Low resource scenarios

AM		LM			
Unlabelled audio	Labelled audio	Unaligned text	Technique/ Problem	Course	
No	Small (<10h)	Yes	Retraining, cotraining, adaptation	Course 1	Standard
Yes	Small (<10h)	Yes	Semi-supervised learning	Course 1	
Yes	No	Yes	Distant supervision	Course 1	
No	Bad labels	?	Weak supervision	Course 2	Non-Standard
Large (>100h)	No	No	Zero resource/ unsupervised	Course 2	
No	No	No	Zero Data / language emergence	Course 2	
Large (>100h)	Small (<10h)	Yes	self-supervised pretraining	Course 3	Standard

This course

Unlabelled audio	Labelled audio	Unaligned text	Technique/ Problem	Course
No	Small (<10h)	Yes	Retraining, cotraining, adaptation	Course 1
Yes	Small (<10h)	Yes	Semi-supervised learning	Course 1
Yes	No	Yes	Distant supervision	Course 1

Standard

Main issue: train a good AM with few (or in the extreme, no) labels
Main ideas:
- Transfer a pretrained AM from other languages
- Create pseudolabels from unlabelled audio
- Use the LM to constrain the discovery of an AM

Unlabelled audio	Labelled audio	Unaligned text	Technique/ Problem
No	Small (<10h)	Yes	Retraining, cotraining, adaptation

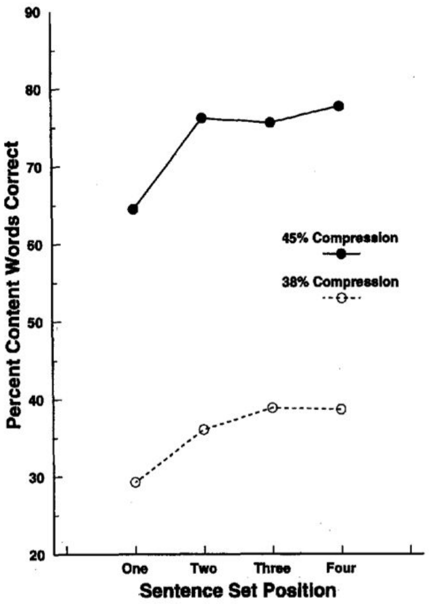
Transfer learning

Distance to a high resource language

- Regional/foreign accent variant
  - Humans: Fast adaptation
  - Machines: adaptation/transfer
- Completely new language
  - Humans: Learning an L2 (painful)
  - Machines:
    - Construct an universal AM
    - retrain/adapt/fine tune the AM
- new language of a given family
  - Humans (somewhat easier)
  - Machines:
    - Mix of the above techniques
    - Joint training

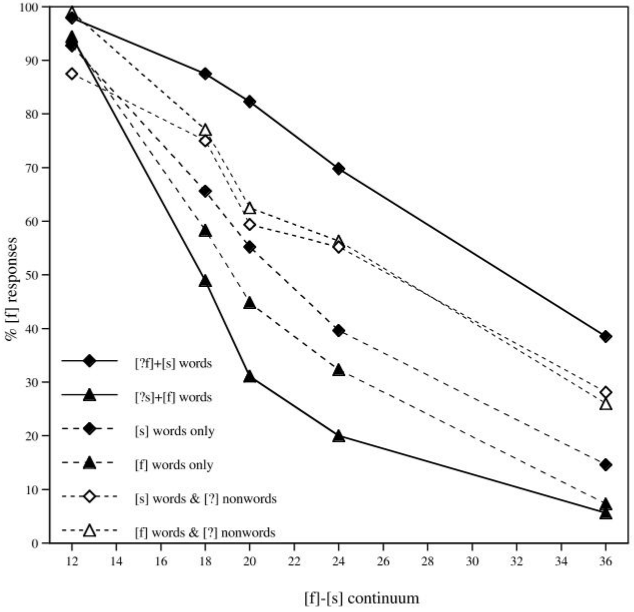
perceptual adaptation to dialect/accents in humans

