SPEECH RECOGNITION AND KEYWORD SPOTTING PERFORMANCE ANALYSIS ACROSS LANGUAGES

Jake Vasilakes, Haipeng Wang, Anton Ragni, Mark Gales, and Kate Knill

Cambridge University Engineering Department

{jav39,hw443,ar527,mjfg,kate.knill}@eng.cam.ac.uk

1 INTRODUCTION

- Cross-lingual performance analysis
  - Expect variation across languages due to, e.g.
    - vocabulary size
    - number of phones
  - Observe unexpected variations
    - How can we improve models?
- Performance prediction
  - Primary: keyword spotting (KWS)
  - Secondary: automatic speech recognition (ASR)
- Estimate quantity of data required to achieve target performance
- Carried out within the IARPA Babel Program
  - Automatic speech recognition (ASR) and keyword spotting (KWS) for low-resource languages

2 PERFORMANCE METRICS

- KWS: Maximum Term Weighted Value (MTWV)

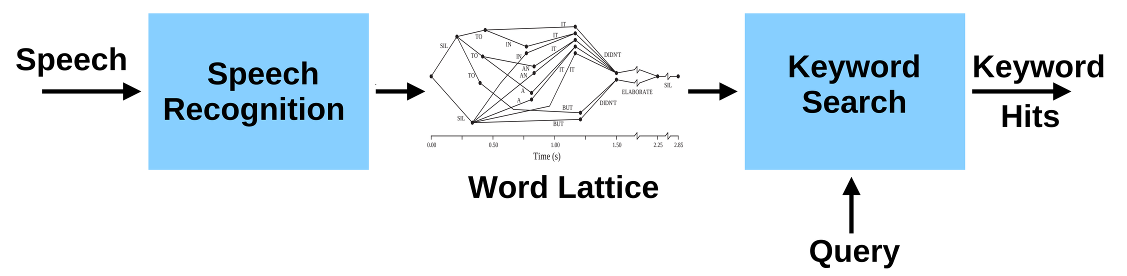
$$\text{MTWV} = \max_{\theta}\{1-(P_{\mathtt{miss}}(\theta)+\beta P_{\mathtt{FA}}(\theta))\}$$

  - where $\beta \approx 1000$
- ASR: Token Error Rate (TER)
- Root grapheme error rate (GER)
  - Represents grapheme confusability
  - Root grapheme: grapheme without unicode attributes

Char	Unicode description	Root grapheme	Grapheme
ı	LATIN SMALL LETTER I	G6	G6;D2D3D6

  - Computation
    - Maximum-likelihood, speaker independent, GMM system with PLP features
    - Weakened LM
    - Computed over training data

3 SYSTEM DESCRIPTION



- Data
  - 80h transcribed audio data
  - Conversational telephone speech
  - Language models built from transcriptions
- ASR
  - Root graphemic lexica
  - Joint decoding of Tandem and Hybrid systems
  - Single decode using combination of log-likelihoods
- KWS
  - Word lattices from output of ASR joint decoder
  - Query represented as a WFSA
  - IV query: composed with word lattice
  - OOV query: composed with grapheme lattice

4 ANALYSIS FRAMEWORK

- Examined correlation of attributes with performance
- Performed over 11 languages from the Babel project

Language	Script	Family	#Phones	$\lvert V\rvert (10^3)$
Cebuano	Latin	Austronesian	28	15.7
Kurmanji Kurdish		Iranian	37	14.9
Lithuanian		Balto-Slavic	60	32.1
Swahili		Niger-Congo	38	24.9
Tagalog		Austronesian	48	23.7
Tok Pisin		Creole	37	6.5
Zulu		Niger-Congo	47	60.9
Kazakh	Non-Latin	Altaic	61	23.3
Pashto		Iranian	44	21.0
Tamil		Dravidian	34	57.8
Telugu		Dravidian	50	36.9

