Merging of native and non-native speech for low-resource accented ASR

Sarah Samson Juan[1], Laurent Besacier[2], Benjamin Lecouteux[2] and Tien-Ping Tan[3]

[1]Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak, Sarawak, Malaysia
[2]Grenoble Informatics Laboratory (LIG), Univ. Grenoble-Alpes, Grenoble, France
[3]School of Computer Science, Universiti Sains Malaysia, Penang, Malaysia

SLSP 2015, Budapest

Outline

Introduction

- Performance of non-native automatic speech recognition (ASR) is poor when few (or no) non-native speech is available for training / adaptation.
- Many approaches have been suggested for handling accented-speech in ASR:
  - acoustic model merging ((Morgan, 2004), (Bouselmi, Fohr, and Haton, 2005), (Tan and Besacier, 2007), (Tan, Besacier, and Lecouteux, 2014)),
  - applying maximum likelihood linear regression (MLLR) for adapting models to each non-native speaker (Huang et al., 2000), or
  - adapting lexicon ((Arslan and Hansen, 1996), (Goronzy, 2002))

Introduction contd.

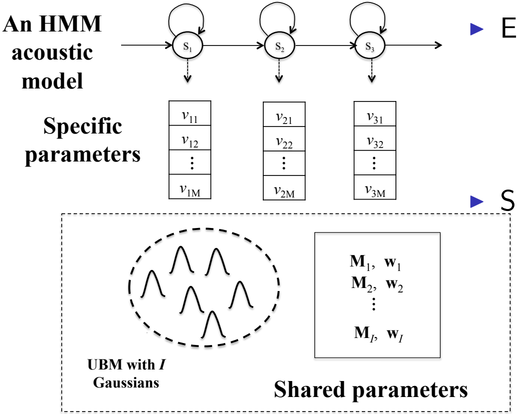
- Multi-accent approach for accented speech:
  - Subspace Gaussian Mixture Model (Mohan, Ghalehjegh, and Rose, 2012) and Deep Neural Network (Huang et al., 2014) - apply pooling data approach
- **Can we finely merge unbalanced corpora (large native data $<->$ small non-native data) for achieving an optimal acoustic model?**

Outline

Subspace Gaussian Mixture Model

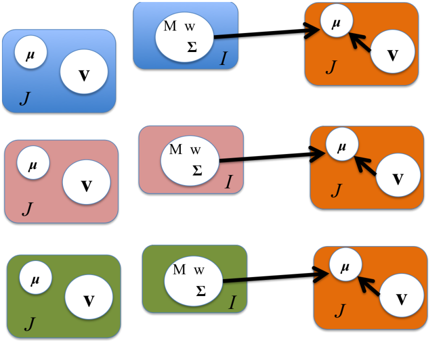
General

Subspace Gaussian Mixture Model (SGMM) (Povey et al., 2010):



- Each HMM state:
  - Defined by a low-dimensional vector v_{jm}
  - Mixture of substates
- Shared parameters:
  - Universal Background Model (UBM)
  - Means, $\mathbf{M}_i$
  - Mixture weights, $\mathbf{w}_i$

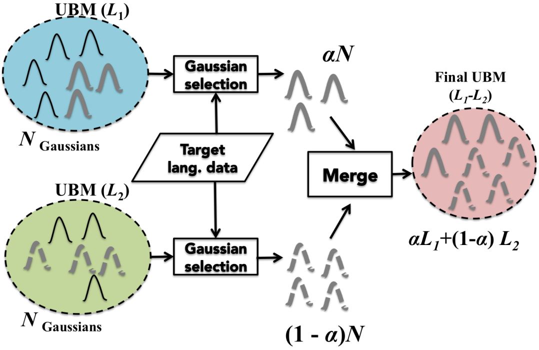
Multi-accent SGMM



Using SGMM:

- Transfer shared parameters from source to target system
- Applied by (Imseng et al., 2014) and (Lu, Ghoshal, and Renals, 2014) - cross-lingual acoustic model for low-resource systems

Language-weighting strategy for multi-accent SGMM

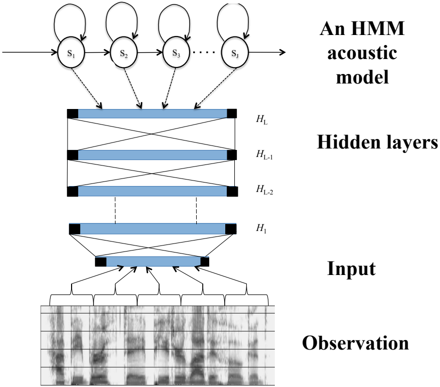


- UBM Gaussians, e.g. $N = 500$
- L_1 = Non-native
  L_2 = Native
- Weights, α= 0.1, 0.2, ..., 0.9.