- Ultra compressed speech (Dupoux & Green 1997)
  - 4 sets of 5 sentences

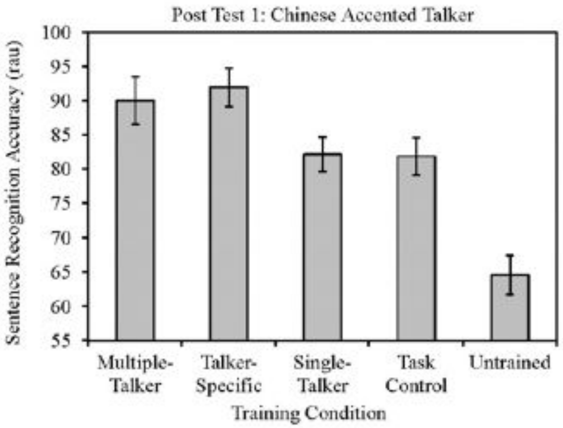


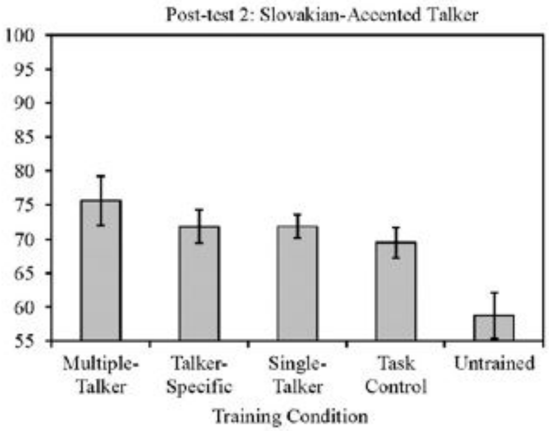
- Artificial dialect (Maye, Aslin, Tanenhaus, 2008)
  - the wicked witch of the west -> the weckup wetch of the wast
  - 20 minutes exposition unsupervised
  - lexical decision: ‘wetch’ -> 69% word

- Adaptation to a shift in a single phoneme (Norris, McQueen Cutler, 2003)
  - 20 words with ambiguous final s/f
  - eg: chrisma[s/f], or belie[s/f]



- accented speech (Bradlow & Bent 2008)
    - training: 5 repetitions of 16 sentences (in noise, no feedback)
    - test: 2 new sets of 16 sentences

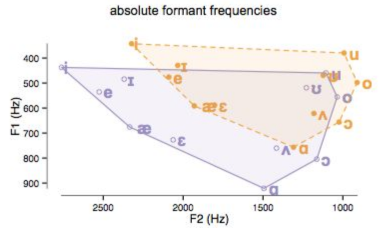




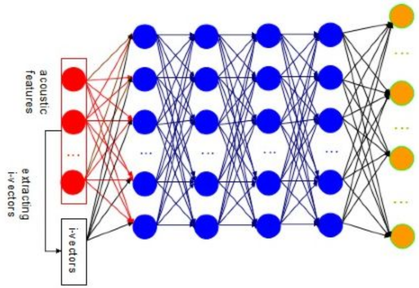
→ note that humans do unsupervised transfer, with few datapoints!

domain adaptation in machines

- fMLLR
  - x□Ax+b to maximize p(x|speaker)

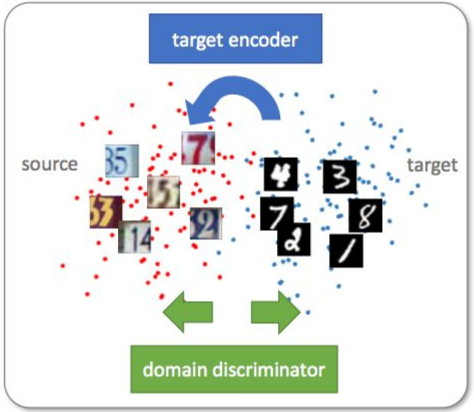


- i-vectors



Chen, Liang et al (2015)

- adversarial domain adaptation



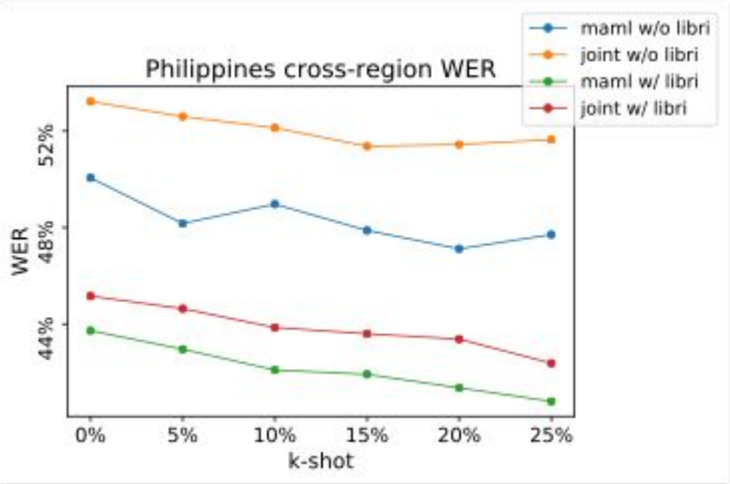
Tzeng, et al 2017

Example of fast domain adaptation

Winata, G. I., Cahyawijaya, S., Liu, Z., Lin, Z., Madotto, A., Xu, P., & Fung, P. (2020). Learning fast adaptation on cross-accented speech recognition. *arXiv preprint arXiv:2003.01901*.

- Common voice data (english annotated by accent)
- Use of meta learning to quickly adapt to a new accent

accents	# sample	duration (hr)
Africa (**af**)	4,065	5.04
Australia (**au**)	19,625	22.86
Bermuda (**be**)	363	0.46
Canada (**ca**)	17,422	20.20
England (**en**)	58,274	64.19
Hong Kong (**hk**)	1,181	1.21
India (**in**)	23,878	29.09
Ireland (**ir**)	3,420	3.71
Malaysia (**my**)	843	1.07
New Zealand (**nz**)	6,070	7.06
Philippines (**ph**)	1,318	1.68
Scotland (**sc**)	4,376	5.08
Singapore (**sg**)	693	1.00
South Atlantic (**sa**)	212	0.23
United States (**us**)	145,692	163.89
Wales (**wa**)	1,128	1.16
Total	288,560	327.93



