

The Aalto submission to the MGB-3 challenge



Peter Smit

Siva Reddy Gangireddy

Sami Virpioja

Mikko Kurimo

<https://petersmit.eu/post/arabicspeech/>

The Winning submission to the MGB-3 challenge

Peter Smit

Siva Reddy Gangireddy

Sami Virpioja

Mikko Kurimo



Aalto University
School of Electrical
Engineering

Peter Smit

inscripta }

Doctoral student from 2012-2018

Chief Scientific Officer since August 2018

`Modern subword-based models for
automatic speech recognition'

ASR for transcription of Medical Dictations.
In any language.

Contents

- MGB-3 challenge system
- Subword based Automatic Speech Recognition
- Demonstration of the MGB speech recognizer

About the MGB-3 challenge

- Organized by Dr. Ali et al. (QCRI / University of Edinburgh). Official challenge for 2017 ASRU workshop.
- ASR TASK 1: Make ASR system for Egyptian Arabic, using 1200h Al Jazeera data + 5h dialectal data
- ASR TASK 2: MGB-2, ASR for 'general' Arabic, using 1200h Al Jazeera data
- Challenges:
 - Very limited amount of dialectal training data
 - Al Jazeera data not perfect
 - No standard orthography for Egyptian
 - No Arabic speakers in the team
 - Limited time to submit system (< 6 weeks)

The challenge

- Arabic MGB-3 transcription challenge: build a speech-to-text system for transcribing Egyptian Arabic.
- Arabic MGB-2 (re-run): transcription system for Modern Standard Arabic (MSA).
- Audio data: 1200 hrs MSA + 6 hrs Egyptian Arabic.
LM data: 121M tokens.

Results

	MGB-3		MGB-2	
	Dev.	Eval.	Dev.	Eval.
Primary	28.2	29.3	14.8	13.2
Contr. 1 (char)	31.8	31.3	16.3	14.4
Contr. 2 (sub-17k)	31.3	30.5	15.9	14.0
Contr. 3 (word)	31.3	31.2	16.3	14.3

TOP-3

	MGB-3		MGB-2
	MR-WER	AV-WER	WER
Aalto	29.3	37.5	13.2
NDSC-THUEE	32.5	40.7	14.5
JHU	32.8	40.7	16.0

Why did we succeed

- Effective acoustic model adaptation.
- Use of multiple different lexical units (words, subwords, characters).
- Wide variety of tools (Kaldi, Morfessor, TheanoLM, VariKN).

Acoustic modeling

Organizer's baseline

- The provided baseline was a model trained on the MGB-2 data. The acoustic model was a Time-Delay Neural Network (TDNN) trained on all MSA data, without any filtering.
- The language model was a **3-gram** trained on the provided background text corpus.
- For acoustic modeling and decoding **Kaldi** was used, for language modeling **SRILM**.

Aalto baseline

relative improvement **+11%**

- Our baseline was the same as the organizer's, except for
- We used only data with high confidence for GMM training.
- All provided training data was **automatically cleaned and segmented** using standard Kaldi scripts. This resulted in **1022 hours** of cleaned data.
- The language model was a small **varigram** model trained with the **VariKN** toolkit.

RNN acoustic model

+11.4%

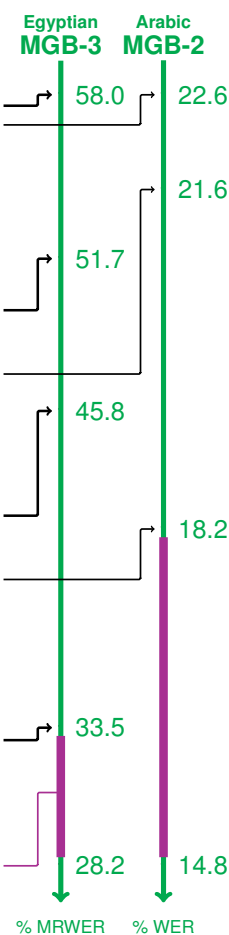
- We experimented with different recurrent and non-recurrent architectures.
- Bidirectional long short-term models combined with regular TDNN layers (**TDNN-BLSTM**) gave the best results. This model had 3 TDNN layers (1024 units) and 3 BLSTM layers in both directions (256 units).
- We used WER on the MGB-2 development set as the stopping criterion.

