Multilingual and Low-Resource Speech Recognition

Ondrej Klejch

Automatic Speech Recognition – ASR Lecture 17
16 March 2026

Languages of the World

- Over 6,000 languages globally....
- In Europe alone
  - 24 official languages and 5 "semi-official" languages
  - Over 100 further regional/minority languages
  - If we rank the 50 most used languages in Europe, then there are over 50 million speakers of languages 26-50 (Finnish – Montenegrin)
- 3,000 of the world's languages are endangered
- Google cloud speech API covers over 125 languages and more than 300 accents/dialects of those languages; Apple Siri covers over 21 languages; Google assistant has over 30

Under-resourced languages

Under-resourced (or low-resourced) languages have some or all of the following characteristics

- limited web presence
- lack of linguistic expertise
- lack of digital resources: acoustic and text corpora, pronunciation lexica, ...

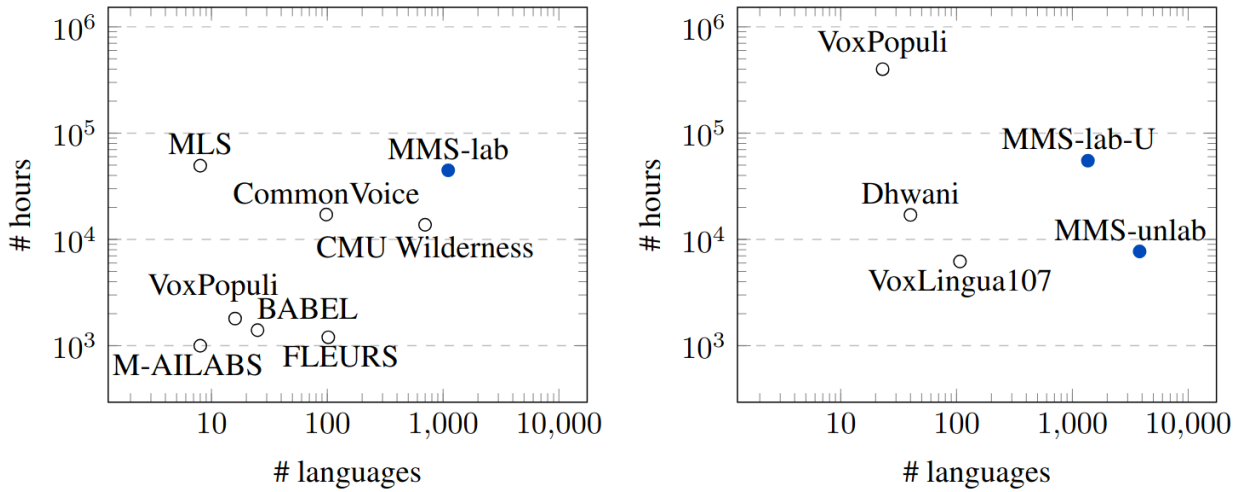
Under-resourced languages thus provide a challenge for speech technology

See Besacier et al (2014) for more

Speech recognition of under-resourced languages

- Limited data to train acoustic and language models
- Language specific characteristics (e.g. morphology)
- Prevalence of code-switching
- Challenge of constructing pronunciation lexica
- Transferring knowledge between languages
- Leveraging unpaired speech and text

Available speech and text datasets



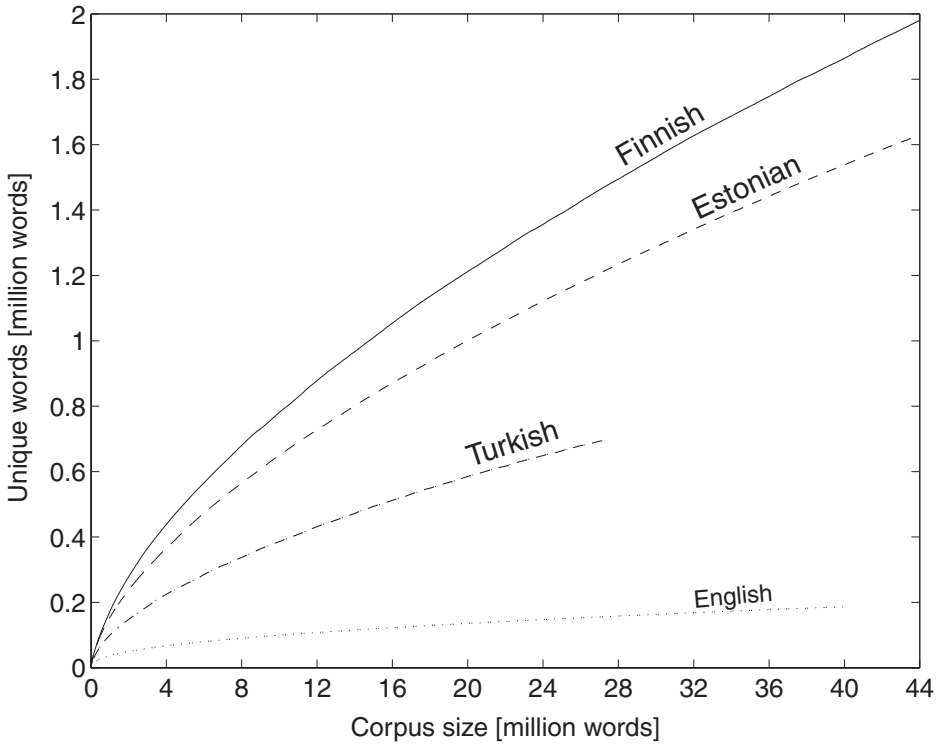
Pratap et al (2024)

- *Wiktionary* - pronunciations for 972 languages. However, the majority of them has less than 100 pronunciations.
- *MADLAD-400* - "a manually audited, 3T token monolingual dataset based on CommonCrawl, spanning 419 languages."
- *FineWeb-2* - "filtered data for 1,893 language-script pairs. Of these, 486 have more than 1MB of text data, and 80 have more than 1GB of filtered data."