Distance to a high resource language

- Regional/foreign accent variant
  - Humans: Fast adaptation
  - Machines: adaptation/transfer
- Completely new language
  - Humans: Learning an L2 (painful)
  - Machines:
    - Construct an universal AM
    - retrain/adapt/fine tune the AM
- new language of a given family
  - Humans (somewhat easier)
  - Machines:
    - Mix of the above techniques
    - Joint training

Second language learning in humans

- In infants:
  - Fast, easy
  - No supervision
  - No mixup/interference between languages

  → we will come back to this

- In adults
  - Slow, difficult
  - Requires supervision
  - Close languages help (but also confuses)

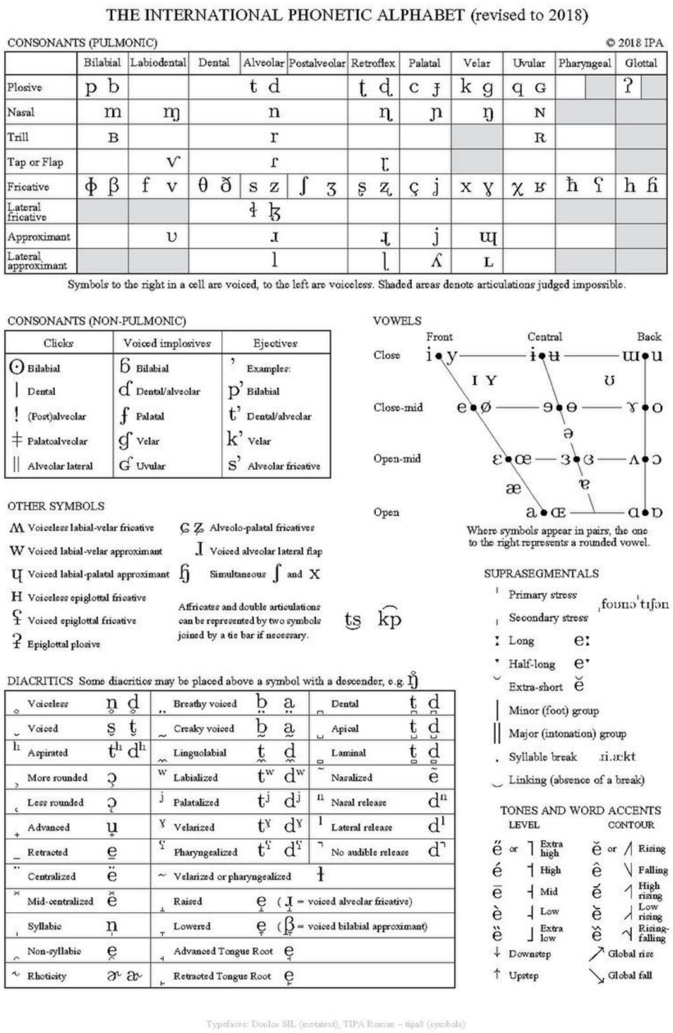
  Eg; bin vs bean (for French learners); right vs light (for Japanese learners), bébe vs bebé (for French listeners)

Learning in machines

- The task: construct a good AM, with few labels
- The main idea:
  1. Construct a universal AM
     a. Universal phoneme sets
     b. Universal articulatory features
     c. Universal embeddings
     d. Universal character sets
  2. Two ideas:
     a. Adapt this AM to the language with few labels
     b. Learn everything jointly

Construct a universal phone set

- International Phonetic Alphabet (IPA)
- resources : phonemizer
- Easy to adapt to a new languages: just learn a new G2P
- Example: Manjunath, K. E., Raghavan, K. S., Rao, K. S., Jayagopi, D. B., & Ramasubramanian, V. (2019). Multilingual Phone Recognition: Comparison of Traditional versus Common Multilingual Phone-Set Approaches and Applications in Code-Switching.



