Multilingual and Low-Resource Speech Recognition

Steve Renals

Automatic Speech Recognition – ASR Lecture 14
7 March 2019

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 60 languages and more than 50 accents/dialects of those languages; Apple Siri covers over 20 languages and about 20 accents/dialects

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 Besaciera et al (2014) for more

Speech recognition of under-resourced languages

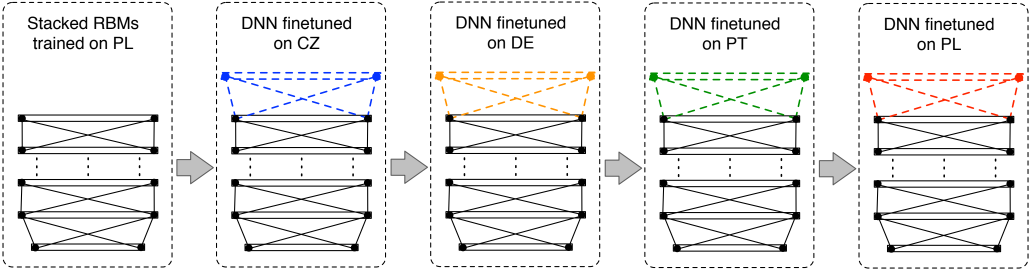
- Training acoustic and language models with limited training data
- Transferring knowledge between languages
- Constructing pronunciation lexica
- Dealing with language specific characteristics (e.g. morphology)

Multilingual and cross-lingual acoustic models

How to share information from acoustic models in different languages?

- General principal – use neural network hidden layers to learn a multilingual representation
- Hidden layers are multilingual – shared between languages
- Output layer is monolingual language specific
- **Hat swap** use a network with multilingual hidden representations directly in a hybrid DNN/HMM systems
- **Block softmax** 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

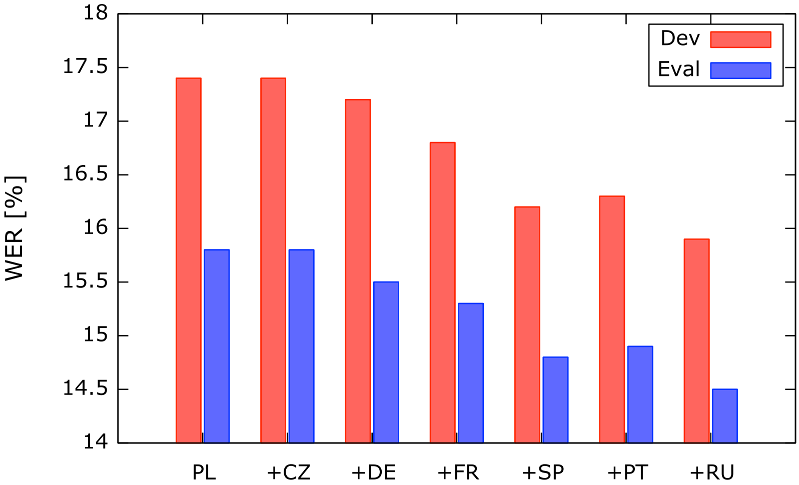
Hat Swap – architecture



Ghoshal et al, 2013

Hat Swap – experiment

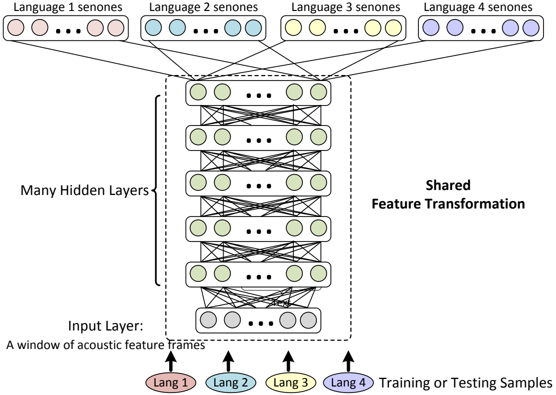
Recognition of GlobalPhone Polish



Block softmax

- In block softmax we train one network for all languages:
    - separate output layer for each language
    - shared hidden layers
- Each training input is propagated forward to the output layer of the corresponding language – only that output layer is used to compute the error used to train the network for that input
- Since the hidden layers are shared, they must learn features relevant to all the output layers (languages)
- Can view block softmax as a parallel version of hat swap