Morphology

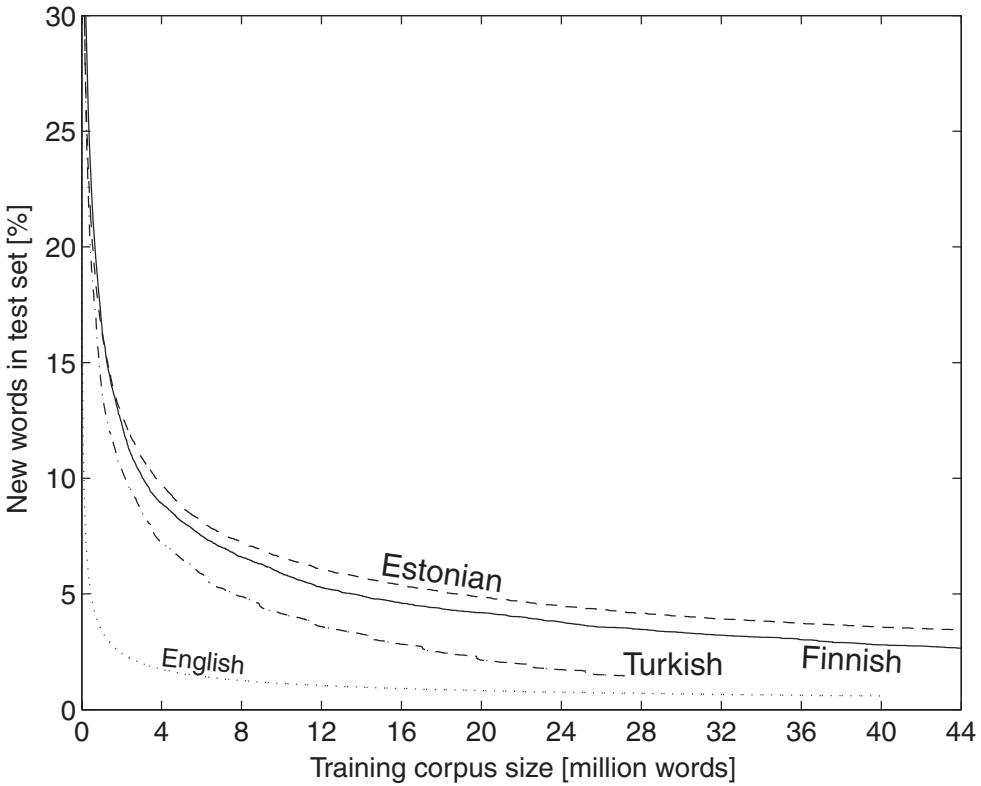
- Many languages are morphologically richer than English: this has a major effect of vocabulary construction and language modelling
- Compounding (eg German): decompose compound words into constituent parts, and carry out pronunciation and language modelling on the decomposed parts
- Highly inflected languages (eg Arabic, Slavic languages): specific components for modelling inflection (eg factored language models)
- Inflecting and compounding languages (eg Finnish, Estonian)
- All approaches aim to reduce ASR errors by reducing the OOV rate through modelling at the morph/subword level; also addresses data sparsity

Vocabulary size for different languages



Creutz et al (2007)

OOV rate for different languages



Creutz et al (2007)

Overcoming OOVs with subword tokenization

- Morfessor (`http://www.cis.hut.fi/projects/morpho/`)
  - "Morfessor is an unsupervised data-driven method for the segmentation of words into morpheme-like units."
  - Aims to identify frequently occurring substrings of letters within either a word list (type-based) or a corpus of text (token-based)
  - Uses a probabilistic framework to balance between few, short morphs and many, longer morphs
- byte-pair encoding (BPE), Greedy Unigram, wordpieces, ...
- If the parameters and context length for the language model are optimized to a comparable level, the actual segmentation method has only a minor effect on the speech recognition performance. (Smit et al 2021)
- Subword language models may require longer context (since multiple tokens correspond to one word)

Code switching

- Code switching can be common in low-resource languages
- Hard to model if only monolingual training data is available
- Can interpolate monolingual language models, but how to predict likely switching points?
- Need to consider if there is a change in phonology

"masithi 3 o'clock ke eclocktower mamela kyk hier ndiyamazi i know him i got him ... ndizithi kuye masiye e waterfront i wont tell him that i'm meeting a friend but ndiyayazi he wont mind xasidibana nawe he will buy us drinks and some lunch then sonwabe wethu"

Phonemes, graphemes and subword tokens

- Can represent pronunciations as a sequence of graphemes (letters) rather than a sequence of phones
- A standard feature of "end-to-end" ASR (Lectures 13 and 14)
- Advantages of grapheme-based pronunciations
  - No need to construct/generate phone-based pronunciations
  - Can use unicode attributes to assist in decision tree construction (Lecture 7)
- Disadvantages:
  - may not always be a direct link between graphemes and sounds (eg. tough, though, through vs. seem, seam, cede, siege)
  - Orthographic representation may not be very efficient (eg. Xhosa, Gaelic)
- Recently, BPE and wordpiece tokens have been used.

Grapheme-based ASR results for 6 low-resource languages

Language	ID	System	WER (%)		
			tg	+cn	cnc
Kurmanji Kurdish	205	Phonetic	67.6	65.8	64.1
		Graphemic	67.0	65.3	
Tok Pisin	207	Phonetic	41.8	40.6	39.4
		Graphemic	42.1	41.1	
Cebuano	301	Phonetic	55.5	54.0	52.6
		Graphemic	55.5	54.2	
Kazakh	302	Phonetic	54.9	53.5	51.5
		Graphemic	54.0	52.7	
Telugu	303	Phonetic	70.6	69.1	67.5
		Graphemic	70.9	69.5	
Lithuanian	304	Phonetic	51.5	50.2	48.3
		Graphemic	50.9	49.5	

IARPA Babel, 40h acoustic training data per language, monolingual training; cnc is confusion network combination, combining the grapheme- and phone-based systems

Gales et al (2015)

Grapheme-based ASR results for English

Model	CD	PD	Dataset	Ph	Gr
5x800	N	Y	*test-clean*	**3.9**	4.1
			test-other	**9.4**	10.7
	Y	N	*test-clean*	4.0	**3.9**
			test-other	**8.9**	9.2
	Y	Y	*test-clean*	3.8	**3.4**
			test-other	9.0	**8.4**

Librispeech Word Error Rate of phonetic (Ph) and graphemic (Gr) ASR under different context dependency (CD) and position dependency (PD) settings.

Le et al (2019)

Multilingual and cross-lingual acoustic models

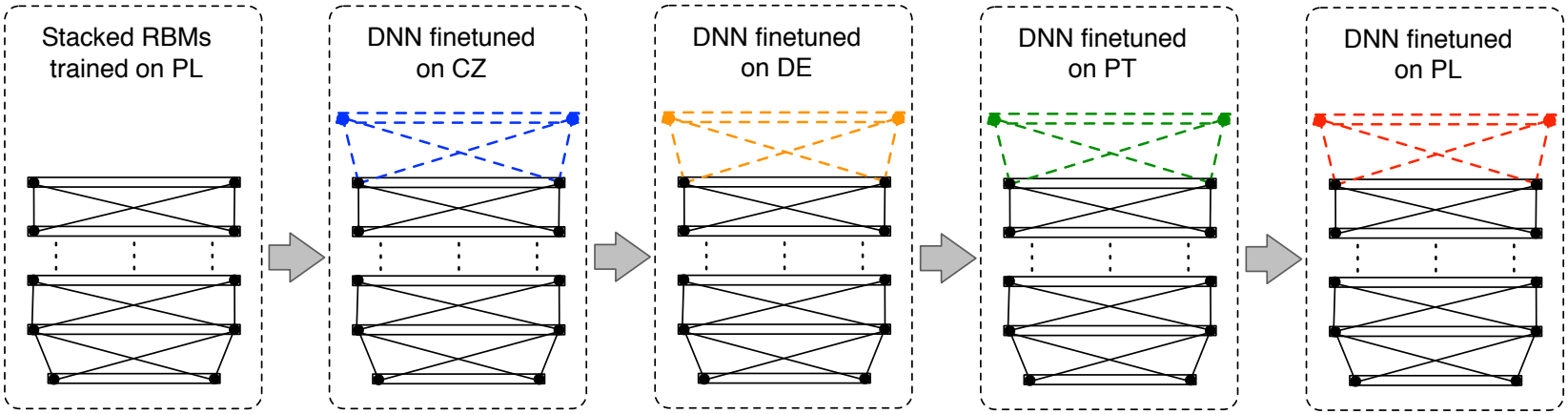
How to share information from acoustic models in different languages?