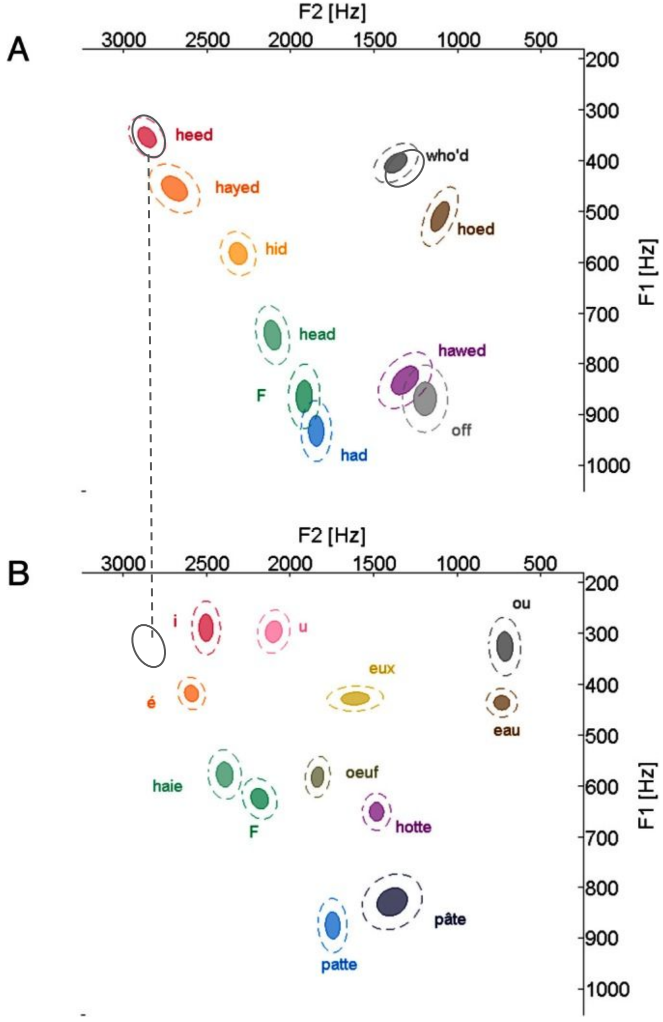
Construct a universal phone set

Problems

- consonants and vowels are distributions over continuous phonetic space
- no two language use the same distributions
- Therefore learning a single symbol for these distributions may just blur them and introduce confusions

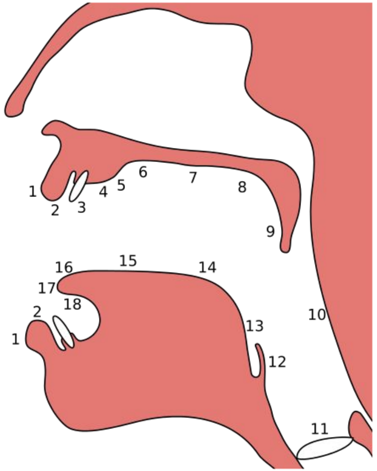
English

French



→ may not work very well, because the universal phone set is not really universal

Construct a universal set of articulatory features



Passive and active places of articulation: (1) *Exo-labial*; (2) *Endo-labial*; (3) *Dental*; (4) *Alveolar*; (5) *Post-alveolar*; (6) *Pre-palatal*; (7) *Palatal*; (8) *Velar*; (9) *Uvular*; (10) *Pharyngeal*; (11) *Glottal*; (12) *Epiglottal*; (13) *Radical*; (14) *Postero-dorsal*; (15) *Antero-dorsal*; (16) *Laminal*; (17) *Apical*; (18) *Sub-apical* or *sub-laminal*.

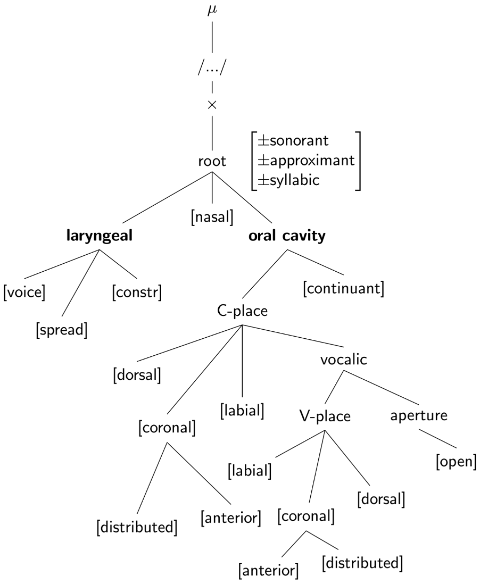


TABLE 10.5

The Distinctive Features of Consonants and Vowels

Consonants and Liquids

Distinctive Feature	p	b	t	d	č	ǰ	k	g	f	v	θ	ð	s	z	š	ž	r	l	m	n	ŋ
Consonantal	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Vocalic	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+	−	−	−
Anterior	+	+	+	+	−	−	−	−	+	+	+	+	+	+	−	−	−	+	+	+	−
Coronal	−	−	+	+	+	+	−	−	−	−	+	+	+	+	+	+	+	+	−	+	−
Voice	−	+	−	+	−	+	−	+	−	+	−	+	−	+	−	+	+	+	+	+	+
Nasal	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+	+
Strident	−	−	−	−	+	+	−	−	+	+	−	−	+	+	+	+	−	−	−	−	−
Continuant	−	−	−	−	−	−	−	−	+	+	+	+	+	+	+	+	+	+	−	−	−

Vowels and Glides

Distinctive Feature	i	ɪ	e	ɛ	æ	ɨ	ə	ʌ	a	u	ʊ	o	ɔ	y	w	h
Vocalic	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−
Consonantal	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
High	+	+	−	−	−	+	−	−	−	+	+	−	−	+	+	−
Back	−	−	−	−	−	+	+	+	+	+	+	+	+	−	+	−
Low	−	−	−	−	+	−	−	+	+	−	−	−	+	−	−	+
Round	−	−	−	−	−	−	−	−	−	+	+	+	+	−	+	−
Tense	+	−	+	−	+	−	−	−	+	+	−	+	−	−	−	−

Source: Clark and Clark, 1977.

