Pushing the frontiers of speech processing — What does it take to tackle new languages and domains? [Working Draft]

Florian Metze, Samuel Thomas,
Bhuvana Ramabhadran, and Brian Kingsbury

Carnegie Mellon University and IBM

Florian Metze

- Associate Research Professor at Carnegie Mellon University (LTI/SCS)
- End-to-end Speech Recognition
- Articulatory Features for Speech Recognition
- Multi-media Analysis
- fmetze@cs.cmu.edu

ルール・マナーを守りましょう
Please follow the rules and be well-mannered
早稲田大学
WASEDA UNIVERSITY

Samuel Thomas

- Researcher in the IBM Watson Group
- Automatic Speech Recognition
- Feature Engineering and Acoustic Modeling
- sthomas@us.ibm.com

Bhuvana Ramabhadran

- Research Manager in the IBM Watson Group
- Automatic Speech Recognition
- Deep Learning for Acoustic and Language Modeling
- Keyword Search
- `bhuvana@us.ibm.com`

Brian Kingsbury

- Research Scientist in the IBM Watson Group
- Deep Learning
- Automatic Speech Recognition
- Keyword Search
- `bedk@us.ibm.com`

Outline (3 hours)

1. Introduction
2. Tackling a New Language
3. Tackling a New Domain
4. Lessons Learnt with Current Neural Network Technologies
5. Research Topics, Challenges, and New Ideas
6. End-to-End Systems
7. Virtual Machines and Tools
8. Conclusions

Introduction

Introduction

Outline

1. The shift in speech based user interfaces
2. Building applications on the information rich speech signal
   - Automatic speech recognition
   - Speaker recognition
   - Speaker diarization
   - Language identification
   - Processing social signals
3. Impact of speech technologies across languages and domains
4. The Speech Recognition Case Study
   - What is under the hood for speech recognition technologies?
   - Building various ASR module and the impact of trascribed data
5. Building ASR Systems in New Languages
   - Building from ASR systems from scratch
   - Is there room for sharing data from other languages?
6. Building ASR Systems in New Domain
   - Adaptation of an existing ASR system

Tackling New Languages

Tackling New Languages

Outline

1. IARPA Babel
2. Audio Keyword Search
3. What Language Characteristics Matter?
    - Morphology and vocabulary growth
    - Writing system
    - Tonal languages
    - Amount of available training data
4. A Recipe for a New Language
    - Pronunciations
    - Flat-start Initialization
    - Multilingual Features
    - Web Text

The IARPA Babel Program

"…to rapidly develop speech recognition capability for keyword search in a previously unstudied language, working with speech recorded in a variety of conditions with limited amounts of transcription."

Rapid Development

Time allowed for surprise language model building

Period	Time
1	4 weeks
2	3 weeks
3	2 weeks
4	1 week

The IARPA Babel Program

"...to rapidly develop speech recognition capability for keyword search in a previously unstudied language, working with speech recorded in a variety of conditions with limited amounts of transcription."

Babel Languages

Period 1	**Period 2**	**Period 3**	**Period 4**
Cantonese	Assamese	Kurmanji Kurdish	Pashto
Pashto	Bengali	Tok Pisin	Guaraní
Turkish	Haitian Creole	Cebuano	Igbo
Tagalog	Lao	Kazakh	Amharic
Vietnamese	Zulu	Telugu	Mongolian
	Tamil	Lithuanian	Javanese
		Swahili	Dholuo
			Georgian

N.B. These will be available from the LDC at $US 25.00 per language for non-members.

The IARPA Babel Program

"...to rapidly develop speech recognition capability for keyword search in a previously unstudied language, working with speech recorded in a variety of conditions with limited amounts of transcription."

Limited resources

Hours of transcribed training data

Period	Hours
1	100
2	10
3	3
4	40

N.B. In Periods 3 and 4, no phonetic lexicons.

The IARPA Babel Program

"...to rapidly develop speech recognition capability for keyword search in a previously unstudied language, working with speech recorded in a variety of conditions with limited amounts of transcription."

What is keyword search, and why focus on it?

Detection task: given

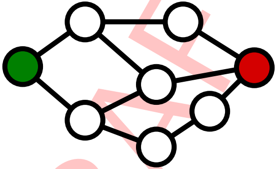
- a word or short phrase and
- a collection of speech data,

where does it occur, and how confident are you?

We can build practical keyword search from unreliable speech recognition.