Dialect adapted Acoustic model

+26.7%

- We used our best TDNN-BLSTM model and **adapted** with a simple procedure.
- The Egyptian data was used to continue training the model for a small number of epochs. Same parameters were used as for the last regular training iteration.
- We used MRWER on the MGB-3 development set as the stopping criterion.

Scale of language modeling improvements in the next column.



Language / lexical modeling

Interpolation and Rescoring

+2.8%

- Besides a small language model on the background text data, we also trained bigger models on the background data, as well as separate models on the transcriptions of the MGB-2 and MGB-3 training sets.
- We **interpolated** these n -gram models, with the interpolation weight optimized for the respective development sets.
- We **rescored** the first-pass results with the larger interpolated language model.

Subwords

-0.1%

- Instead of words as units, we also tried **subwords** (trained with **Morfessor**) and **characters** as language modeling units.
- Subwords reduce the vocabulary size while increasing the coverage of the language, as unseen words can be constructed.
- For MGB-3 this did not give an immediate improvement, but in combination with later steps, it did improve over word models.

Recurrent Neural Network Language Model

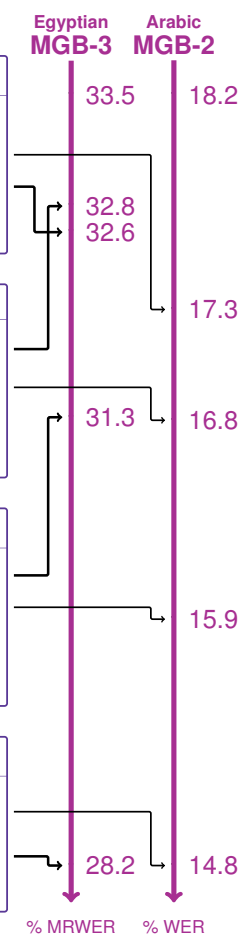
+4.5%

- We build Recurrent Neural Network Language Models (**RNNLM**) with the **TheanoLM** toolkit.
- Our models were **LSTM** models with a projection layer (300/200 units), a hidden layer (1500/1000 units) and a highway layer (1500/1000 units) with a \tanh activation function (units for subwords/words). The models were trained with Adagrad.
- We rescored the lattices and interpolated them with the n -gram lattices.

System combination

+9.9%

- Different systems were built, including TDNN-LSTM and TDNN-BLSTM acoustic models, and models with various lexical units and boundary marking methods.
- Systems were combined and decoded using Minimum Bayesian Risk (**MBR-decoding**).
- In total 40 systems were used.