example

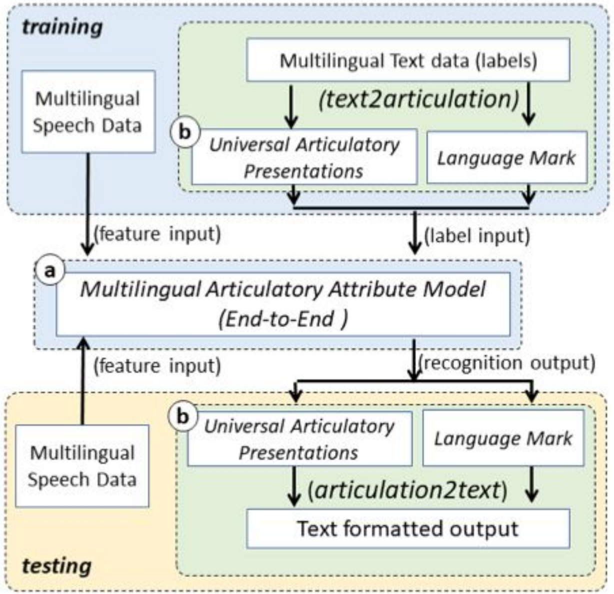
Li S, Ding C, Lu, X. Shen, P., Kawahara, T, Kawai H (2019). End-to-End Articulatory Attribute Modeling for Low-resource Multilingual Speech Recognition. Interspeech

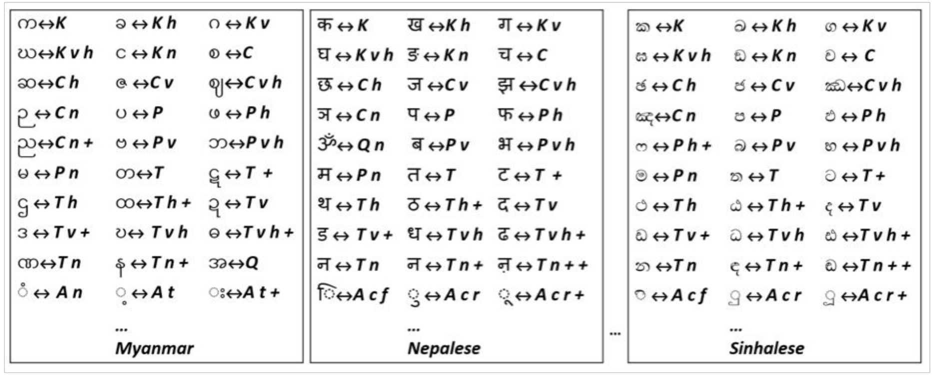
(see also Li, X, Dalmia, S, Mortensen, DR, Li, J, Black, AW, Metze F (2020). Towards Zero-shot Learning for Automatic Phonemic Transcription)

→ may not work very well, because the universal feature set is not really universal

Table 1: *Universal Articulatory Representations*

Catergories	Attributes
Consonants (placement)	Velar (**K**) Palatal (**C**) Coronal (**T**) Labial (**P**) Glottal (**Q**)
Consonants (manner)	Aspired (**h**) Voiced (**v**) Nasal (**n**) Trill (**R**) Lateral approximant (**L**) Labial/Labio-velar approximant (**W**) Palatal approximant (**Y**) Sibilant fricative (**S**) Non-sibilant fricative (**H**)
Vowel (**A**)	Round (**r**) Front (**f**) Close (**c**) Tonal (**t**) Visarga (**h**) Anunasika (**n**)
Special Marks	Repeat Removal (**+**)





Common character sets

- Liu, C. Zhang, Q. Zhang, X., Singh, K. Saraf, Y., Zweig, G. (2020). Multilingual Graphemic Hybrid ASR with Massive Data Augmentation
  - hybrid system (DNN, HMM - WFST)
  - clustered trigrapheme units (some overlap in char sets but also char specific)
  - $H \circ C \circ L \circ G$. (red: language independant; blue, language specific or not)

→ multilingual better than monolingual

→ language independant decoding works better than language specific

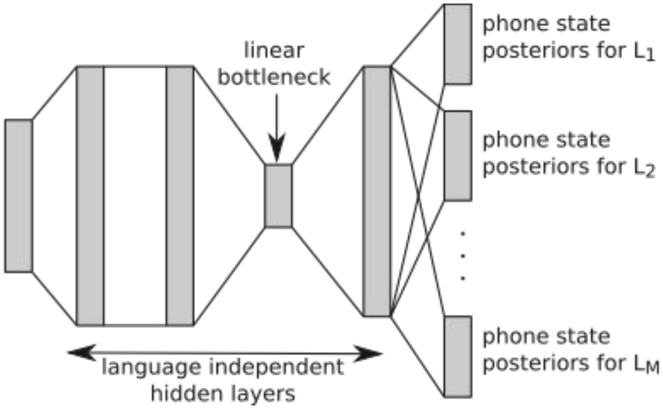
		%gain of 7 lang vs monolang
Kannada	125.5	7.1
Malayalam	127.7	5.0
Sinhala	160.0	4.6
Tamil	176.9	5.0
Bengali	160.0	7.8
Hindi	160.0	10.3
Marathi	148.6	7.6

final wer: ~ 50%

- A bit brutal: in the spirit of end-to-end models
- May preserve language specificity (though context dependant units)
- Does not require phonetic annotations

