

# Investigating techniques for low resource conversational speech recognition

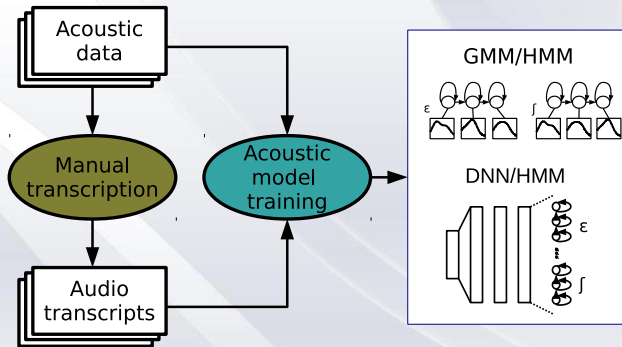
Antoine Laurent, Thiago Fraga da Silva,  
Lori Lamel, Jean-Luc Gauvain

March 23rd, 2016

VOCAPIA  
*research*



# Speech recognition system development



Speech-to-text (STT) and Keyword Search (KWS) system development usually requires a large amount of transcribed data

# Low resource STT/KWS

---

- Limited amount of transcribed data
- Little knowledge about the language: pronunciation, grammar, structure, writing system...

# IARPA Babel Program challenges

To develop speech technologies for **rapidly** creating effective KWS systems for a **large variety of languages** and with significantly **less training data** than has been used in the current state-of-the-art systems

Characteristic	Year 1	Year 2	Year 3
Full Language Pack	80h	60h	40h
(Very) Limited Language Pack	10h	10h	3h
Dev time (surprise language)	4 weeks	3 weeks	2 weeks
Pronunciation dictionary	yes	yes	no

# Techniques addressed in this work

---

**The goal of this work** is to investigate various techniques in order to build effective STT/KWS systems for low resource conversational speech

# Techniques addressed in this work

---

**The goal of this work** is to investigate various techniques in order to build effective STT/KWS systems for low resource conversational speech

- 1 Subword keyword search
- 2 Data selection for acoustic model training
- 3 Semi-supervised acoustic model training
- 4 Webtext data retrieval for language modeling
- 5 Acoustic data augmentation
- 6 Using neural network language models

## Available data for Year 3

---

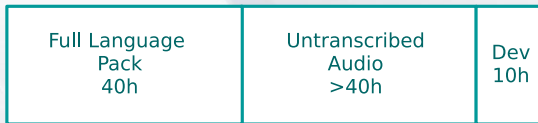
**40h condition**

Full Language Pack 40h	Untranscribed Audio >40h	Dev 10h
------------------------------	--------------------------------	------------

- Monolingual bottleneck features (Grézl, Karafiát, 2013)

# Available data for Year 3

40h condition



3h condition  
(VLLP vs ALP)



- Multilingual bottleneck features
- **VLLP**: Very Limited Language Pack
- **ALP**: Active Learning Pack



# Speech-to-text systems

---

- Swahili conversational telephone speech (IARPA-Babel202b-v1.0d)
- Graphemic dictionaries
- GMM/HMM acoustic models (also DNN/HMM for comparison)
- Bottleneck features (provided by BUT)
- Backoff and neural network  $n$ -gram language models
- Webtexts (pre-processed by BBN) for language modeling

# Keyword search

---

- Keyword list: defined according to transcription statistics
- Out-of-vocabulary (OOV) keywords: any keyword having at least one OOV word

# Keyword search

---

- Keyword list: defined according to transcription statistics
- Out-of-vocabulary (OOV) keywords: any keyword having at least one OOV word
- Search: based on the method described in (Hartmann et al, 2014)
- Keyword search:
  - Decode with word and subword based systems
  - Confusion networks are searched to locate all sequence of words/subwords that correspond to each keyword
  - Keyword hits are combined based on time codes

# Keyword search

---

- Keyword list: defined according to transcription statistics
- Out-of-vocabulary (OOV) keywords: any keyword having at least one OOV word
- Search: based on the method described in (Hartmann et al, 2014)
- Keyword search:
  - Decode with word and subword based systems
  - Confusion networks are searched to locate all sequence of words/subwords that correspond to each keyword
  - Keyword hits are combined based on time codes
- Subword units are obtained iteratively via language model perplexity optimization