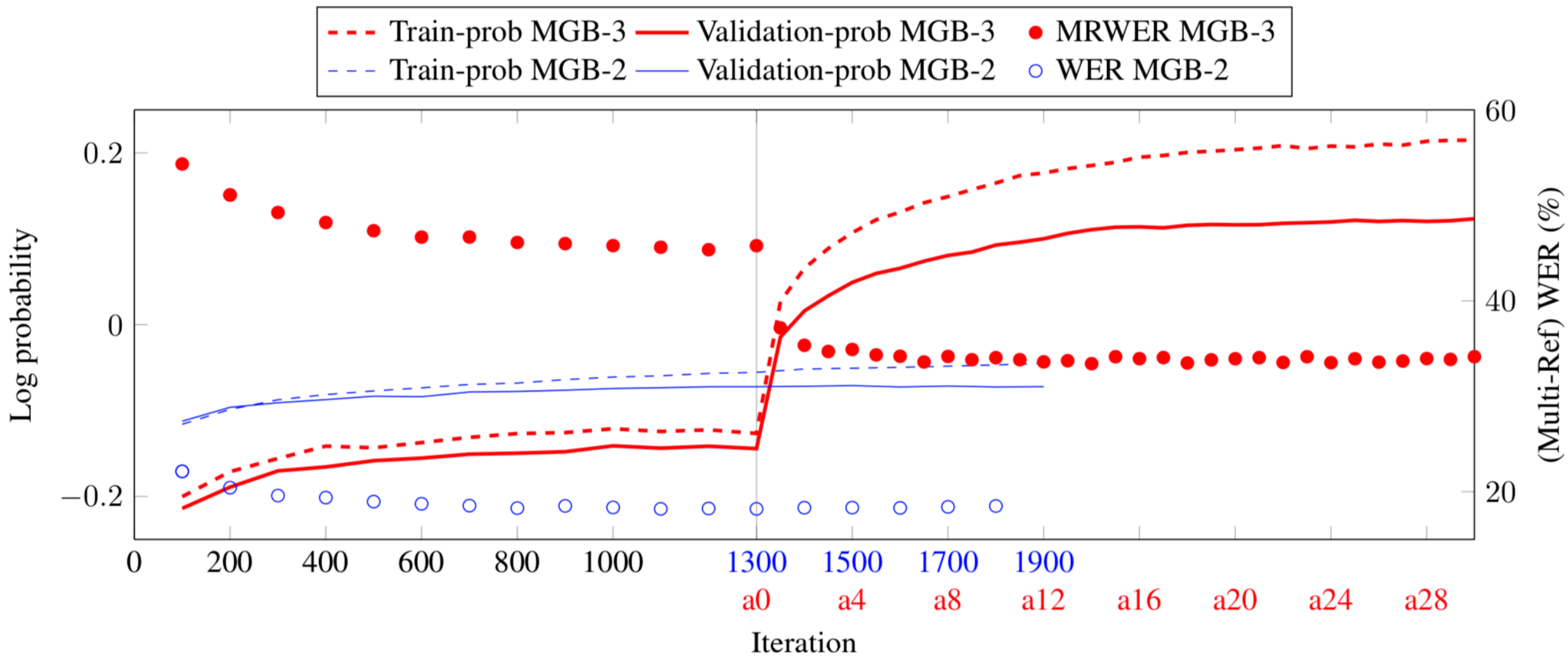


Steps for building our system

- Train a grapheme based GMM
- Data cleaning
- Build BLSTM acoustic model
- Adapt model (MGB-3 only)

Model dialect adaptation

- Train DNN with only background data
- After convergence, continue training complete model with dialect data only, keeping all hyperparameters (learning rate etc.) static.
- Train with all different transcriptions for dialect data
- Stop training once dev-set error rates start climbing.



Steps for building our system

- Train a grapheme based GMM
- Data cleaning
- Build BLSTM acoustic model
- Adapt model (MGB-3 only)
- Create subword lexicon
- Variable-order n-gram language model
- RNNLM rescoring
- System combination

Results after each step

Stage	Egyptian (MGB-3)		Arabic (MGB-2)	
	MRWER %	Rel impr. %	WER %	Rel impr. %
Baseline Organizers	58.0	-	22.6	
Baseline Aalto (TDNN, word n-gram)	51.7	10.9	21.6	4.2
BLSTM Acoustic model	45.8	11.4	18.2	15.9
Adapted BLSTM (Egyptian)	33.5	26.7		
N-gram interpolation	32.6	2.8	17.3	5.2
Subwords	32.8	-0.6	16.8	2.8
Neural network language model rescoring	31.3	4.5	15.9	5.3
System combination (36 systems)	28.2	9.9	14.8	6.8

Challenge results (from Ali et al. 2017)