Construct universal embeddings

- Preserves all continuous information
- Adaptation:
  - Learn a classifier/decoder for a new language
  - Apply domain adversarial training to move the target language closer

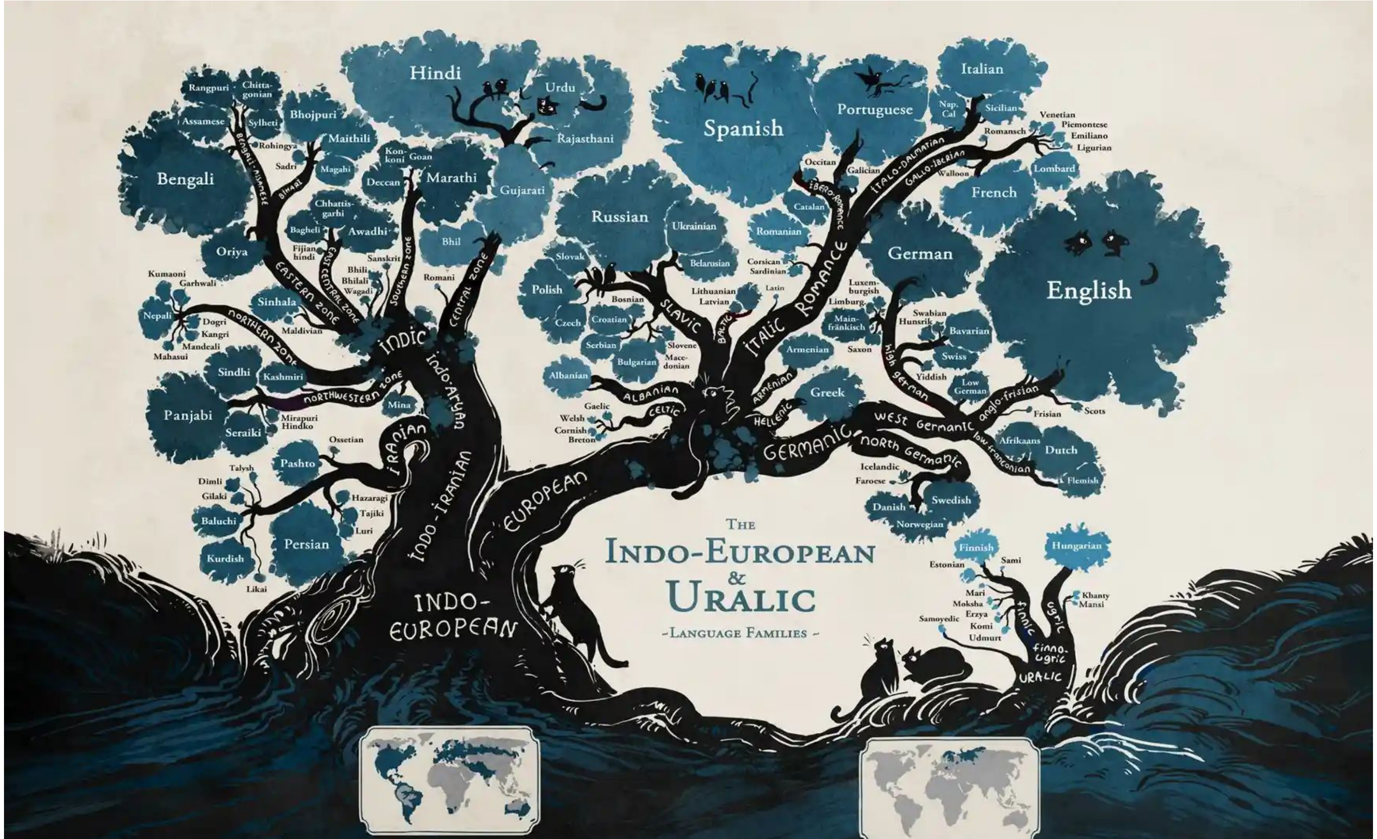


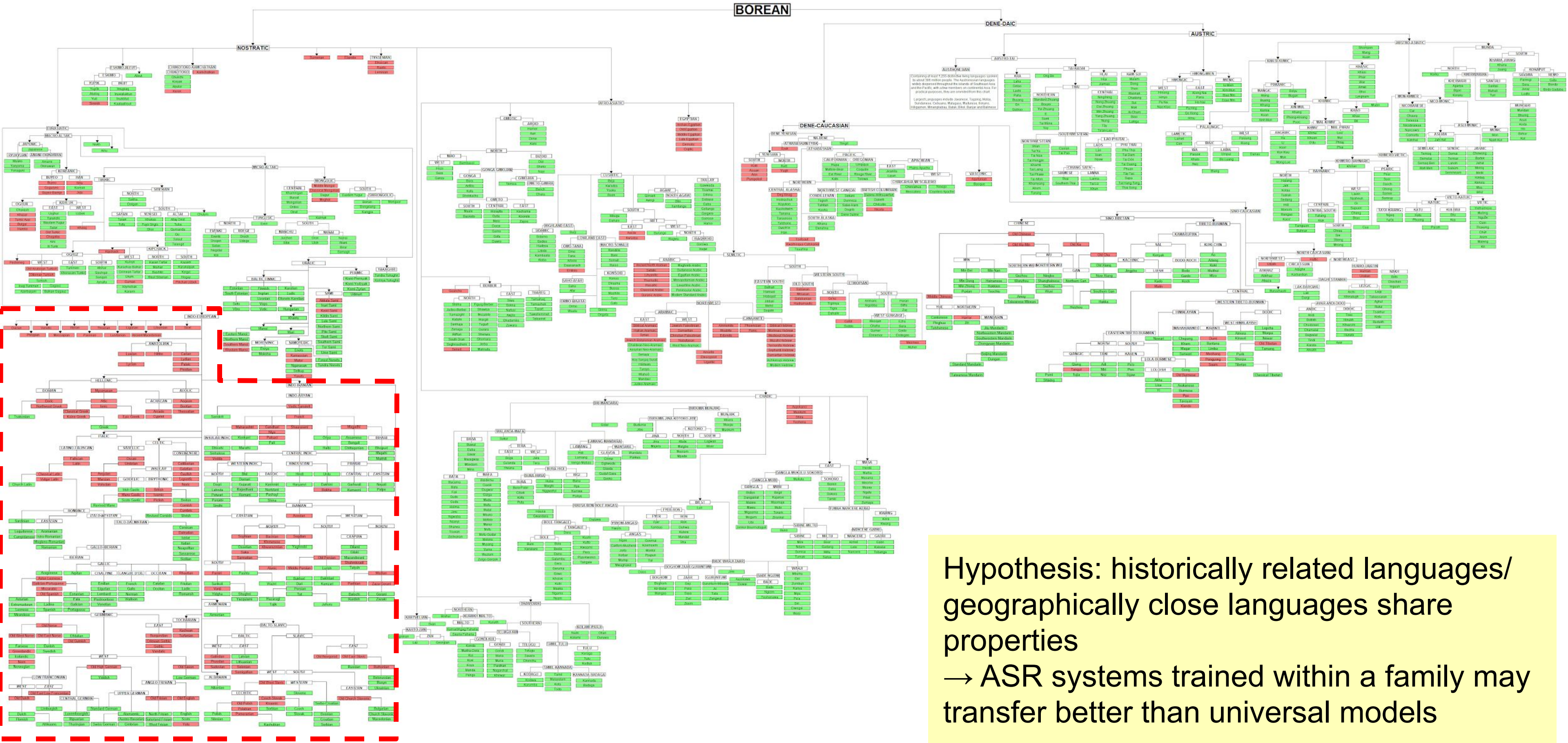
Fer, R., Matějka, P., Grézl, F., Plchot, O., Veselý, K., & Černocký, J. H. (2017). Multilingually trained bottleneck features in spoken language recognition. *Computer Speech & Language*, *46*, 252-267.

Distance to a high resource language

- Regional/foreign accent variant
  - Humans: Fast adaptation
  - Machines: adaptation/transfer
- Completely new language
  - Humans: Learning an L2 (painful)
  - Machines:
    - Construct an universal AM
    - retrain/adapt/fine tune the AM
- new language of a given family
  - Humans (somewhat easier)
  - Machines:
    - Mix of the above techniques (restricting to the family)
    - Joint training

Language families





Hypothesis: historically related languages/ geographically close languages share properties
→ ASR systems trained within a family may transfer better than universal models

Joint learning of everything

- Liu, C. Zhang, Q. Zhang, X., Singh, K. Saraf, Y., Zweig, G. (2020). Multilingual Graphemic Hybrid ASR with Massive Data Augmentation
    - hybrid system (DNN, HMM - WFST)
    - clustered trigrapheme units (some overlap in char sets but also char specific)
    - $H \circ C \circ L \circ G$. (red: language independant; blue, language specific or not)