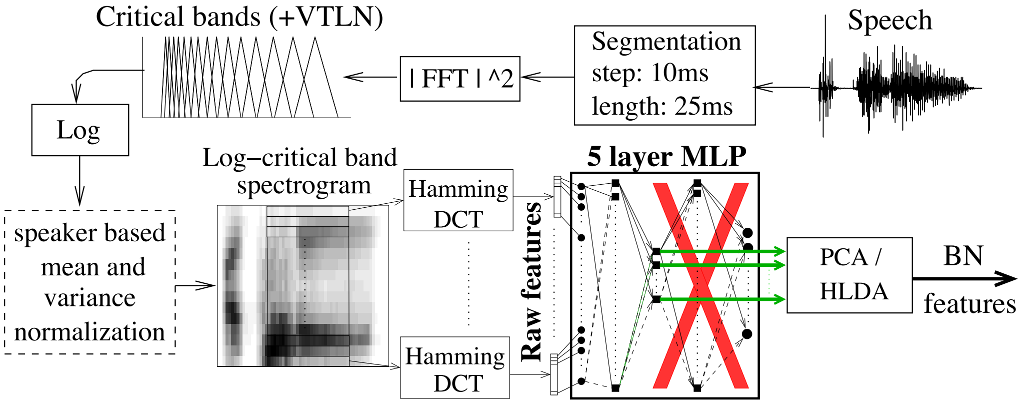
Block softmax – architecture



Huang et al, 2013

NB: A senone is a context-dependent tied state

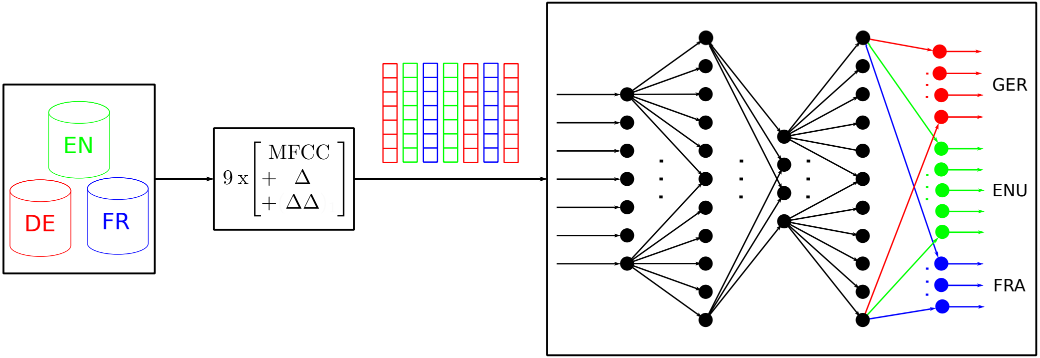
Bottleneck features



Grezl and Fousek (2008)

- Use a "bottleneck" hidden layer to provide features as input to a GMM or an NN
- Decorrelate the hidden layer using PCA (or similar)

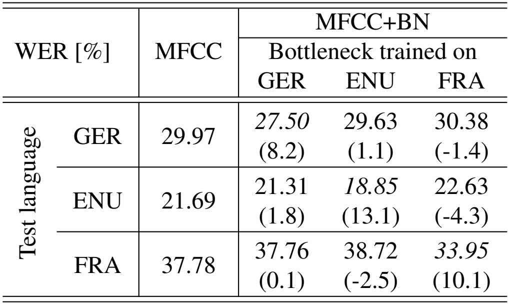
Multilingual bottleneck features – architecture



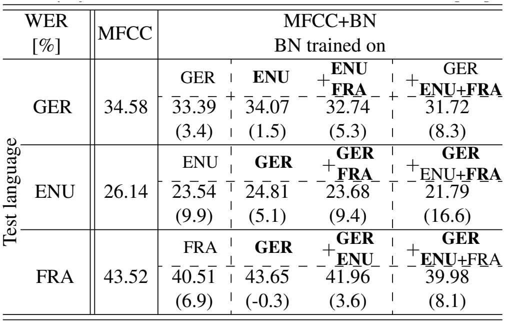
Tüske et al, 2013

Multilingual bottleneck features – experiments

GMM-based acoustic models. (Similar results obtained using multilingual bottleneck features with NN-based acoustic models.)



	WER [%]	MFCC	MFCC+BN		
			Bottleneck trained on		
			GER	ENU	FRA
Test language	GER	29.97	*27.50* (8.2)	29.63 (1.1)	30.38 (-1.4)
	ENU	21.69	21.31 (1.8)	*18.85* (13.1)	22.63 (-4.3)
	FRA	37.78	37.76 (0.1)	38.72 (-2.5)	*33.95* (10.1)



	WER [%]	MFCC	MFCC+BN BN trained on			
Test language	GER	34.58	GER	**ENU**	**ENU** + **FRA**	GER + **ENU+FRA**
			33.39 (3.4)	34.07 (1.5)	32.74 (5.3)	31.72 (8.3)
	ENU	26.14	ENU	**GER**	**GER** + **FRA**	**GER** + ENU+**FRA**
			23.54 (9.9)	24.81 (5.1)	23.68 (9.4)	21.79 (16.6)
	FRA	43.52	FRA	**GER**	**GER** + **ENU**	**GER** + **ENU**+FRA
			40.51 (6.9)	43.65 (-0.3)	41.96 (3.6)	39.98 (8.1)

(Mismatched acoustic environment)

Tüske et al, 2013

Graphemes and phonemes

- Can represent pronunciations as a sequence of graphemes (letters) rather than a sequence of phones
- Advantages of grapheme-based pronunciations
  - No need to construct/generate phone-based pronunciations
  - Can use unicode attributes to assist in decision tree construction
- Disadvantages: not always direct link between graphemes and sounds (most of in English)

