Improved low-resource Somali speech recognition by semi-supervised acoustic and language model training

Astik Biswas, Raghav Menon, Ewald van der Westhuizen,Thomas Niesler
Stellenbosch University, South Africa

Summary

We present improvements in automatic speech recognition (ASR) for Somali, a currently extremely under-resourced language. This forms part of a continuing United Nations (UN) effort to employ ASR-based keyword spotting systems to support humanitarian relief programmes in rural Africa. Using just 1.57 hours of annotated speech data as a seed corpus, we increase the pool of training data by applying semi-supervised training to 17.55 hours of untranscribed speech. Three semi-supervised training passes were performed, where the decoded output from each pass was used for acoustic model training in the subsequent pass. The automatic transcriptions (AutoT) from the best performing pass were used for language model augmentation. To ensure the quality of automatic transcriptions, decoder confidence is used as a threshold.

Background

- The success of Ugandan radio browsing system deployed by the United Nation for humanitarian relief application inspired to do the same for Somali.
- Somali is an Afroasiatic language. It is the official language of Somalia and widely used its neighbouring countries.
- Somali is an agglutinative language, the number of unique word tokens is large
- The preprocessed audio stream is passed to the ASR system which generates lattices which are subsequently searched for predefined keywords.
- Human analysts further process the data which aid in humanitarian decision making and situational awareness.
- Given the amount of Somali transcriber speech, multilingual acoustic model training found promising.

Acoustic Data

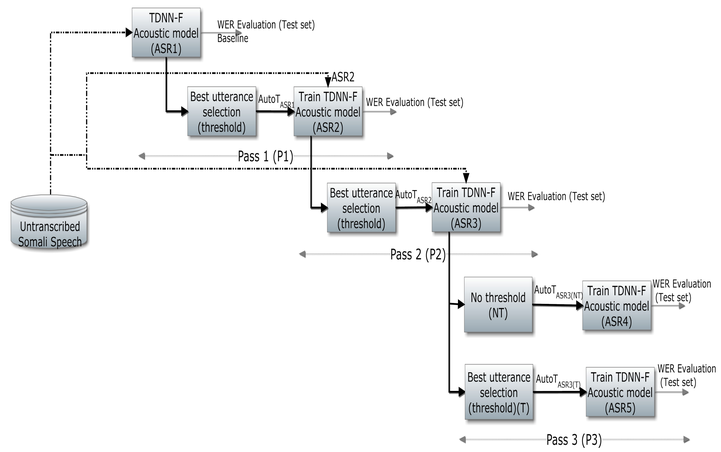
- This small dataset of speech captured from broadcast Somali radio phone-in programmes and transcribed manually (ManT), contains only:
  - 1.57 hours training speech that is available for training and
  - 10 minutes for testing.
- Other transcribed language resources used were:
  - Luganda (9.6h)
  - Acholi (9.2h)
  - Ugandan English (6.0h)
  - South African English (20.0h)
- For semi-supervised training, approximately 17.55 hours of untranscribed Somali speech, collected from phone-in programme.

Text Data for Language Modeling

Somali text resources used for language modelling

Corpus	Word tokens	Word types	Sentences
Somali transcriptions	15.1k	4.7k	1.3k
Somali news text	1.92M	82.8k	59.2k
Facebook posts	1.55M	92.9k	54.9k
Facebook comments	3.5M	356.7k	215.3k
LCC newspaper text	2.37M	300k	100k
LCC Wikipedia text	200k	50.7k	10k
LSTM generated text	11.29M	4.7k	775.3k

Somali Semi-supervised Training



Semi-supervised training framework for Somali ASR. -·-·-·→ represents untranscribed speech is being fed to transcriber

- Three iterations of semi- supervised training were implemented.
- Average decoder Confidence threshold was applied on each iteration to select best transcriptions.
- ASR2 and ASR3 were retrained with 9.11h and 9.58h of AutoT speech respectively.
- Final pass (ASR3), evaluated with two different configurations: 17.55h (threshold=0) and 9.86h AutoT data respectively.
- All acoustic models were factorised time-delay neural network (10 time-delay layers followed by a rank reduction layer).

Perplexity Evaluation

LM	Sources	Optimized on	AutoT	PPval	PPtst
LMbase	T1, T2-T4, T7	Test set	no	–	269.80
LM2	T1, T2-T7	Test set	no	–	253.60
LM3	T1, T2-T7	Validation set	no	576.98	321.31
LM4	T1, T2-T7	Test set	yes(ASR2)	–	260.94
LM5	T1, T2-T7	Test set	yes (ASR2)	500.49	300.25

Results

Word error rate(%) on Somali test set of different types of Somali ASR.

System	Type	Training data size (h)			PPval
		ManT	AutoT		
ASR1	Supervised	46.37		0.00	53.68
ASR2	Semi-supervised	46.37	AutoT$_{ASR1}$	9.11	51.91
ASR3		46.37	AutoT$_{ASR2}$	9.58	**50.95**
ASR4		46.37	AutoT$_{ASR3}$	17.55	51.71
ASR5		46.37	AutoT$_{ASR3}$	9.86	51.09

- DNN results reveal that additional training data from same language family (N-N) results in a relative improvement of 2.71% compared to bilingual baseline.
- Gain observed with additional training data from a different language family (N-S) is lower at 2.09%.
- An improvement of 4.85% relative to baseline DNN system observed when training on all 4 CS language pairs.
- TDNN-LSTMs perform better than DNNs but similar show trend.
- TDNN-LSTM system for 4 CS pairs shows a 7.88% relative improvement in WER compared to TDNN-LSTM baseline.

Conclusions

- These experiments represent first multilingual acoustic models trained on multiple code-switched datasets, and the first such investigation for African languages.
- System performance benefits most when additional training data originates from a closely related language, in this case another Nguni language rather than a Sotho language.
- Best overall performance was achieved when data from all four code-switched language pairs were combined.
- Best system shows a WER improvement of 7.88% relative to the baseline.