- General principle: share model parameters across languages, learning a multilingual representation of speech
- In neural network acoustic models, share hidden layers between languages
- Can share phone sets or map them between languages...
- ... but output layers were often monolingual, language specific

Multilingual and cross-lingual acoustic models

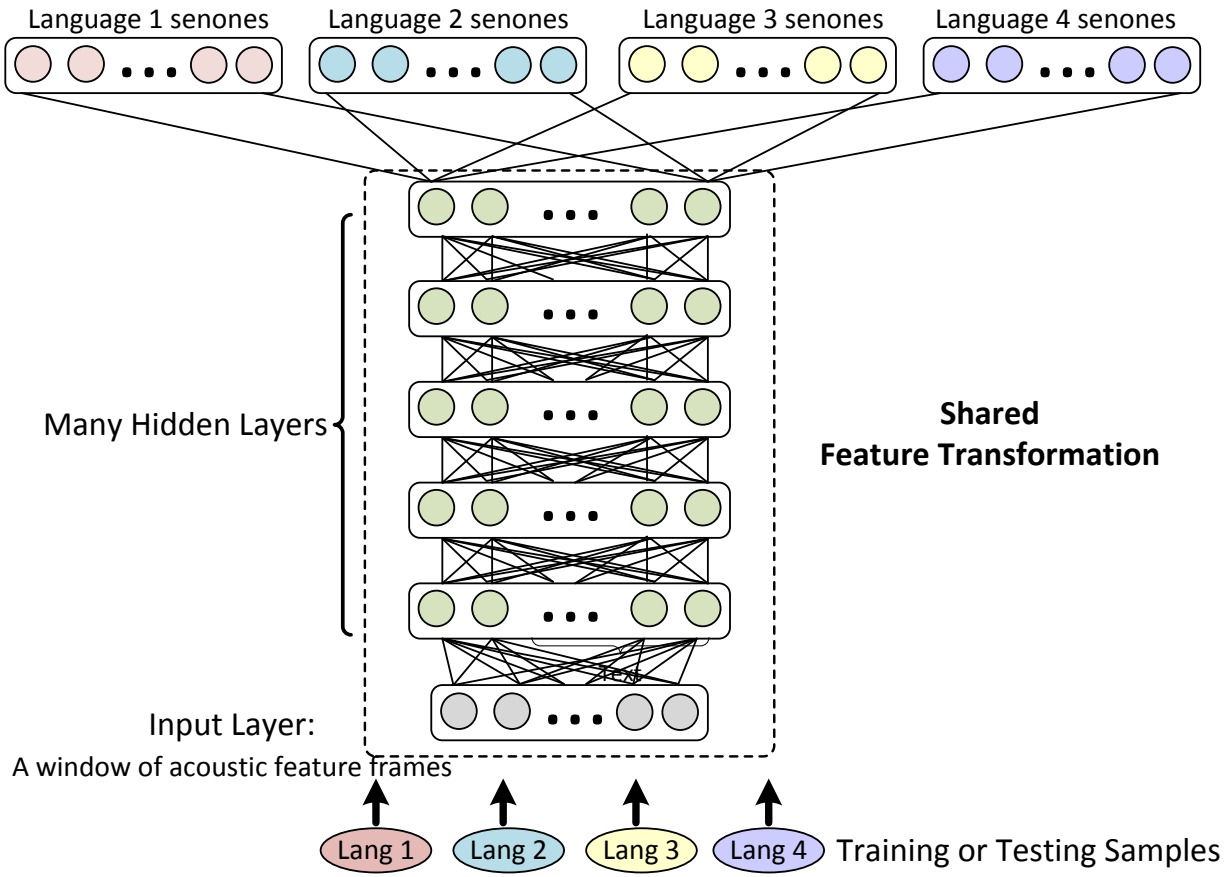
- **Multi-lingual phone sets** – use a network with multilingual hidden representations directly in a hybrid DNN/HMM system
- **Hat-swap/multi-task** – train a network with an output layer for each language, but shared hidden layers
- **Multilingual bottleneck** – use a bottleneck hidden layer (trained in a multilingual) way as features for either a GMM- or NN-based system
- **Pre-train** without phonetic labels in a language-independent manner