→ multilingual better than monolingual

→ language independant decoding works better than language specific

→ **family specific works better (less data, more relevant)**

		%gain of 7 lang vs monolang	%gain of 3-4 lang vs monolang
Kannada	125.5	7.1	**7.5**
Malayalam	127.7	5.0	**5.3**
Sinhala	160.0	**4.6**	**4.6**
Tamil	176.9	**5.0**	**5.0**
Bengali	160.0	7.8	**9.5**
Hindi	160.0	**10.3**	**10.3**
Marathi	148.6	**7.6**	**7.6**

final wer: ~ 50%

Summary

- Using labels from other languages!
  - Language transfer: using a very close language
  - Universal models: using as many languages as possible (more data)
  - A compromise: language family
- What to learn from these external labels
  - universal embeddings
  - universal phone set (IPA)
  - universal articulatory feature set
  - ‘universal ‘ grapheme set
- How to use the few labels in the target language
  - Transfer
    - Learn a classifier (from fixed embeddings)
    - Fine tune the embeddings
    - Map the units in the target to the units from the source(s)
  - Joint learning
  - Domain adversarial training

Unlabelled audio	Labelled audio	Unaligned text	Technique/ Problem
No	Small (<10h)	Yes	Retraining, cotraining, adaptation
Yes	Small (<10h)	Yes	Semi-supervised learning

Semi-supervised training (or, how to create pseudo labels)

Main idea

- Train a (small) system on your few labels
- Use it to generate labels on new unlabeled data
- (optional) remove the suspicious pseudo-labels
- Train a larger/retrain your system on the total data
- Iterate!

