadran, B., Wang, G. and Moreno, P., 2021. Injecting text in self-supervised speech pretraining. *arXiv preprint arXiv:2108.12226*, ASRU 2021

References

25. Bapna, Ankur, et al. "SLAM: A Unified Encoder for Speech and Language Modeling via Speech-Text Joint Pre-Training." *arXiv preprint arXiv:2110.10329* (2021).
26. Hsu, Wei-Ning, et al. "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training." *arXiv preprint arXiv:2104.01027* (2021)
27. Chan, William, et al. "SpeechStew: Simply mix all available speech recognition data to train one large neural network." *arXiv preprint arXiv:2104.02133* (2021).
28. Wang, G., Rosenberg, A., Chen, Z., Zhang, Y., Ramabhadran, B., Moreno, P.J., SCADA: Stochastic, Consistent and Adversarial Data Augmentation to Improve ASR. Proc. Interspeech 2020.
29. G. Wang, A. Rosenberg, Z. Chen, Y. Zhang, B. Ramabhadran, Y. Wu, P. Moreno, "Improving Speech Recognition Using Consistent Predictions on Synthesized Speech," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020.
30. Zhehuai Chen, Yu Zhang, Andrew Rosenberg, Bhuvana Ramabhadran, Pedro Moreno, Gary Wang”Tts4pretrain 2.0: “Advancing the use of text and speech in ASR pretraining with consistency and contrastive losses”, ICASSP 2022
31. Zhehuai Chen, Yu Zhang, Andrew Rosenberg, Bhuvana Ramabhadran, Pedro Moreno, Ankur Bapna, Heiga Zen, “MAESTRO: Matched Speech Text Representations through Modality Matching”, *arXiv preprint arXiv:2204.03409* (2022).
32. Baevski, Alexei, et al. "Data2vec: A general framework for self-supervised learning in speech, vision and language." 2022 from Facebook.
33. Ao, Junyi, et al. "Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing." 2021 from Microsoft.
34. Yuan, Xin, et al. "Multimodal contrastive training for visual representation learning." 2021 from Adobe.
35. Renduchintala, Adithya, et al. "Multi-modal data augmentation for end-to-end asr." 2018 from JHU.
36. Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, Alexis Conneau, “mSLAM: Massively multilingual joint pre-training for speech and text”, arXiv:2202.01374(2022).

References

37. Bapna, Ankur, et al. "SLAM: A Unified Encoder for Speech and Language Modeling via Speech-Text Joint Pre-Training." *arXiv preprint arXiv:2110.10329* (2021).
38. Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. 2018. Mine: mutual information neural estimation. In Proc. ICML.
39. Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A Alemi, and George Tucker. 2019. On variational bounds of mutual information. arXiv preprint arXiv:1905.06922.
40. Kawakami, Kazuya, et al. "Learning robust and multilingual speech representations." arXiv preprint arXiv:2001.11128 (2020).
41. Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Proc. of NAACL.
42. Olivier J Hénaff, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. 2019. Data-efficient image recognition with contrastive predictive coding. arXiv preprint arXiv:1905.09272..
43. Adi, Yossi, et al. "Fine-grained analysis of sentence embeddings using auxiliary prediction tasks." arXiv preprint arXiv:1608.04207 (2016). ICLR 2017
44. Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, Yonghui Wu, Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages, March 2023
45. Babu, Arun, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh et al. "XLS-R: Self-supervised cross-lingual speech representation learning at scale." arXiv preprint arXiv:2111.09296 (2021).
46. Leonard, M.K. and Chang, E.F., 2014. Dynamic speech representations in the human temporal lobe. Trends in cognitive sciences, 18(9), pp.472-479.
47. Lin, Guan-Ting, Cheng-Han Chiang, and Hung-yi Lee. "Advancing large language models to capture varied speaking styles and respond properly in spoken conversations." arXiv preprint arXiv:2402.12786 (ACL 2024).

References

48. Guan-Ting Lin, Wei-Ping Huang, Hung-yi Lee, “Continual Test-time Adaptation for End-to-end Speech Recognition on Noisy Speech”, arXiv:2406.11064v2, EMNLP 2024.
49. Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai, “Continual test-time domain adaptation”, In Proceedings of the IEEE/CVF Conference on Computer, Vision and Pattern Recognition, pages 7201–7211, 2022.
50. Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan, “Efficient test-time model adaptation without forgetting”, In International conference on machine learning, pages 16888–16905. PMLR, 2022a.
51. Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan, “Towards stable test-time adaptation in dynamic wild world”, In The Eleventh International Conference on Learning Representations, 2022b.
52. Ori Press, Steffen Schneider, Matthias Kümmerer, and Matthias Bethge, “ Rdumb: A simple approach that questions our progress in continual test-time adaptation”, Advances in Neural Information Processing Systems, 36, 2024.
53. MTEB: Massive Text Embedding Benchmark
54. Niklas Muennighoff, Nouamane Tazi, Loïc Magne, Nils Reimers, “MTEB: Massive Text Embedding Benchmark”, arXiv:2210.07316v3, 2022.