Hat Swap – architecture



Ghoshal et al, 2013

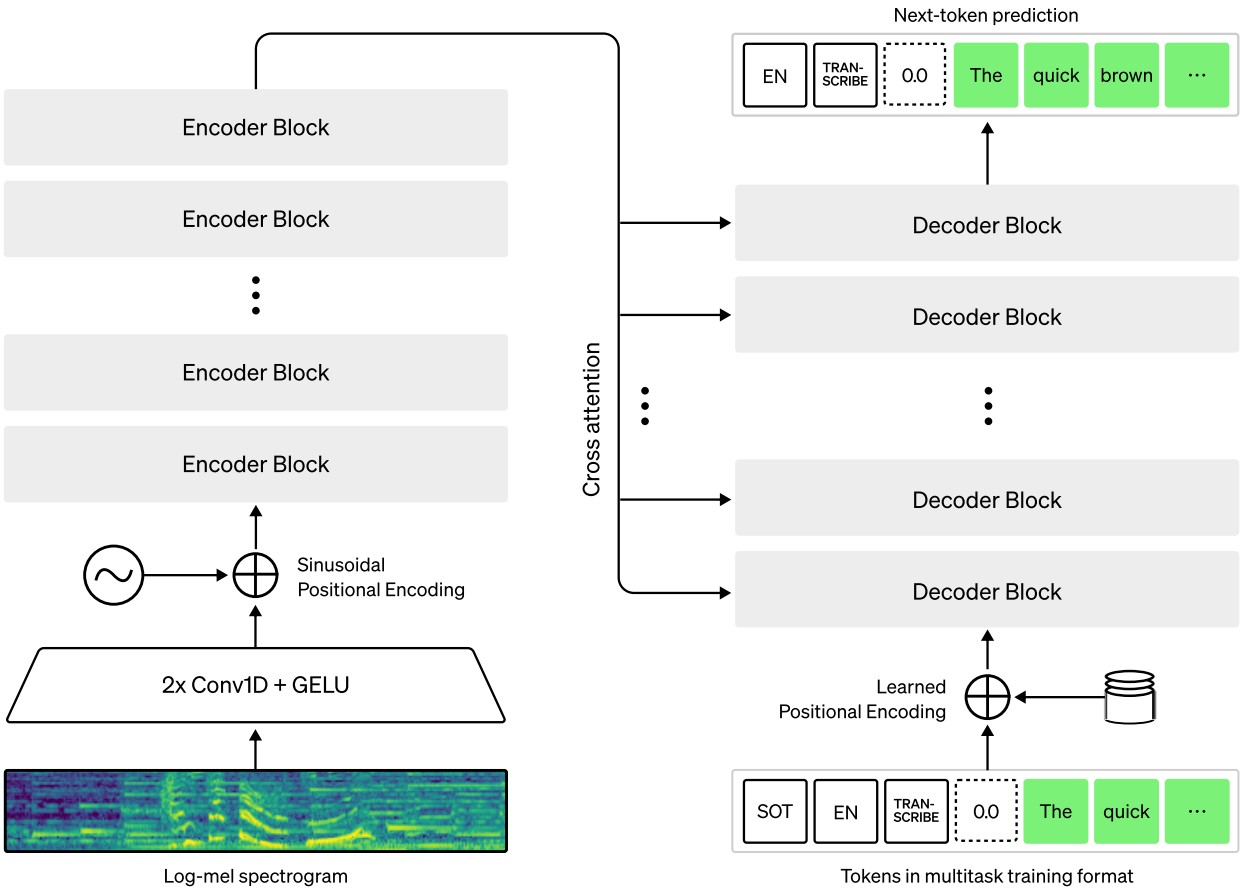
Multi-lingual networks ("block softmax")



Huang et al, 2013

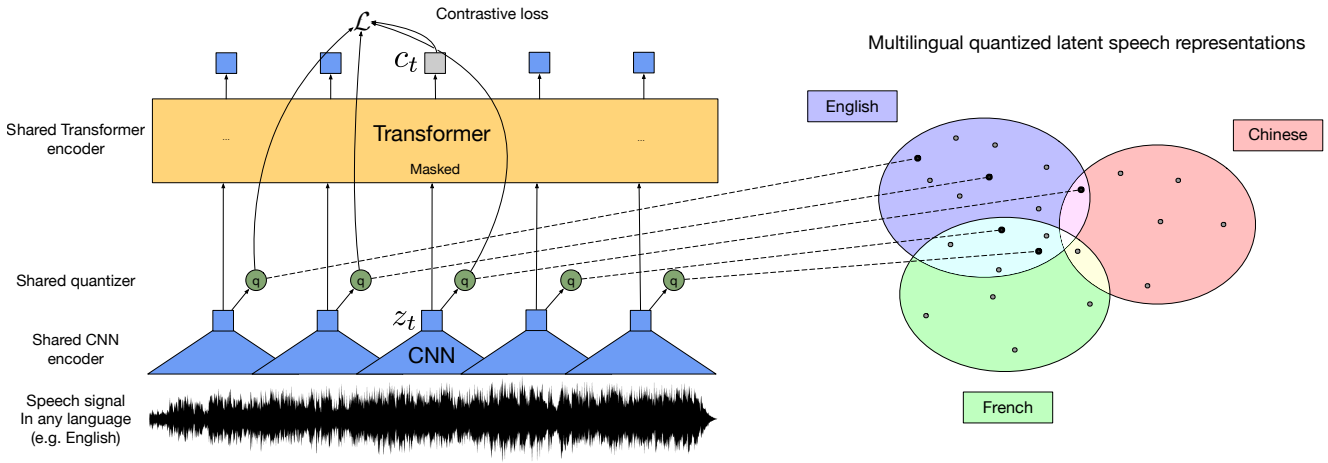
NB: A senone is a context-dependent tied state

End-to-end multi-lingual networks – architecture



Radford et al, 2023

Self-supervised pre-training



Conneau et al, 2020

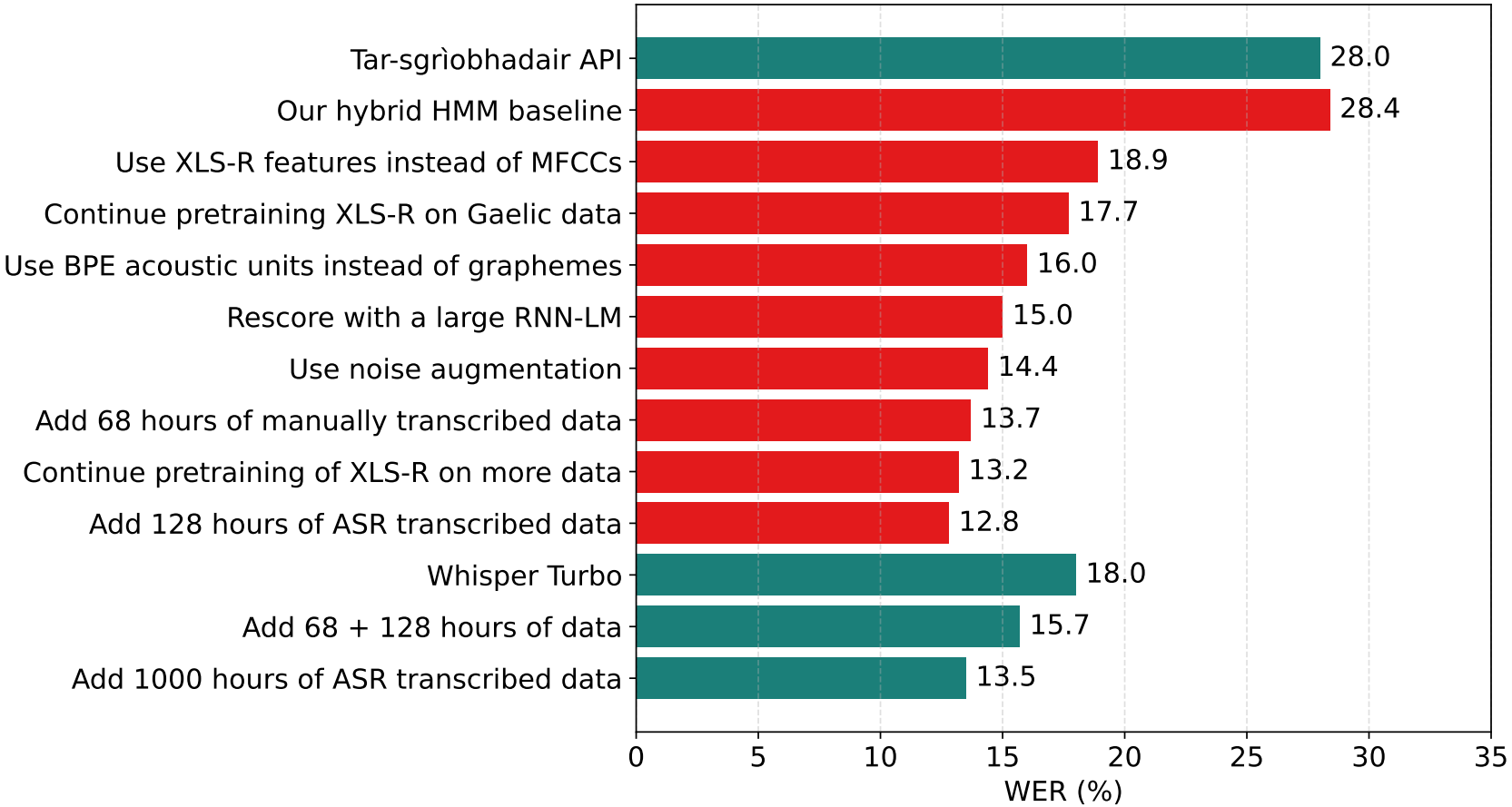
- Self-supervised training was taught in Lecture 16.
- Multilingual models like XLSR-53, XLS-R and XEUS are very good for low-resource languages.
- Continued self-supervised pre-training can leverage unlabelled data in the target language (Lam-Yee-Mui et al, 2023).