	Dutch			Romanian		
	clean	noisy	extreme	clean	noisy	extreme
Supervised baseline						
Hybrid (LFMMI)	23.6	23.3	32.8	17.5	19.5	32.9
CTC	26.7	26.3	36.8	20.8	22.2	38.3
Enc-Dec	27.2	27.0	39.0	25.5	26.8	46.0
Self-labeling using frame-level distillation						
Top 3 (CE)	23.3	22.8	32.3	15.7	17.9	29.6
+ LFMMI	21.3	20.7	29.8	14.7	17.0	28.8
Top 1 (CE)	22.9	22.6	32.0	15.9	17.8	29.7
+ LFMMI	21.7	21.6	30.9	14.7	16.9	28.6
Self-labeling using sequence-level distillation						
Hybrid (CE)	23.6	23.3	32.6	16.3	18.4	30.7
+ LFMMI	20.9	20.8	29.7	14.7	16.8	28.3
SL-LFMMI	22.1	21.8	31.4	15.6	17.6	29.5
CTC	22.5	22.2	31.4	14.9	17.2	29.4
Enc-Dec	18.4	18.5	27.9	13.1	15.6	27.3

+37% rel

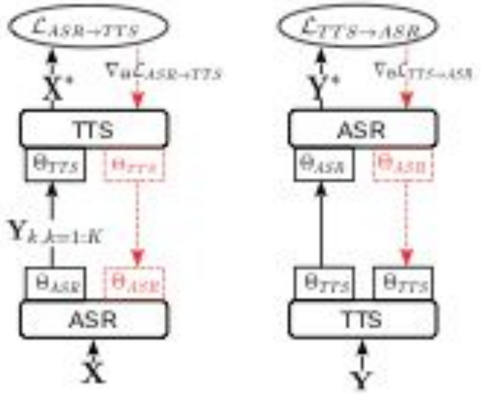
Singh, K. Manohar, V., Xiao A. Edunov, S. Girshick, R., Liptchinsky, V., Fuegen, C. Saraf, Y., Zweig, G Mohamed, A. (2020) Large scale weakly and semi-supervised learning for low-resource video ASR

- frame -level distillation -- for hybrid models (replace the ground truth labels by probability distributions from the teacher)
- Sequence level distillation -- fot CTC or seq to seq (train with the 1-best decoding from the teacher)
- Training: 290 h (dutch); 193h (romanian) + 6000-8000 hours of (noisy), unlabelled speech
- 3 iterations of relabelling

Other ideas

- Baskar, MK, Watanabe†, S., Astudillo, R., Hori, T., Burget, L., Cˇernocky, J. (2020). Semi-supervised Sequence-to-sequence ASR using Unpaired Speech and Text

  → Use a TTS to generate speech for unpaired text; use ASR to generate text for unlabelled speech. (Similar to back translation in neural translation)



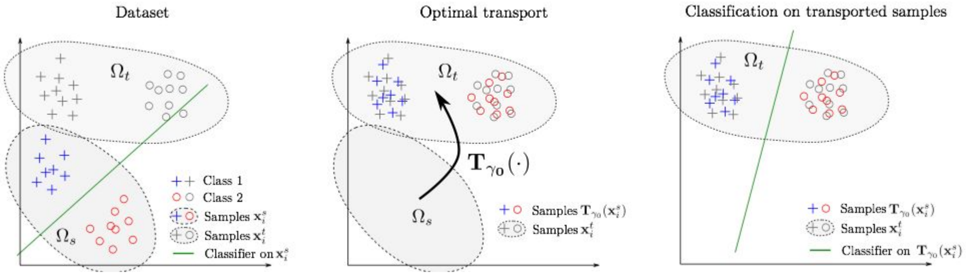
See also Ren, Y, Tan, X. Qin, T. Zhao, S. Zhao, Z.; Liu TY (2019). Almost Unsupervised Text to Speech and Automatic Speech Recognition

Unlabelled audio	Labelled audio	Unaligned text	Technique/ Problem
No	Small (<10h)	Yes	Retraining, cotraining, adaptation
Yes	Small (<10h)	Yes	Semi-supervised learning
Yes	No	Yes	Distant supervision

Distant supervision

Main idea

- The problem is similar to that of deciphering a coded message
  - you don’t know the message, you don’t know the code
  - But if you know the language, you may break the code
- Similar ideas are used in unaligned (unsupervised) translation:
  - Learning within modality embedding space (one for text, one for speech), and realign them
- Two JSALT workshops



→ very difficult! Not yet, there!

(see Li, X, Dalmia, S, Mortensen, DR, Li, J, Black, AW, Metze F (2020). Towards Zero-shot Learning for Automatic Phonemic Transcription)

summary

- The few labels case is difficult
- Requires to understand the underlying problem (no longer brute force machine learning)
- Linguistics is useful
- Having a look at how humans do it could be inspiring
- Some interesting techniques from domain adaptation (vision), and translation (NLP) are being tried.