ARL-IR-0000 • Apr 2018

US Army Research Laboratory

A Minimal Example of MTL for ASR

by John Morgan, Stephen LaRocca and Michelle Vanni

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-IR-0000 ● APR 2018

US Army Research Laboratory

A Minimal Example of MTL for ASR

by John J Morgan
Computational and Information Sciences Directorate, ARL

Stephen A LaRocca and Michelle Vanni
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
April 2018	Internal Report	October 2016-November 2016

4. TITLE AND SUBTITLE
A Minimal Example of MTL for ASR

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
John Morgan, Stephen LaRocca and Michelle Vanni

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
US Army Research Laboratory
ATTN: RDRL-CII-T
Adelphi Laboratory Center, MD 20783-1138

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-IR-0000

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
primary author's email: <john.j.morgan50.civ@mail.mil>.

14. ABSTRACT
Multitask Learning was applied to a Large corpus of English and a small corpus of Modern Standard Arabic read speech for the purpose of improving the performance of an Automatic Speech Recognition system. An improvement in Word Error Rate over the best Singletask Learning method was observed.

15. SUBJECT TERMS
Automatic Speech Recognition of Accented Speech, Multitask Learning

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			John J Morgan
Unclassified	Unclassified	Unclassified	UU	16	19b. TELEPHONE NUMBER (Include area code) 301-394-1902

Contents

List of Tables

Acknowledgments

John Morgan wishes to sincerely thank his co-author, Dr. Stephen LaRocca.

INTENTIONALLY LEFT BLANK.

1. INTRODUCTION

For the experiments described here, we used the Neural Network (nn) as a framework for acoustic modeling in Automatic Speech Recognition (asr). Recently, large asr systems have been trained on tens of thousands of hours of speech data.[1] nn models have performed well when these amounts of data are available. We assumed we are under conditions of severe training data paucity in the Target Language (tl). We also assumed that we have access to a large corpus of speech in another language. We refer to this other language as the Background Language (bl).

Multitask Learning (mtl)[2] is a framework that enables the advantages of Deep Learning (dl) to be applied in the situation where we have access to large amounts of data in the bl and scarce resources in the tl.

an nn was built with two kinds of layers.

1. Shared Layers.
2. Language Specific Layers.

The shared layers were trained on all the training data from all the languages.

The language specific layers were trained only on data from the tl.

Our research question was formulated as follows.

Can nn acoustic models trained with mtl on data from two different languages improve performance of an asr system for a Low Resource (lr) tl?

2. DATA

We ran an experiment to test the mtl method on two specific publically available corpora: the large Librispeech English corpus and a small corpus of Tunisian Accented msa.

2.1 Librispeech

LibriSpeech is a corpus of read speech, based on LibriVox's public domain audio books. The corpus is available at:

http://www.openslr.org/resources/12.

We used the cleaned training fold of 960 hours of speech.

2.2 Tunisian msa:

This is a corpus of ten hours of msa as spoken by 120 male and female Tunisians in 2003. The informants provided recitations and answers to questions. It can be downloaded at:

http://www.openslr.org/resources/46.

3. EXPERIMENT

We used the kaldi toolkit[3] to build our ASR systems. We derived our setup from the kaldi babel multilang recipe. We wrote our recipe so that it only contain steps that are required to implement the mtl method. Thus, We did not use i-vectors which are standard in many kaldi recipes. We did include however a bottleneck layer.

We built baseline Speaker Adaptive Training (sat) Gaussian Mixture Model (gmm) Hidden Markov Model (hmm) Acoustic Model (am)s for both the Librispeech and Tunisian msa corpora. For the Tunisian msa baseline system we derived our pronouncing dictionary from the 2 million entries from the Qatar Computing Research Institute (qcri) vowelized dictionary[4] available at: http://alt.qcri.org/resources/speech/dictionary/ar-ar_lexicon_2014-03-17.txt.bz2 We added the Out Of Vocabulary (oov) words from the Tunisian msa training set and the test set. We trained our 3-gram language model with the Stanfor Research Institute Language Model (srilm) toolkit[5] on the transcripts from the training and test data. Our best Word Error Rate (wer) results for Tunisian msa were obtained with online chain models.

We followed the kaldi standard recipe for the librispeech cleaned 960 hours of speech task with one exception. We extracted and trained with Perceptual Linear Prediction $(\text{plp})_{\text{pitch}}$ features instead of Mel Frequency Cepstral Coefficients (mfcc) features. We followed this path since we derived our scripts from the kaldi babel multilang recipe.

3.1 Neural Network Configuration

The nn for both languages had ten layers.

1. One input layer,
2. 6 hidden layers,
3. One Bottleneck layer,
4. One affine layer, and
5. One soft max layer.

The Rectified Linear Unit (relu) function was used to compute activations. The dimension of the hidden layers was 1024. The dimension of the Bottleneck layer was 512.

The final layer implemented a soft max function that output a probability density function over the clustered triphones.

The frame Context was set to: 16 frames to the left and 12 frames to the right.

3.2 Training

A bilingual raw deep nn was trained on the combined set of training examples from the English Librispeech and Tunisian msa corpora. The data from the Tunisian msa corpus was used to readjust the parameters in the last two layers of the bilingual Deep Neural Network (dnn) model to produce a new monolingual Tunisian msa acoustic model. Similarly, a new monolingual English model was produced. These two models shared the parameters in their first eight layers, only their final 2 layers were different.

3.3 Decoding

The monolingual system with the Tunisian msa am was used to decode a test set of speech from four speakers, 3 Libyan males and one Tunisian female. The same Finite State Transducer (fst) decoding graph that was built for the Tunisian msa sat gmm hmm system was used for decoding with the mtl am set.

4. RESULTS

TThe Single-Task unisian msa Baseline system with chain model trained ams yielded a wer of 11.03.

After MTL, the Tunisian msa system gave a wer of 7.12.

5. References

1. Heigold G, Vanhoucke V, Senior A, Nguyen P, Ranzato M, Devin M, Dean J. Multilingual acoustic models using distributed deep neural networks,âĂİ ICASSP. 2013.

2. Caruana R. Multitask learning: A knowledge-based source of inductive bias. In: Proceedings of the Tenth International Conference on Machine Learning; 1993 ; Venue Unknown. Morgan Kaufmann; 1993. p. 41–48.

3. Povey D, Ghoshal A, Boulianne G, Goel N, Hannemann M, Qian Y, Schwarz P, Stemmer G. The kaldi speech recognition toolkit. In: In IEEE 2011 workshop; 2011 ; Venue Unknown.

4. Ali A, Zhang Y, Cardinal P, Dahak N, Vogel S, Glass J. A complete kaldi recipe for building arabic speech recognition systems. In: Spoken Language Technology Workshop (SLT), 2014 IEEE; 2014 Dec; Venue Unknown. p. 525–529.

5. Stolcke A. Srilm - an extensible language modeling toolkit. In: SRILM - An Extensible Language Modeling Toolkit; 2002 ; Interspeech. p. 901–904.

INTENTIONALLY LEFT BLANK.

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
2 (PDF)	DIRECTOR US ARMY RESEARCH LAB RDRL CIO L IMAL HRA MAIL & RECORDS MGMT
1 (PDF)	GOVT PRINTG OFC A MALHOTRA

INTENTIONALLY LEFT BLANK.