	MGB2 WER	MGB3 WER per transcriber				MGB3	
		WER1	WER2	WER3	WER4	AV-WER	MR-WER
<i>2016-best system</i>	14.7						
Aalto	13.2	38.0	37.7	37.4	36.9	37.5	29.3
NDSC-THUEE	14.5	41.5	40.1	40.7	40.8	40.75	32.5
JHU	16.0	42.1	42.4	41.4	41.1	40.7	32.8
MIT	17.5	45.4	45.4	45.5	44.2	44.9	36.8
BUT	24.7	55.0	55.2	54.3	54.4	53.4	46.8
RDI-CU	16.0	63.2	63.4	62.6	62.7	62.5	57.7

Table 7: Summary of speech-to-text transcription results for MGB-2 and MGB-3 data. For MGB-3, WERs are given for each of the four references (produced by different transcribers), as well as AV-WER and MR-WER across the four references.

Questions about our MGB-
submission?

Subword based ASR

What are subwords?

- WORDS: Different types of complexions
- Morphological?: Different | types | of | complex ions
- Statistical?: Different | types | of | complex ions
- Duochars: Different | types | of | complex ions
- Character: Different | types | of | complex ions

Why subwords?

Words

- Very large vocabularies (1M? 4M+?)
- Never full coverage
- Impossible to recognize words outside vocabulary
- Many words are sparse in LM training data

Subwords

- Pick the vocab size (10k, 100k)
- Full coverage of language
- Able to recognize words that do not exist (yet)
- Most subwords appear many times in language data

Why no subwords?

- Need to create subword segmentation
 - Morphological: Need analyzer/expert. Often ambiguous (especially for surface forms)
 - Statistical: Training algorithm (Morfessor, BPE)
 - Characters: Easy
- Need to use grapheme lexicon
 - Very possible in sequence-trained acoustic models
- Need language model that takes longer modeling context
 - Variable order n-grams
 - Neural network lm

Grapheme vs Phoneme lexicon

Language	Grapheme	Phoneme	
Finnish	10.0	10.0	Phonemes == Graphemes
English	21.4	20.5	
Arabic	23.7	20.4	

Variable order n-grams

- Normal n-gram estimation: Create full n-order model and prune to desired size
- Variable-order:
 - Grow model selectively, only add n-grams of next order if they add something to the model
 - Still prune after growing step
- Vesa Siivola, Teemu Hirsimäki and Sami Virpioja, "On Growing and Pruning Kneser-Ney Smoothed N-Gram Models", 2007

Example counts for VariKN-model

ngram 1=99229
ngram 2=18132565
ngram 3=12736038
ngram 4=7646470
ngram 5=3425873
ngram 6=1749286
ngram 7=999495
ngram 8=478707
ngram 9=209144
ngram 10=85423

ngram 11=33067
ngram 12=13881
ngram 13=2741
ngram 14=2280
ngram 15=99
ngram 16=89
ngram 17=74
ngram 18=71
ngram 19=4
ngram 20=9

NNLM models with subwords

- Huge vocabularies give huge input layers, and many (untrainable) parameters.
- Normally things like Hierarchical Softmax is needed for output layers
- Subwords make network simple, and all parameters have enough training samples for robust training
- Even character-based models are possible

Subword results (from Smit et al. 2019)

Unit	TDNN	TDNN+LSTM	TDNN+BLSTM	AM comb
char	18.9	17.8	16.5	15.9 (3)
morf 0.001	17.7	16.7	15.7	14.9 (3)
morf 0.01	17.7	16.6	15.6	14.9 (3)
morf 0.1	18.4	17.1	16.2	15.4 (3)
word	18.7	17.6	16.5	15.7 (3)
				15.5 (5)
LM/AM				
Comb	14.7 (10)			
				14.6 (15)

(b) Arabic

Questions?

Arabic ASR demonstration

Modern Standard Arabic example

Egyptian Arabic example

Live demo?

The Aalto submission to the MGB-3 challenge



Peter Smit

Siva Reddy Gangireddy

Sami Virpioja

Mikko Kurimo

<https://petersmit.eu/post/arabicspeech/>