Self-supervised training results on MLS

Model	#pt	#ft	en	de	nl	fr	es	it	pt	pl
Number of training hours			44.7k	2k	1.6k	1.1k	918	247	161	104
LibriVox	1	1-1h	13.2	25.5	30.3	37.0	22.5	23.3	37.8	38.4
LibriVox	1	1-10h	10.6	14.5	19.6	19.6	13.5	16.7	25.0	32.0
XLSR-53	53	1-1h	17.3	10.6	15.6	17.0	10.4	15.1	21.4	31.9
XLSR-53	53	1-10h	14.6	8.4	12.8	12.5	8.9	13.4	18.2	21.2
XLSR-53	53	1-100h	13.2	7.4	10.9	9.8	7.9	12.0	15.7	18.9
XLSR-53	53	1-full	-	7.0	**10.8**	7.6	6.3	**10.4**	**14.7**	**17.2**
Pratap et al. (2020)	-	1-full	**5.88**	**6.49**	12.02	**5.58**	**6.07**	10.54	19.49	20.39

XLSR-53 fine-tuned on 1h, 10h, 100h and full data to evaluate the few-shot capability of the model (Conneau et al, 2020).

Case-study: Scottish Gaelic ASR



Klejch et al (2025)

Very large multilingual models

Model	Year	Number of Languages
OpenAI Whisper	2022	97
Google USM	2023	100
Meta MMS	2023	1100
Meta Omnilingual ASR	2025	1600

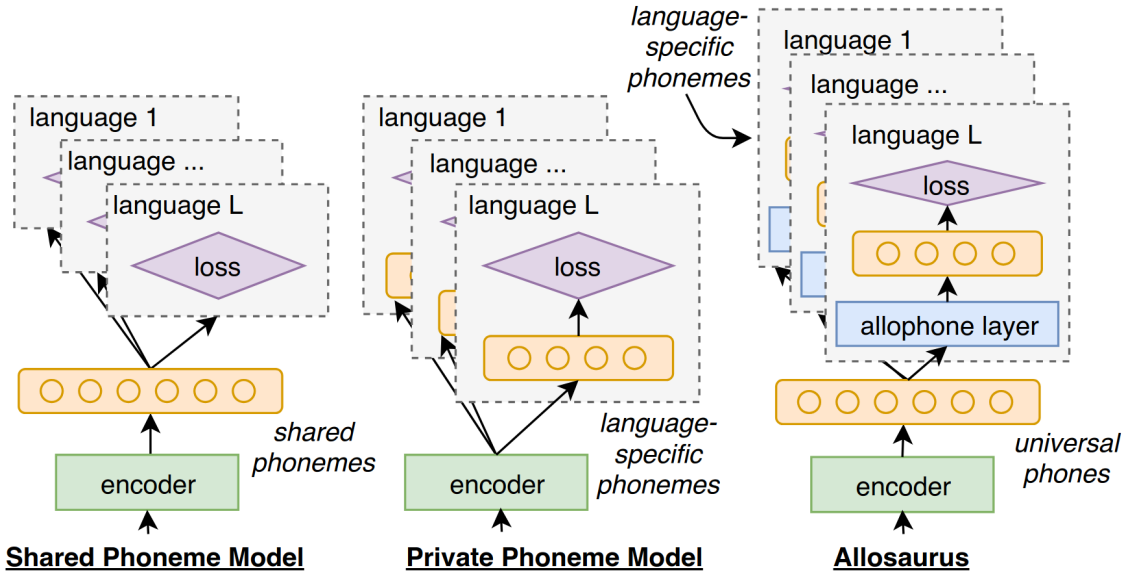
Table: Number of languages supported by large-scale ASR models.

Meta Omnilingual ASR

- Various data collection efforts to increase language coverage.
- Scaled self-supervised pre-training to 7B parameters.
- CTC and LLM-ASR variants (Lecture 18).
- In-context learning for ASR on unseen languages.

Grouping	# of lang	Avg CER	CER ≤ 10	%
Afroasia	92	11.8	61	66%
Amazbasi	83	2.0	82	99%
Amerande	67	2.0	66	99%
Atlacong	389	9.3	280	72%
Austasia	35	5.4	31	89%
Austrone	239	5.1	193	81%
Caucasus	35	3.9	35	100%
Dravidia	22	7.3	18	82%
Indoeuro	209	9.1	154	74%
Mesoamer	159	7.8	115	72%
Newguine	77	5.5	63	82%
Nilosaha	56	4.4	50	89%
Norameri	42	4.8	37	88%
Sinotibe	65	8.2	52	80%
Total	1570	7.1	1237	78%

