DEEP NEURAL NETWORKS BASED AUTOMATIC SPEECH RECOGNITION FOR FOUR ETHIOPIAN LANGUAGES

Solomon Teferra Abate[12], Martha Yifiru Tachbelie[12] and Tanja Schultz[1]

@

[1]CSL, University of Bremen, Germany
[2]SIS, Addis Ababa University, Ethiopia
abate, marthayifiru, tanja.schultz@uni-bremen.de

The authors would like to thank the Alexander von Humboldt Foundation for research fellowship

Outline

- Introduction
  - Challenges
  - Background
  - The Languages
- Experimental Setups
  - Corpora
  - Tools and Techniques
- Experimental Results
  - Amharic and Tigrigna
  - Oromo and Wolaytta
- Discussions
- Conclusions and Future Directions

Introduction: Challenges

- Ethiopia has more than 80 languages and a population of about 110 Million
  - Its illiteracy rate is about 42%
    - Speech technologies are of high demand for all of its languages
    - But they have not been developed for almost all of the languages
- Challenges that hinder the development of speech technologies:
  - Lack of speech and language resources,
  - Lack of computational resources
- Our opportunities to tackle these challenges:
  - Development of read speech corpora for Amharic, Tigrigna, Oromo and Wolaytta
  - Computational resources at CSL of the University of Bremen

Introduction: Background

- We have developed DNN based ASR systems for four Ethiopian languages
  - The languages are from three language families:
    - Semitic language family - Amharic, Tigrigna
    - Cushitic language family – Oromo
    - Omotic language family – Wolaytta
  - Used very large decoding vocabularies for the Semitic languages
    - ➢ To minimize the effect of high OOV rates

Introduction: The Languages

- Phonology:
  - Amharic and Tigrigna have 28 and 31 consonants, respectively and 7 vowels
  - Oromo and Wolaytta have 28 and 26 consonants, respectively and 5 vowels
- Morphology:
  - All these languages are morphologically complex
    - They have inflectional and derivational morphology
  - The morphology of Amharic and Tigrigna is more complexity than that of Oromo and Wolaytta
  - Their OOV rates on a comparable test text show this nature of the languages

Languages	Training Vocabulary	OOV	OOV With 21,232
Amharic	28,661	24.99	33.37
Tigrigna	31,759	16.33	19.75
Oromo	21,232	11.73	11.73
Wolaytta	25,267	9.34	10.09

Experimental Setups: Corpora

- The speech corpora we used:

Languages	Training [hrs:min]	Development [hrs:min]	Evaluation [hrs:min]
Amharic	20:00	1:30	1:33
Oromo	22:48	1:11	1:04
Tigrigna	22:06	1:03	1:02
Wolaytta	29:42	1.32	1.43

- Our language model training texts consist of:
  - 4 million tokens for Amharic and Tigrigna, each,
  - 1.5 million tokens for Oromo and
  - 226k tokens for Wolaytta
- The lexical models of all the languages have been generated using automatic grapheme to phoneme (G2P) conversion
  - The writing systems of the languages reflect their phonetic properties