Deep Neural Networks

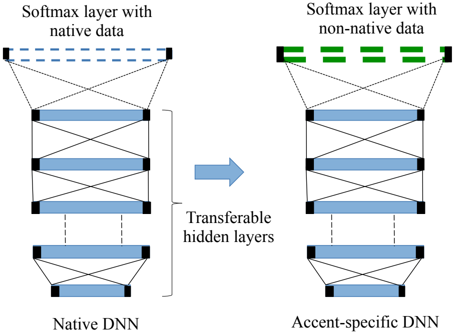
General

Deep Neural Networks (DNN) (Hinton et al., 2012):



- Alternative to HMM/GMM systems
- Feedforward neural network
- Intialization of DNN weights:
  - Random
  - Pretraining - Restrictive Boltzmann Machines (RBM) (Hinton, 2010)
- Adjust weights - Stochastic Gradient Descent

Accent-specific top layer DNN



1. Train DNN on Native / Non-native data:
2. Remove last layer (softmax layer) from DNN with native speech
3. Fine-tune hidden layers on non-native training data

Outline

Experimental setup

- Non-native - Malaysian English (Tan, Besacier, and Lecouteux, 2014):
  - Train: 2h transcribed; 9h untranscribed (UBM - 11h)
  - Test: 4h
- "Native" - TED English[1] (TED-LIUM) (Rousseau, Deléglise, and Estève, 2012)
  - Train: 118h
  - Test: 4h
- Toolkit: Kaldi
- Systems:
  - HMM/GMM
  - HMM/SGMM :
    - UBM 500
    - Merging: $\alpha = 0.1, ..., 0.9$
    - substates 800 to 8750
  - HMM/DNN : 6 hidden layers with 1024 units

[1] Even if non-native speakers exist in the corpus

Outline

Introduction

Acoustic Model Merging Approach

Experimental Setup

Performance of Non-native ASR

Conclusions

Baseline results (WER %)

English ASR results for native and non-native speech
- no speaker adaptation (fMLLR) at this stage

Train	Test	
	Native (4h)	Non-native (4h)
Native 118h	30.55 (GMM) 28.05 (SGMM) **19.10 (DNN)**	
Non-native 2h		

Baseline results (WER %)

English ASR results for native and non-native speech
- no speaker adaptation (fMLLR) at this stage

Train	Test	
	Native (4h)	Non-native (4h)
Native 118h	30.55 (GMM) 28.05 (SGMM) **19.10 (DNN)**	57.09 (GMM) 45.84 (SGMM) **40.70 (DNN)**
Non-native 2h		

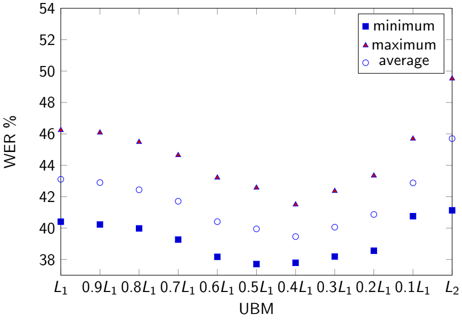
Baseline results (WER %)

English ASR results for native and non-native speech
- no speaker adaptation (fMLLR) at this stage

Train	Test	
	Native (4h)	Non-native (4h)
Native 118h	30.55 (GMM) 28.05 (SGMM) **19.10 (DNN)**	57.09 (GMM) 45.84 (SGMM) **40.70 (DNN)**
Non-native 2h		41.47 (GMM) 40.41 (SGMM) **32.52 (DNN)**

Multi-accent SGMM results

L_1: Malaysian English, L_2: TED English



- 4h test data
- **Best WER: 37.7%** - Baseline: 40.4%
- $\alpha = 0.5$ (250 Gaussians from L_1/L_2)
- Increase substates degrades results

Accent-specific top layer for DNN

DNN with accent-specific top layer	WER (%)
Baseline - standard DNN	32.52
No speaker adaptation	24.89
Speaker adaptation	21.48

Outline

Conclusions

- Proposed two approaches for optimal merging of native and non-native data in order to improve accented ASR with limited training data:
  1. Language weighting strategy for multi-accent compact SGMM acoustic models - used language weights to control the number of UBM Gaussians.
  2. Fine-tuning hidden layers of native DNN on the non-native training data
- Observed improvements on non-native ASR performance:
  - Relative improvement: 15% for SGMM (multi-accent UBM500 - $\alpha = 0.5$) and 34% for DNN (accent-specific with speaker adaptation).