Multilingual phone recognition

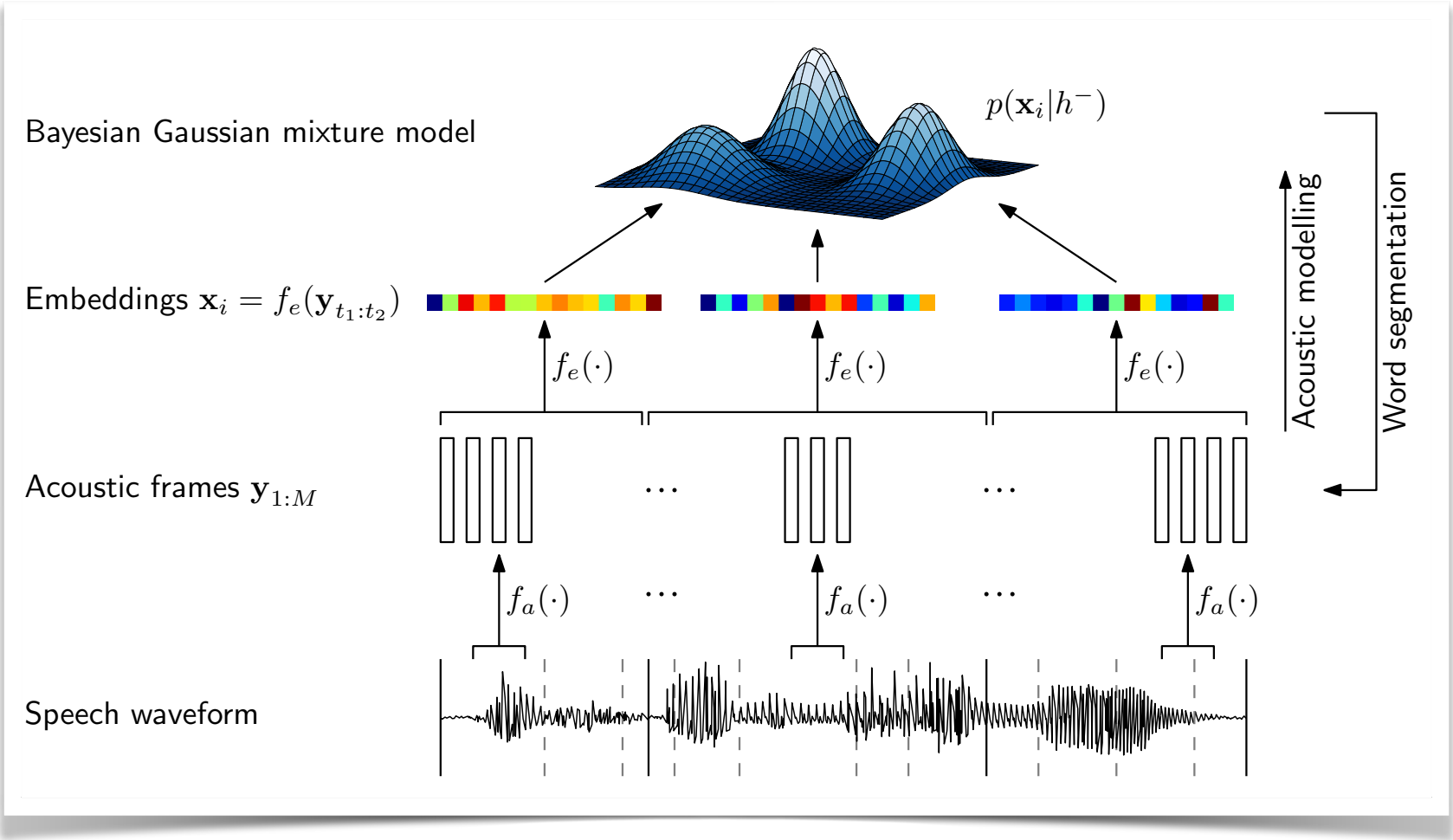


Li et al, 2020

Use-cases:

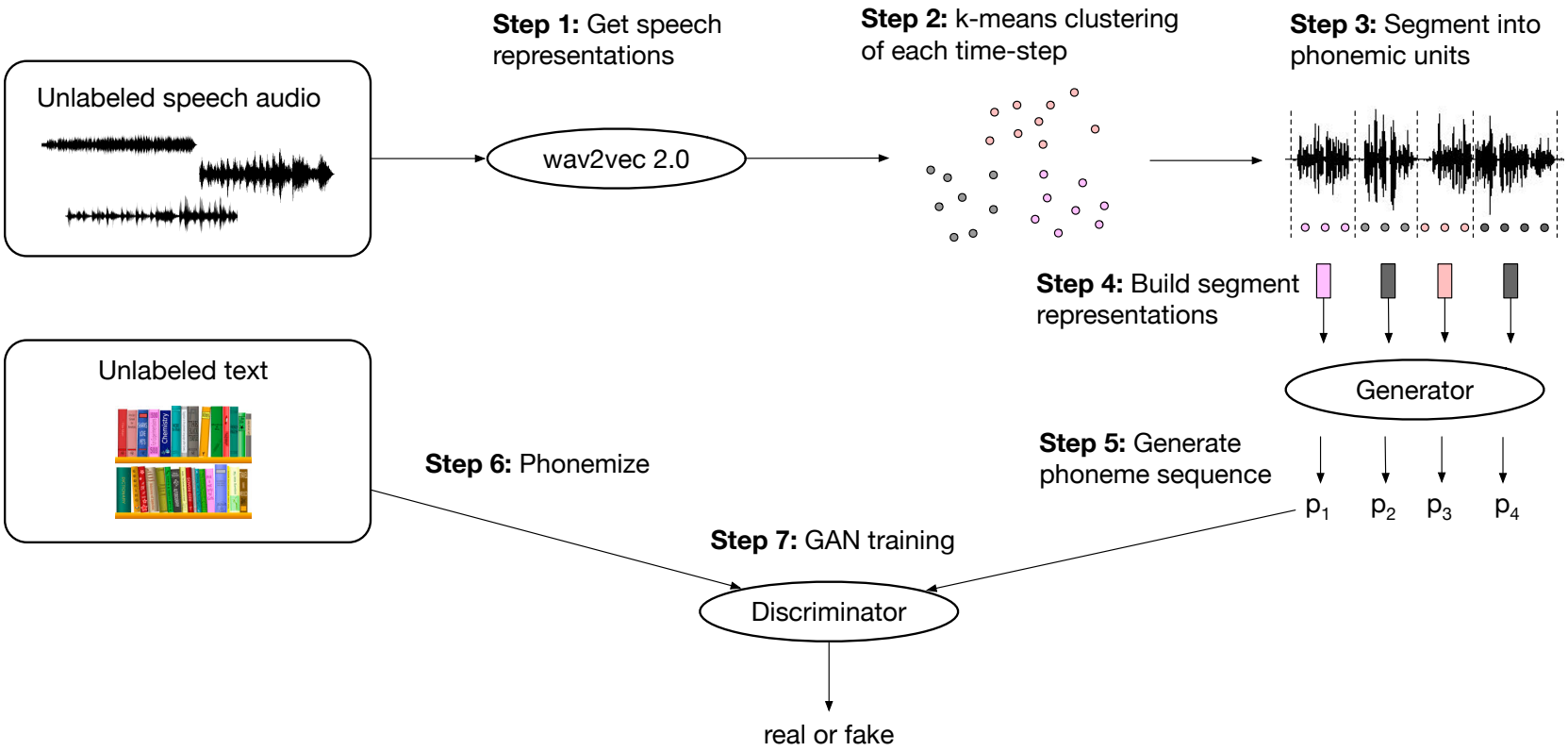
- Scaling ASR to low-resource languages without transcribed speech (Li et al, (2022))
- Voice-search for unwritten languages (Reitmaier et al, (2024))