Building an index

1. Generate a lattice for each segment in the collection.

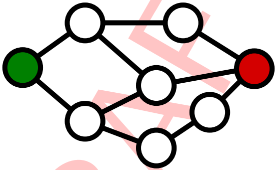


Lattice

Nodes	times
Edges	words (or phones) and posterior probabilities

Building an index

1. Generate a lattice for each segment in the collection.

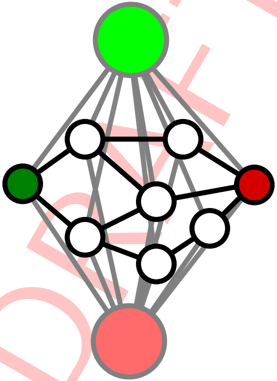


Transducer

Inputs	words (or phones)
Outputs	times
Costs	negative log-posteriors

Building an index

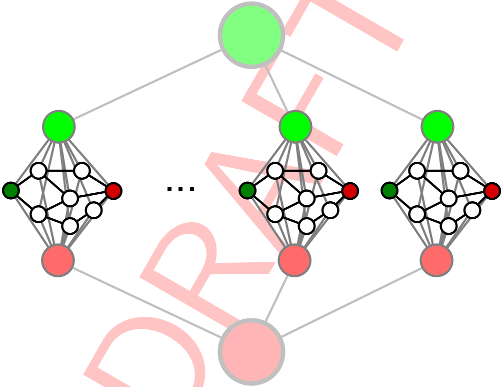
2. Produce the factor automaton for each segment.



Added edges have ϵ inputs, ϵ outputs, and no costs.

Building an index

3. Connect all the factor automata in parallel.



Added edges from start have ϵ inputs, segment ID outputs, and no costs. Added edges to end have ϵ inputs, ϵ outputs, and no costs.

Tackling a New Domain

Tackling a New Domain

Outline

1. Building Acoustic models in New Domains
    - With no adaptation data
        - Robust Features
        - Feature compensation and test-time adaptation
        - Multicondition training
    - With adaptation data
        - Model adaptation
2. Improving Language models for New Domains
    - In-domain data
    - Data from related data sources - Web data
3. Recipes for New Domains

Research Directions and New Modeling Techniques

Research Topics

Challenges

New Ideas

End-to-End Systems

Hands-On Experience with Virtual Machines

Practicalities

- We want to give you hands-on experience with building ASR systems
- You will be able to train a system on a Babel language (most likely 201 Haitian)
- You can then experiment with other Babel languages, or port the system to other domains
- To facilitate experimentation, we will distribute a Virtual Machine (VM)
- Read on to see how you can prepare

Virtual Machines and Tools

- Think of a VM as a “virtual” computer, in our case running Linux
- VMs allow sharing reproducible experiments easily
- https://github.com/srvk, http://speechkitchen.org as repositories
- https://www.vagrantup.com/ to build VMs
- https://www.virtualbox.org/ to run VMs (along with https://aws.amazon.com/)
- An “image” is a computer when it is turned off, it becomes an “instance” when you turn it on

Exercises

- We will share a Vagrantfile, plus an image on AWS (most likely), and/ or a Virtualbox OVA (less likely)
- Your best bet is to run the exercise on AWS
- So, you may want to sign up for an account first (`https://aws.amazon.com/getting-started/`)
- Familiarize yourself with how to start a Linux VM on "EC2" using a pre-configured Amazon Machine Image (AMI)
- Training a DNN-based recognizer on a GPU will cost some money, but the cost should not be dramatic
- Once you reproduced the basics, you can continue on AWS, or you can migrate to your own infrastructure

Eesen

- We will use the “Eesen” toolkit (https://github.com/srvk/eesen) for end-to-end speech recognition
- It is based on Kaldi (http://kaldi-asr.org/), but a bit smaller and easier to handle
- More details to follow

Conclusions

Thank You!

Any Questions?

References

[1] Xavier Anguera, Luis Javier Rodríguez-Fuentes, Andi Buzo, Florian Metze, Igor Szöke, and Mikel Peñagarikano.
QUESST 2014: Evaluating query-by-example speech search in a zero-resource setting with real-life queries.
In *Proc. ICASSP*, Brisbane; Australia, April 2015. IEEE.

[2] Rebecca Bates, Eric Fosler-Lussier, Florian Metze, Martha Larson, Gina-Anne Levow, and Emily Mower Provost.
Experiences with shared resources for research and education in speech and language processing.
In *Proc. INTERSPEECH*, San Francisco, CA; U.S.A., September 2016. ISCA.
Accepted.