thank you
danke
謝謝
ngiyabonga
teşekkür ederim
gracias
merci
obrigado
спасибо
рахмат
Баярлалаа
dziękuję
bedankt
go raibh maith agat
mochchakkeram
sukriya
kop khun krap
terima kasih
감사합니다
ありがとう
arigatô
dakujem
grazie
tapadh leat
мерси
хвала
εὐχαριστώ
dank je
hvala
mauruuru
sagolun
takk
taiku
mersi
barka
welalin
tack
spas
kia ora
vinaka
спасибі
blagodaram
misaotra
matondo
paldies
grazzi
mahalo
faafetai lava
kiitos
dankie
dhanyavad
nanni
nandri
köszönöm
enkosi
bayarlalaa
gràcie
sobodi
dêkuji
akun
dankon
ačiū
djiere dieuf
tau
дякую
mamnun
asante
manana
obrigada
murakoze
tenki
chokrane
chnorakaloutioun
gratias ago
gràcies
sulpáy
trugarez
mèsi
didi madloba
najis tuke
kam sah hamnida
rahmat
tanemirt
rahmet
diolch
dhanyavadagalu
shukriya
mercé
xièxie
তোমাকে ধন্যবাদ

References I

Arslan, M. J. and J. L. Hansen (1996). "A Study of the Temporal Features and Frequency Characteristics in American English Foreign Accent". In: *Journal of the Acoustic Society.*

Bouselmi, G., D. Fohr, and J. P. Haton (2005). "Fully Automated Non-native Speech Recognition using Confusion-based Acoustic Model Intergration". In: *Proceedings of Eurospeech.* Lisboa, pp. 1369–1372.

Goronzy, S. (2002). "Robust Adaptation to Non-native Accents in Automatic Speech Recognition". In: *Springer.*

Hinton, Geoffrey et al. (2012). "Deep Neural Networks for Acoustic Modeling in Speech recognition". In: *IEEE Signal Processing Magazine* 29.6, pp. 82–97.

Hinton, Geoffrey E. (2010). *A Practical Guide to Training Restricted Boltzmann Machines*. UTML TR 2010-003. Dept. Computer Science, University of Toronto.

References II

Huang, C. et al. (2000). "Accent Modeling based on Pronunciation Dictionary Adaptation for Large Vocabulary Mandarin Speech Recognition". In: *Proceedings of ICLSP*. Vol. 2, pp. 818–821.

Huang, Yan et al. (2014). "Multi-accent Deep Neural Network Acoustic Model with Accent-specific Top Layer using the KLD-regularized Model Adaptation". In: *Proceedings of INTERSPEECH*.

Imseng, David et al. (2014). "Using Out-of-language Data to Improve Under-resourced Speech Recognizer". In: *Speech Communication* 56.0, pp. 142–151.

Lu, Liang, Arnab Ghoshal, and Steve Renals (2014). "Cross-lingual Subspace Gaussian Mixture Models for Low-resource Speech Recognition". In: *IEEE/ACM Transactions on Audio, Speech and Language Processing*. Vol. 22, pp. 17–27.

References III

Mohan, Aachan, Sina Hamidi Ghalehjegh, and Richard C. Rose (2012). "Dealing with Acoustic Mismatch for Training Multlingual Subspace Gaussian Mixture Models for Speech Recognition". In: *Proceedings of ICASSP*. Kyoto: IEEE, pp. 4893–4896.

Morgan, J. J. (2004). "Making a Speech Recognizer Tolerate Non-native Speech through Gaussian Mixture Merging". In: *Proceedings of ICALL'04*. Venice.

Povey, Daniel et al. (2010). "Subspace Gaussian Mixture Models for Speech Recognition". In: *Proceedings of ICASSP*.

Rousseau, Anthony, Paul Deléglise, and Yannick Estève (2012). "TED-LIUM: An Automatic Speech Recognition Dedicated Corpus". In: *Proceedings of LREC*. European Language Resources Association (ELRA), pp. 125–129.

References IV

Tan, Tien-Ping and Laurent Besacier (2007). "Acoustic Model Interpolation for Non-native Speech Recognition". In: *Proceedings of ICASSP*.

Tan, Tien-Ping, Laurent Besacier, and Benjamin Lecouteux (2014). "Acoustic Model Merging using Acoustic Models from Multilingual Speakers for Automatic Speech Recognition". In: *Proceedings of International Conference on Asian Language Processing (IALP)*.