Bottom-up approaches (1)



Kamper et al, 2017

Bottom-up approaches (2)



Baevski et al, 2021

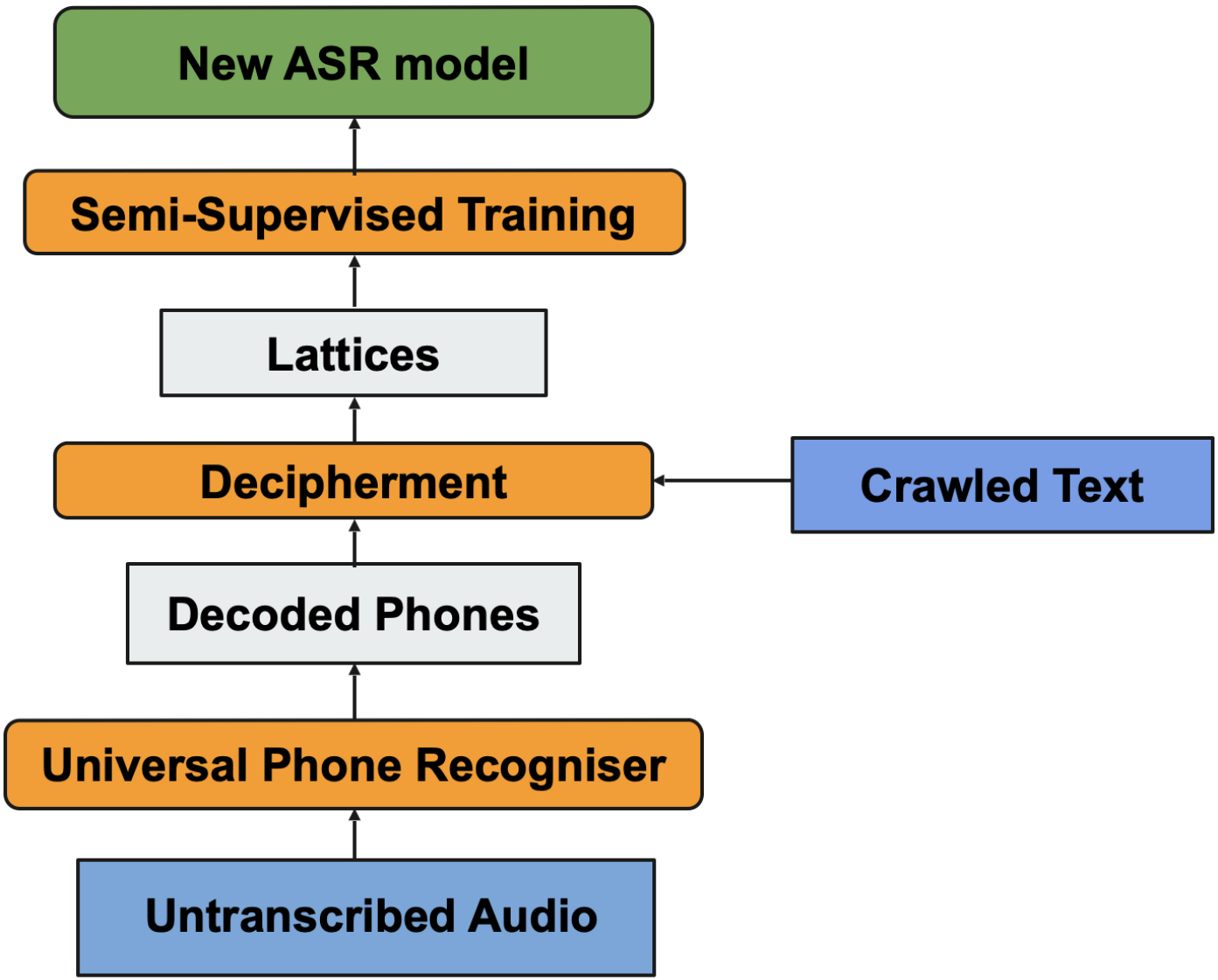
Top-down approaches

- Use a multilingual phone set (IPA, XSAMPA)
- Need a pronunciation model to map from words/graphemes to phones for each new language

$$P(W|X) \propto \sum_{Q} \underbrace{P(X|Q)}_{\substack{\text{Language} \\ \text{universal}}} \times \underbrace{P(Q|W)}_{\substack{\text{Learn for} \\ \text{each language}}} \times \underbrace{P(W)}_{\substack{\text{Train as} \\ \text{normal}}}$$

- Obtain pronunciation model in an unsupervised manner:
  - Universal grapheme-to-phoneme model with a phylogenetic tree to associate closely-related languages (Li et al, 2022)
  - Use graphemes and transliteration for new scripts
  - Learn the mapping using a "decipherment" approach (Klejch et al, 2022)

Decipherment: full pipeline



Decipherment model

Noisy channel model

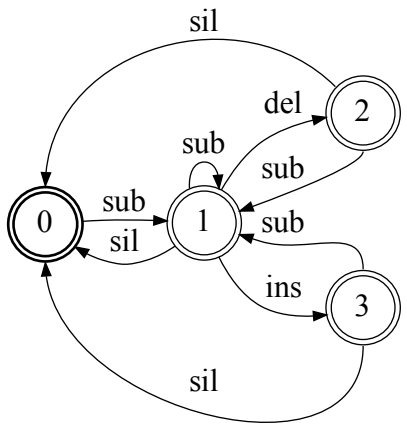
$$\hat{Y} = \underset{Y}{argmax}\, P_{\text{lex}}(X \mid Y) P_{\text{lm}}(Y)$$

Training with the Baum-Welch algorithm and decoding with the Viterbi algorithm (Lecture 5).