Experimental Setups: Tools and Techniques

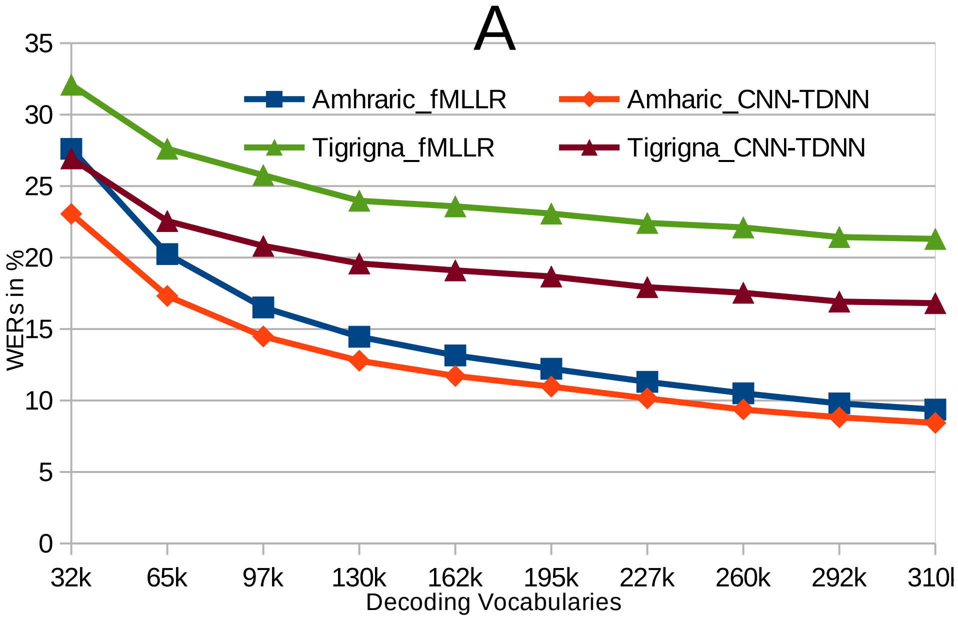
Cognitive Systems Lab

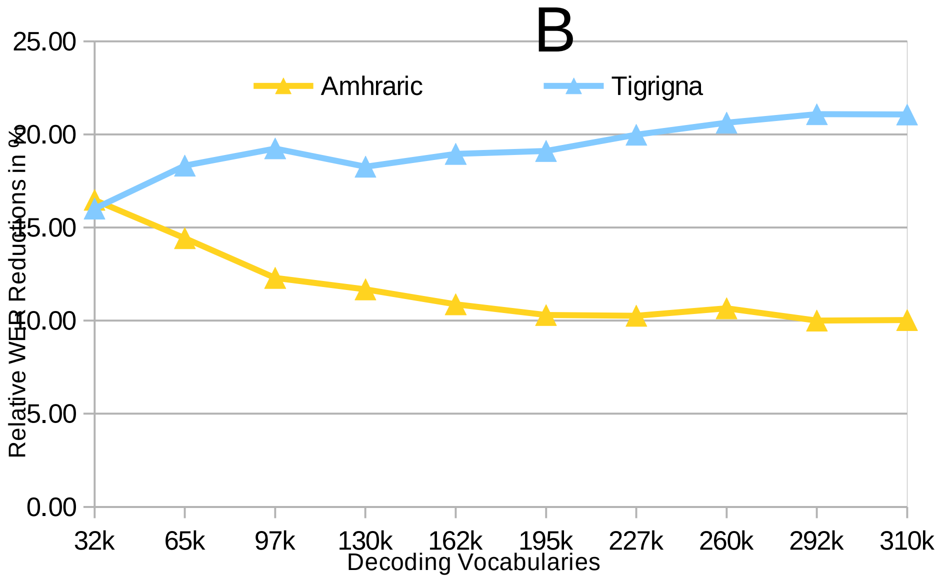
- Trigram LMs are developed using SRILM toolkit
  - Smoothed with unmodified Kneser-Ney smoothing
- HMM-GMM and HMM-DNN Acoustic Models are developed using Kaldi toolkit
  - We trained different HMM-GMM Acoustic Models
    - Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) feature transformation for each of the models
    - Speaker Adaptive Training (SAT) using an affine transform, feature space Maximum Likelihood Linear Regression (fMLLR)
  - HMM-DNN acoustic models are developed using Factored Time Delay Neural Networks with additional Convolutional layers (CNN-TDNNf)
    - Three-fold data augmentation has been used
    - 40-dimensional MFCCs without derivatives, with 3-dimensional pitch features and 100-dimensional i-vectors
    - 15 hidden layers (6 CNN and 9 TDNNf) of 1024 units

Experimental Results: Amharic and Tigrigna

Cognitive Systems Lab

- Amharic and Tigrigna have more complex morphology
  - We used different sizes of decoding lexicons ranging from 32.5k to 310k
    - OOV rates of 3.06% for Amharic and 4.89% for Tigrigna
  - WERs of 8.43% and 16.82% for Amharic and Tigrigna, respectively with DNN AMs (Figure A)
  - Tigrigna benefited more from DNN than Amharic (Figure B)