what a peaceful place → what a peaceful place

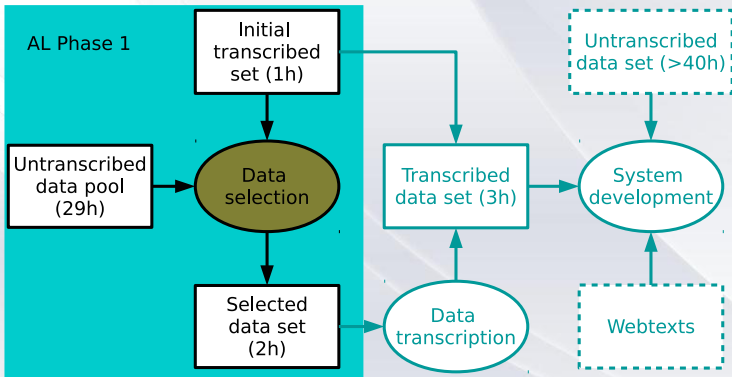
# Keyword search

VLLP system : with SST and Webdata

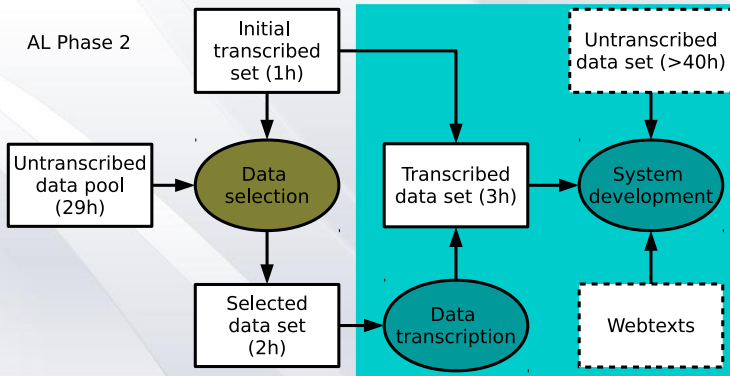
Keyword hits	All	In-vocabulary	Out-of-vocabulary
Word	0.436	0.458	0.268
Sub-word (5-gram)	0.371	0.367	0.409
Sub-word (6-gram)	0.375	0.369	0.419
Sub-word (7-gram)	0.367	0.362	0.409
4-way combination	0.458	0.461	0.456
<b>Absolute gain</b>	<b>2.2%</b>	<b>0.3%</b>	<b>18.8%</b>

$$ATWV(k, t) = 1 - P_{MISS}(k, t) - \beta P_{FA}(k, t)$$

# Data selection



# Data selection



**Question:** Is the automatic selection (AL) better than the baseline selection (VLLP)?

# Data selection

---

- Two stage selection method: (Fraga-Silva et al, 2015)
  - Use the letter density to select a subset of data (e.g. 10h)
  - Select 2h within this subset by maximizing the HMM state entropy



# Data selection

- Two stage selection method: (Fraga-Silva et al, 2015)
  - Use the letter density to select a subset of data (e.g. 10h)
  - Select 2h within this subset by maximizing the HMM state entropy

System	WER	ATWV
VLLP (baseline)	58.5	0.419
AL (data selection)	57.4	0.421
VLLP + SST + Webdata	50.5	0.458
AL + SST + Webdata	50.2	0.458
<b>Absolute gain</b>	<b>0.3-1.1%</b>	<b>&lt;0.3%</b>

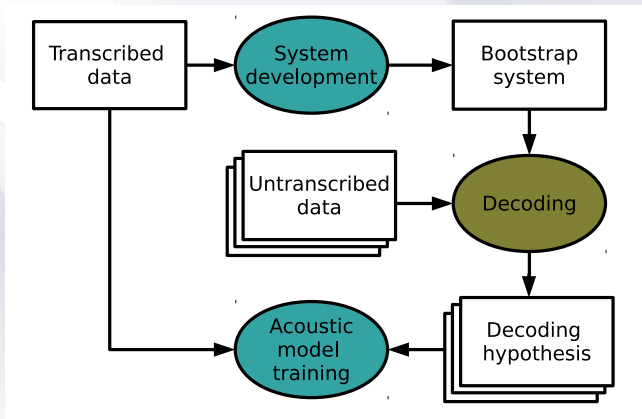
# Data selection

- Two stage selection method: (Fraga-Silva et al, 2015)
  - Use the letter density to select a subset of data (e.g. 10h)
  - Select 2h within this subset by maximizing the HMM state entropy

System	WER	ATWV
VLLP (baseline)	58.5	0.419
AL (data selection)	57.4	0.421
VLLP + SST + Webdata	50.5	0.458
AL + SST + Webdata	50.2	0.458
<b>Absolute gain</b>	<b>0.3-1.1%</b>	<b>&lt;0.3%</b>

- On the 6 Year-3 IARPA-Babel languages: WER gains 0.1%-2.4% and ATWV gains 0.7%-4.0% (ATWV)

# Semi-supervised acoustic model training



# Semi-supervised acoustic model training

System	Without Web	
	WER	ATWV
VLLP (3h)	58.5	0.419
VLLP (3h) + SST (70h)	57.9	0.421
<b>Absolute gain</b>	<b>0.6%</b>	<b>0.2%</b>

# Semi-supervised acoustic model training

System	Without Web	
	WER	ATWV
VLLP (3h)	58.5	0.419
VLLP (3h) + SST (70h)	57.9	0.421
<b>Absolute gain</b>	<b>0.6%</b>	<b>0.2%</b>

No gain on FLP: 40h transcribed + 40h untranscribed

# Adding webdata for language modeling

---

- Texts automatically retrieved from the Web (Zhang et al, 2015)
- Conversational-like queries submitted to a search engine
- 16M words, 200k word vocabulary

# Adding webdata for language modeling

- Texts automatically retrieved from the Web (Zhang et al, 2015)
- Conversational-like queries submitted to a search engine
- 16M words, 200k word vocabulary