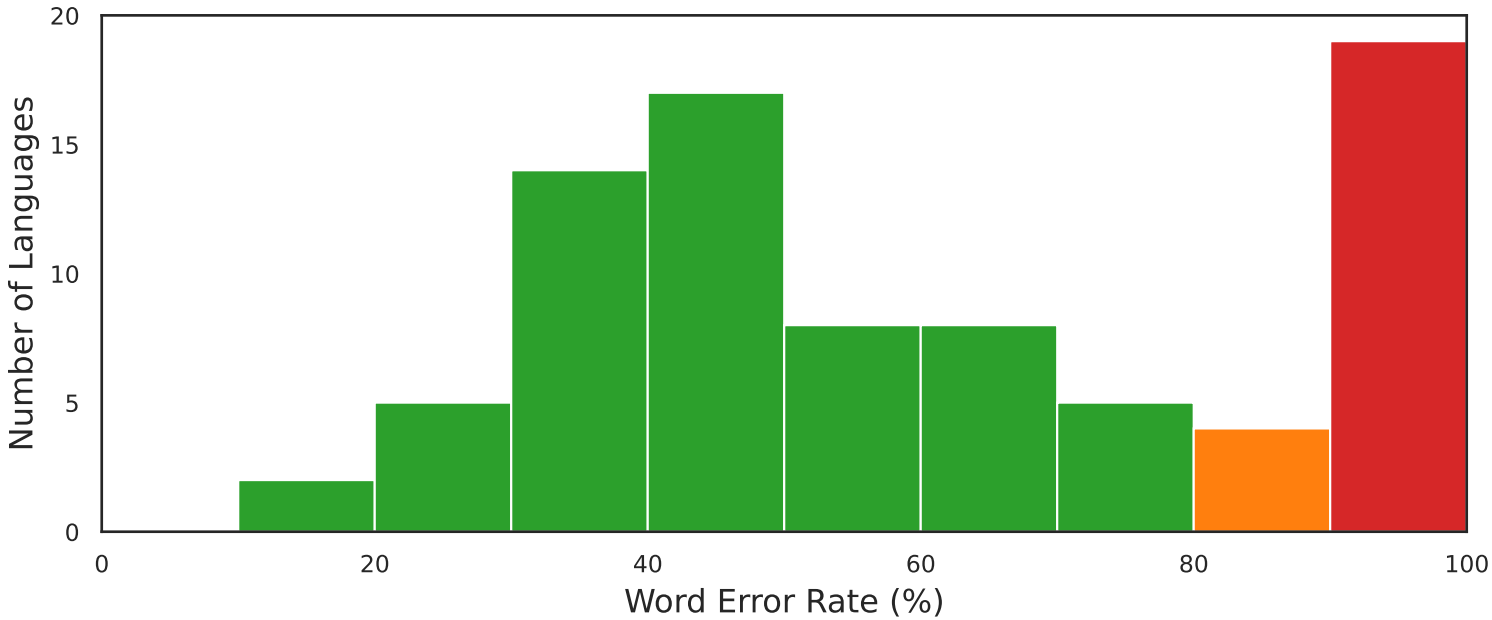
Unsynchronised nondeterministic decipherment

$$\hat{Y} = \underset{Y,A}{argmax}\, P_{\text{lex}}(X \mid Y, A) P_{\text{lm}}(Y) P_{\text{ali}}(A)$$

$$= shortestpath(X \circ (L \circ A) \circ G)$$



Decipherment: FLEURS-100 results



Undeciphered languages: Amharic, Arabic, Fula, Irish, Hebrew, Khmer, Kannada, Lao, Malayalam, Burmese, Nepali, Sindhi, Tamil, Thai, Urdu, Vietnamese, Yoruba.

Speech recognition systems for low-resource languages

- Subword language modeling
- Transferring data between acoustic models based on multilingual hidden representations
- Grapheme/subword-based pronunciation lexica
- Massively multilingual models
- Leveraging untranscribed speech

Reading (1)

- L Besacier et al (2014). "Automatic speech recognition for under-resourced languages: A survey", Speech Communication, 56:85–100. http://www.sciencedirect.com/science/article/pii/S0167639313000988
- Z Tüske et al (2013). "Investigation on cross- and multilingual MLP features under matched and mismatched acoustical conditions", ICASSP. http://ieeexplore.ieee.org/abstract/document/6639090/
- A Ghoshal et al (2013). "Multilingual training of deep neural networks", ICASSP-2013. http://ieeexplore.ieee.org/abstract/document/6639084/
- J-T Huang et al (2013). "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers", ICASSP. http://ieeexplore.ieee.org/abstract/document/6639081/.
- M Gales et al (2015). "Unicode-based graphemic systems for limited resource languages", ICASSP. http://ieeexplore.ieee.org/document/7178960/

Reading (2)

- M. Creutz et al (2007). "Morph-based speech recognition and modeling OOV words across languages", *ACM Trans Speech and Language Processing*, 5(1). http://doi.acm.org/10.1145/1322391.1322394
- P. Swietojanski et al. (2012), "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR", In Proc. IEEE SLT. https://ieeexplore.ieee.org/document/6424230
- A. Conneau, et al. (2020). "Unsupervised cross-lingual representation learning for speech recognition", arXiv:2006.13979. https://arxiv.org/abs/2006.13979
- V. Manohar, et al. (2018) "Semi-supervised training of acoustic models using lattice-free MMI". In Proc. IECC ICASSP (pp. 4844-4848). https://ieeexplore.ieee.org/abstract/document/8462331
- E. Wallington, et al. (2021) "On the learning dynamics of semi-supervised training for ASR". In Proc. Interspeech. https://www.isca-speech.org/archive/interspeech_2021/wallington21_interspeech.html

Reading (3)

- H. Kamper et al (2017). "A segmental framework for fully-unsupervised large vocabulary speech recognition", *Computer Speech & Language*, 46 (pp. 154-174). https://www.sciencedirect.com/science/article/pii/S0885230816301905
- A. Baevski et al (2021), "Unsupervised speech recognition", In NeurIPS 34 https://arxiv.org/abs/2105.11084
- V. Pratap et al (2024), "Scaling speech technology to 1,000+ languages." *Journal of Machine Learning Research*, 25.97 (pp. 1-52). http://www.jmlr.org/papers/v25/23-1318.html
- O. Klejch et al (2022), "Deciphering Speech: a Zero-Resource Approach to Cross-Lingual Transfer in ASR", Proc. Interspeech. https://www.isca-speech.org/archive/interspeech_2022/klejch22_interspeech.html
- L.M. Lam-Yee-Mui et al (2023). "Comparing Self-Supervised Pre-Training and Semi-Supervised Training for Speech Recognition in Languages with Weak Language Models." In Proc. Interspeech. https://www.isca-archive.org/interspeech_2023/lamyeemui23_interspeech.html

Reading (4)

- P. Smit et al (2021), "Advances in subword-based HMM-DNN speech recognition across languages", *Computer Speech & Language*, 66 (pp. 101158). https://www.sciencedirect.com/science/article/pii/S0885230820300917
- X. Li et al (2020). "Universal phone recognition with a multilingual allophone system." In Proc. ICASSP. https://ieeexplore.ieee.org/abstract/document/9054362/
- X. Li et al (2022), "Zero-shot Learning for Grapheme to Phoneme Conversion with Language Ensemble", Findings of the ACL (pp. 2106-2115). https://aclanthology.org/2022.findings-acl.166/
- X. Li et al (2022). "ASR2K: Speech recognition for around 2000 languages without audio." In Proc. Interspeech. https://www.isca-archive.org/interspeech_2022/li22aa_interspeech.html
- P. Wu et al (2021). "Cross-lingual transfer for speech processing using acoustic language similarity." In Proc. ASRU. https://ieeexplore.ieee.org/abstract/document/9688276/
- T. Reitmaier et al. (2024). "Cultivating spoken language technologies for unwritten languages." In Proc. CHI. https://dl.acm.org/doi/full/10.1145/3613904.3642026