Grapheme-based ASR results for 6 low-resource languages

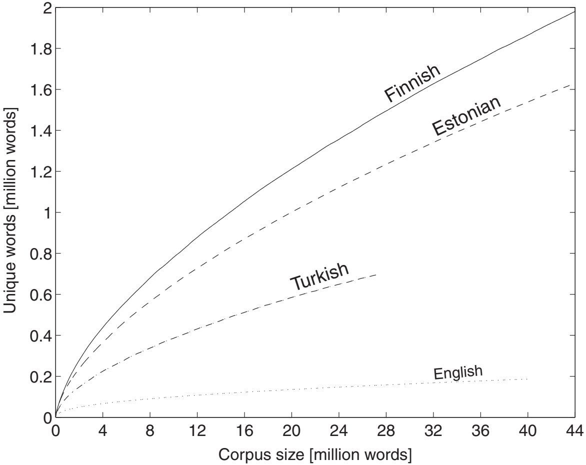
Language	ID	System	WER (%)		
			tg	+cn	cnc
Kurmanji Kurdish	205	Phonetic	67.6	65.8	64.1
		Graphemic	67.0	65.3	
Tok Pisin	207	Phonetic	41.8	40.6	39.4
		Graphemic	42.1	41.1	
Cebuano	301	Phonetic	55.5	54.0	52.6
		Graphemic	55.5	54.2	
Kazakh	302	Phonetic	54.9	53.5	51.5
		Graphemic	54.0	52.7	
Telugu	303	Phonetic	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)

Morphology

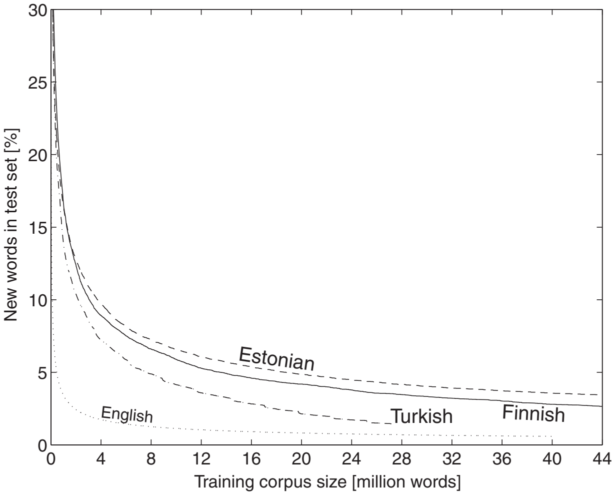
- Many languages are morphologically richer than English: this has a major effect of vocabulary construction and language modelling
- Compounding (eg German): decompose compund words into constituent parts, and carry out pronunciation and language modelling on the decomposed parts
- Highly inflected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

Vocabulary size for different languages



Creutz et al (2007)

OOV Rate for different languages



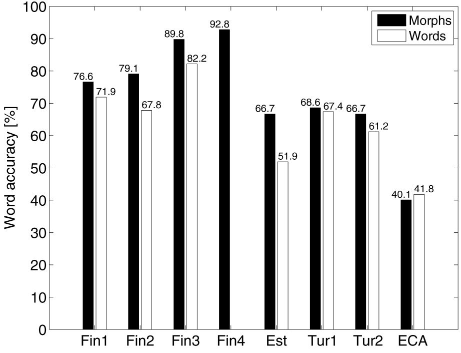
Creutz et al (2007)

Segmenting into morphs

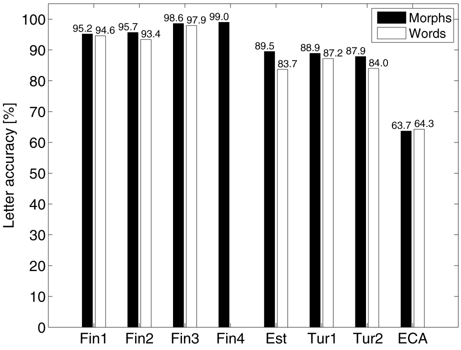
- Linguistic rule-based approaches – require a lot of work for an under-resourced language!
- Automatic approaches – use automatically segment and cluster words into their constitutent morphs
- Morfessor (http://www.cis.hut.fi/projects/morpho/)
  - "Morfessor is an unsupervised data-driven method for 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
- Morph-based language modelling uses morphs instead of words – may require longer context (since multiple morphs correspond to one word)

Morph-based vs Word-based ASR

Speech recognition accuracies on Finnish (Fin1-Fin4), Estonian (Est), Turkish (Tur), and Egyptian Arabic (ECA), using morph- and word-based n-gram language models.



word accuracies



letter accuracies

Creutz et al (2007)

Speech recognition systems for low-resource languages

- Transferring data between acoustic models based on multilingual hidden representations
- Grapheme-based pronunciation lexica
- Morph-based language modeling

Reading

- L Besaciera 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/
- 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