[3] Yashesh Gaur, Walter S. Lasecki, Florian Metze, and Jeffrey P. Bigham.

The effects of automatic speech recognition quality on human transcription latency.
In *Proc. Web for All (W4A)*, Montreal; Canada, April 2016.
Best Paper.

[4] Yashesh Gaur, Florian Metze, and Jeffrey P. Bigham.
Manipulating word lattices to incorporate human corrections.
In *Proc. INTERSPEECH*, San Francisco, CA; U.S.A., September 2016. ISCA.
Accepted.

[5] Yashesh Gaur, Florian Metze, Yajie Miao, and Jeffrey P. Bigham.
Using keyword spotting to help humans correct captioning faster.

In *Proc. INTERSPEECH*, Dresden, Germany, September 2015. ISCA.

[6] Florian Metze, Eric Riebling, Eric Fosler-Lussier, Andrew Plummer, and Rebecca Bates.
The speech recognition virtual kitchen turns one.
In *Proc. INTERSPEECH*, Dresden, Germany, September 2015. ISCA.

[7] Florian Metze, Eric Riebling, Anne S. Warlaumont, and Elike Bergelson.
Virtual machines and containers as a platform for experimentation.
In *Proc. INTERSPEECH*, San Francisco, CA; U.S.A., September 2016. ISCA.
Accepted.

[8] Yajie Miao, Mohammad Gowayyed, and Florian Metze.

EESEN: End-to-End Speech Recognition using Deep RNN Models and WFST-based Decoding.
In *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, Scottsdale, AZ; U.S.A., December 2015. IEEE.
https://github.com/srvk/eesen.

[9] Yajie Miao, Mohammad Gowayyed, Florian Metze, Xingyu Na, Tom Ko, and Alex Waibel.
An empirical exploration of CTC acoustic models.
In *Proc. ICASSP*, Shanghai; China, March 2016. IEEE.

[10] Yajie Miao and Florian Metze.
Distance-aware DNNs for robust speech recognition.
In *Proc. INTERSPEECH*, Dresden, Germany, September 2015. ISCA.

[11] Yajie Miao and Florian Metze.

On speaker adaptation of long short-term memory recurrent neural networks.
In *Proc. INTERSPEECH*, Dresden, Germany, September 2015. ISCA.

[12] Yajie Miao and Florian Metze.
New Era for Robust Speech Recognition – Exploiting Deep Learning, volume 1, chapter End-to-End Architectures for Speech Recognition.
Springer, 2016.
To appear.

[13] Yajie Miao and Florian Metze.
Open-domain audio-visual speech recognition: A deep learning approach.
In *Proc. INTERSPEECH*, San Francisco, CA; U.S.A., September 2016. ISCA.
Accepted.

[14] Yajie Miao, Hao Zhang, and Florian Metze.
Speaker adaptive training of deep neural network acoustic models using i-vectors.
IEEE/ACM Transactions on Audio, Speech and Language Processing, 23(11):1938–1949, November 2015.

[15] Yun Wang and Florian Metze.
Recurrent support vector machines for audio-based multimedia event detection.
In *Proc. ICMR*, New York, NY; U.S.A., June 2016. ACM.

[16] Yun Wang, Leonardo Neves, and Florian Metze.
Audio-based multimedia event detection using deep recurrent neural networks.
In *Proc. ICASSP*, Shanghai; China, March 2016. IEEE.

[17] Shinji Watanabe, Marc Delcroix, Florian Metze, and John R. Hershey, editors.

New Era for Robust Speech Recognition – Exploiting Deep Learning, volume 1.
Springer, 2016.
To appear.

[18] Shoou-I Yu, Lu Jiang, Zhongwen Xu, Zhenzhong Lan, Shicheng Xu, Xiaojun Chang, Xuanchong Li, Zexi Mao, Chuang Gan, Yajie Miao, Xingzhong Du, Yang Cai, Lara Martin, Nikolas Wolfe, Anurag Kumar, Huan Li, Ming Lin, Zhigang Ma, Yi Yang, Deyu Meng, Shiguang Shan, Pinar Duygulu Sahin, Susanne Burger, Florian Metze, Rita Singh, Bhiksha Raj, Teruko Mitamura, Richard Stern, Alexander Hauptmann, Zhiyong Cheng, Jialie Shen, Xingzhong Du, and Xiaofang Zhou.
Cmu informedia@trecvid 2015: Med/ sin/ lnk/ sed.
In *Proc. TrecVID*, Gaithersburg, MD; U.S.A., December 2015. NIST.