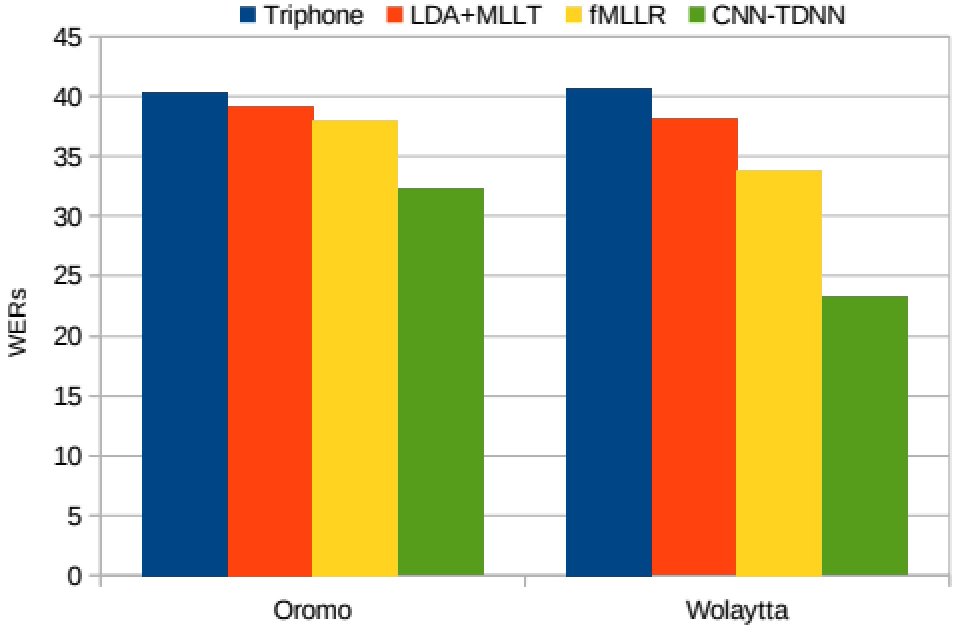




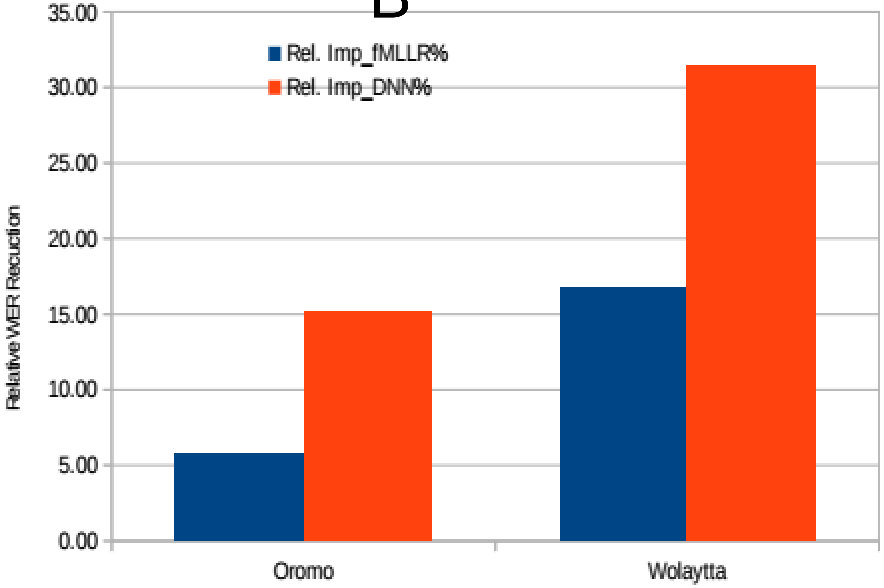
Experimental Results: Oromo and Wolaytta

- For Oromo and Wolaytta, we used the training lexicon for decoding
  - We have achieved WERs of 32.28% and 23.23% for Oromo and Wolaytta, respectively with DNN AMs (Figure A)
  - The relative WER reduction for Wolaytta is higher than for Oromo (Figure B)

A



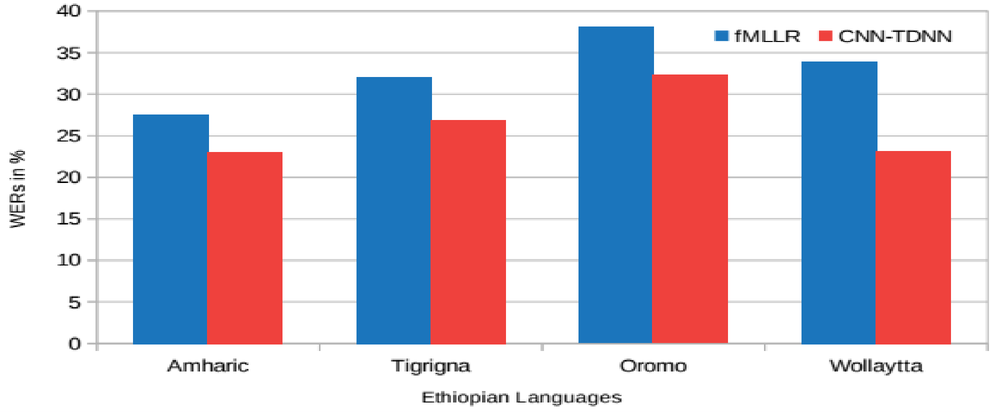
B



Discussions

Cognitive Systems Lab

- From our experiments and the results we have achieved
  - Application of CNN-TDNNf reduces WER in the development of an ASR for the four Ethiopian languages
    - As a result of using large training speech, Wolaytta benefited most from DNN
      - WER reduction from 33.89% (achieved using GMM) to 23.23% ➔ relative reduction of 31.45%
  - Our results showed the fact that strength of LM bring a significant impact on WER reduction
    - As reflected in the lower WER of Tigrigna than WER of Oromo and Wolaytta
      - That calls up on preparation of text for language model training



Conclusions and Future Directions

- From this work we conclude that HMM-DNN ASRs outperform the HMM-GMM based ones for all the languages
  - irrespective of the size of training speech,
  - decoding vocabulary and
  - strength of the language models
- Wolaytta benefited most from the use of HMM-DNN
  - Which might be due to the large amount of training speech in Wolaytta
- Based on our results, we recommend development of strong language models for Oromo and Wolaytta
- Following state-of-the-art approaches and the phonetic relationship among these languages, we are working on the development of multilingual ASR