System	Without Web		With Web	
	WER	ATWV	WER	ATWV
VLLP (3h)	58.5	0.419	52.4	0.454
+ SST (70h)	57.9	0.421	50.5	0.458

# Adding webdata for language modeling

- Texts automatically retrieved from the Web (Zhang et al, 2015)
- Conversational-like queries submitted to a search engine
- 16M words, 200k word vocabulary

System	Without Web		With Web		Absolute gain	
	WER	ATWV	WER	ATWV	WER	ATWV
VLLP (3h)	58.5	0.419	52.4	0.454	<b>6.1%</b>	<b>3.5%</b>
+ SST (70h)	57.9	0.421	50.5	0.458	<b>7.4%</b>	<b>3.7%</b>



# Adding webdata for language modeling

- Texts automatically retrieved from the Web (Zhang et al, 2015)
- Conversational-like queries submitted to a search engine
- 16M words, 200k word vocabulary

System	Without Web		With Web		Absolute gain	
	WER	ATWV	WER	ATWV	WER	ATWV
VLLP (3h)	58.5	0.419	52.4	0.454	<b>6.1%</b>	<b>3.5%</b>
+ SST (70h)	57.9	0.421	50.5	0.458	<b>7.4%</b>	<b>3.7%</b>
<b>Absolute gain</b>	0.6%	0.2%	<b>1.9%</b>	0.4%		

# Adding webdata for language modeling

- Texts automatically retrieved from the Web (Zhang et al, 2015)
- Conversational-like queries submitted to a search engine
- 16M words, 200k word vocabulary

System	Without Web		With Web		Absolute gain	
	WER	ATWV	WER	ATWV	WER	ATWV
VLLP (3h)	58.5	0.419	52.4	0.454	<b>6.1%</b>	<b>3.5%</b>
+ SST (70h)	57.9	0.421	50.5	0.458	<b>7.4%</b>	<b>3.7%</b>
<b>Absolute gain</b>	<b>0.6%</b>	<b>0.2%</b>	<b>1.9%</b>	<b>0.4%</b>		
FLP (40h)	43.1	0.507	41.5	0.520	<b>1.6%</b>	<b>1.3%</b>

# Acoustic data augmentation - VLLP

---

- Systems with SST and Webdata
- First, the multilingual BN DNN was fine tuned to Swahili VLLP (3h)
- 4 copies of data created by adding noise
- Additional 4 copies created by varying pitch

# Acoustic data augmentation - VLLP

- Systems with SST and Webdata
- First, the multilingual BN DNN was fine tuned to Swahili VLLP (3h)
- 4 copies of data created by adding noise
- Additional 4 copies created by varying pitch

<b>DNN bottleneck features</b>	<b>WER</b>	<b>ATWV</b>
Multilingual + fine tuning (3h)	48.2	0.439
+ noise (x4)	47.0	0.458

# Acoustic data augmentation - VLLP

- Systems with SST and Webdata
- First, the multilingual BN DNN was fine tuned to Swahili VLLP (3h)
- 4 copies of data created by adding noise
- Additional 4 copies created by varying pitch

<b>DNN bottleneck features</b>	<b>WER</b>	<b>ATWV</b>
Multilingual + fine tuning (3h)	48.2	0.439
+ noise (x4)	47.0	0.458
+ pitch variation (x4)	46.7	0.453
<b>Absolute gain (wrt fine tuned)</b>	<b>1.5%</b>	<b>1.9%</b>

# Acoustic data augmentation - FLP

---

- FLP systems without SST and with Webdata
- Monolingual features (40h)
- 4 copies of data created by adding noise

# Acoustic data augmentation - FLP

- FLP systems without SST and with Webdata
- Monolingual features (40h)
- 4 copies of data created by adding noise

DNN bottleneck features	WER	ATWV
Monolingual (40h)	41.5	0.520
+ noise (4x)	40.5	0.538
<b>Absolute gain</b>	<b>1.0%</b>	<b>1.8%</b>

# Neural network language models

---

- Feed-forward 4-gram neural network language models
- 4 layers, 12k word short list



# Neural network language models

- Feed-forward 4-gram neural network language models
- 4 layers, 12k word short list

System	With Web	
	WER	ATWV
FLP (40h)	41.5	0.520
+ noise (4x)	40.5	0.538
+ NNLM	39.1	0.540
<b>Absolute gain</b>	<b>1.4%</b>	<b>0.2%</b>

# Summary

Absolute gain

★ < 0.5%

★★ 0.5 – 1.5%

★★★ 1.5 – 3.0%

★★★★ > 3.0%

Technique	3h systems		40h systems	
	WER	ATWV	WER	ATWV
Subword KWS	-	★★★★	-	★★★★
Data selection	★	none	-	-
SST	★★★★	★	none	none
Webtexts	★★★★★	★★★★★	★★★★	★★
Data augmentation	★★★★	★★★★	★★	★★★★
NNLMs	-	-	★★	★

# Thank you

## Acknowledgments

This research was in part supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0013. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.