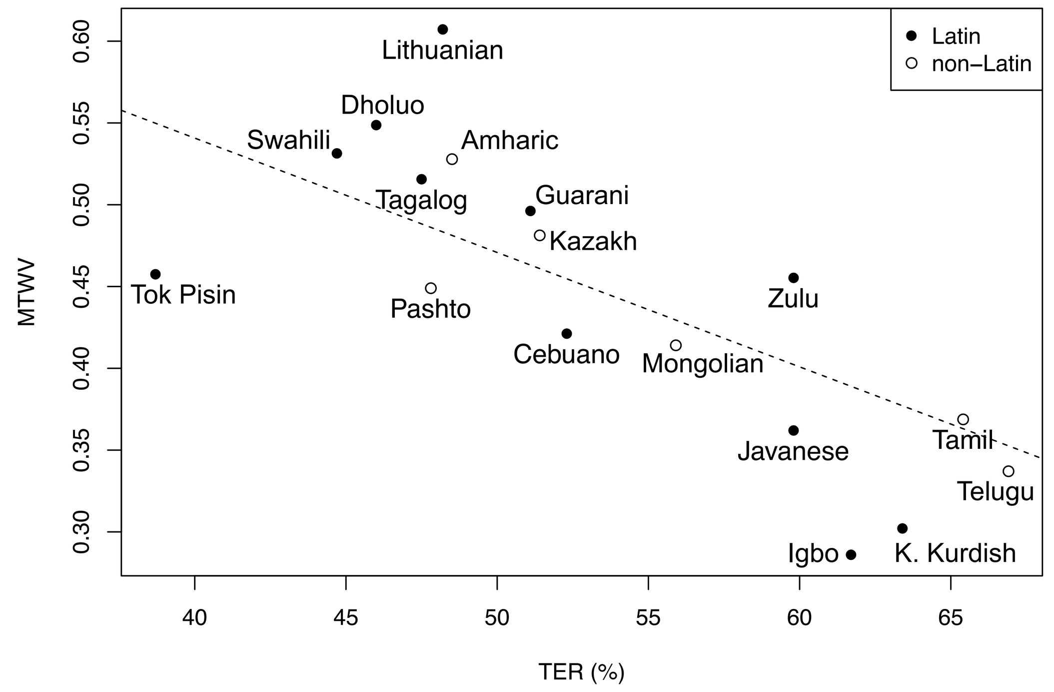
- 3 groups of attributes investigated

Linguistic	# Graphemes # Phones
Data	Signal-noise ratio (SNR) Mean opinion score (MOS) Vocabulary size # Frames
Model	Language model perplexity % out-of-vocabulary terms for ASR Root grapheme error rate (GER)

  - TER also investigated with respect to MTWV
  - Pearson's correlation (PCC) measured between each attribute and both TER and MTWV

5 ANALYSIS

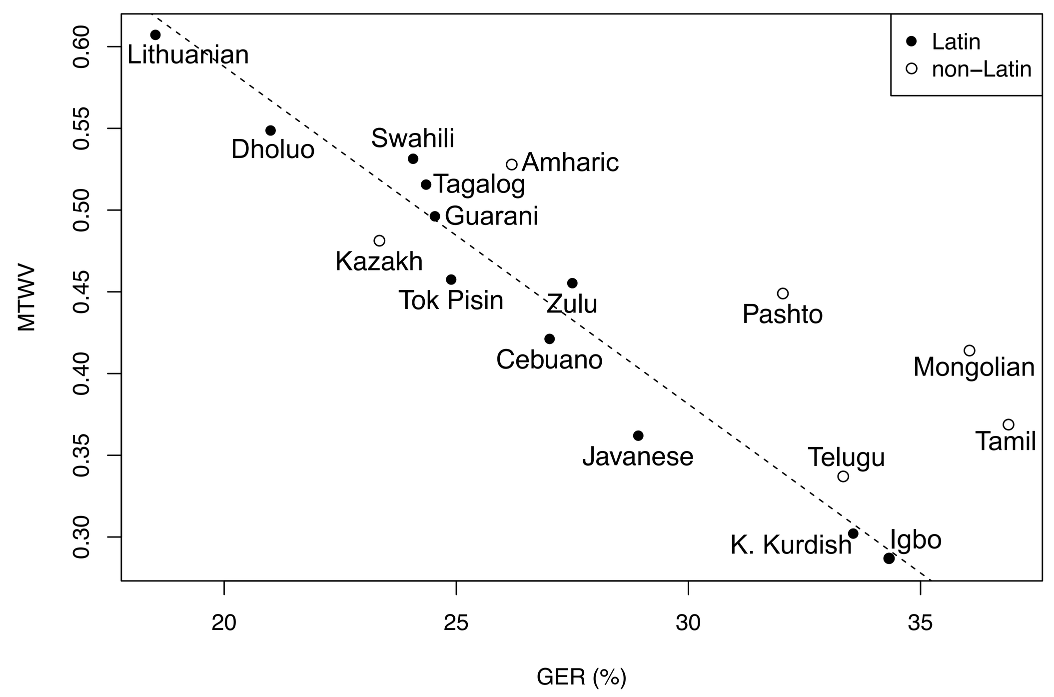
- TER against MTWV
  - PCC $= -.730$
  - But some languages do not behave as expected.
  - e.g. Lithuanian, Tok Pisin



- Other attributes
  - Most attributes were at best weakly correlated with performance
  - GER strongly correlated to MTWV

Script	PCC	
	MTWV	TER
Latin	$-.972$	$.684$
All	$-.887$	$.722$

  - TER correlation also shows possibility of prediction
- GER against MTWV
  - Lithuanian and Tok Pisin are no longer outliers
  - Non-Latin script outliers due to difference in grapheme set?



6 PREDICTIONS

- Predicted TER and MTWV
  - 6 held-out Babel languages
- Linear regression
  - GER as independent variable
  - MTWV predictions: Latin script languages only
  - Regression equations:

$$\mathtt{TER} = 19.68 + 1.21 \times \mathtt{GER}$$

$$\mathtt{MTWV} = 1.00063 - 0.02065 \times \mathtt{GER}$$

- Predictions well approximate observed values!

Language	%GER	%TER		MTWV	
		pred	obs	pred	obs
Dholuo	20.9	≈45	46.0	≈0.57	0.549
Guarani	24.5	≈49	51.1	≈0.50	0.496
Igbo	34.3	≈61	61.7	≈0.29	0.286
Javanese	28.5	≈54	59.8	≈0.41	0.362
Amharic†	25.6	≈51	48.5	≈0.47	0.528
Mongolian†	35.3	≈62	55.9	≈0.27	0.414

† Non-Latin script

7 CONCLUSIONS

- Variation in ASR and KWS performance across languages
  - Even given same system configuration
- Linguistic, data, and model attributes investigated
  - Most weakly correlated with performance
- Root grapheme error rate (GER)
  - Available at early stage of the system build
  - Strong correlation with MTWV
  - Correlated with TER
  - Able to predict performance for Latin script languages

LANGUAGE RELEASES

Cebuano (301) IARPA-babel301b-v1.0b; Kurmanji Kurdish (205) IARPA-babel205b-v1.0a; Lithuanian (304) IARPA-babel304b-v1.0b; Swahili (202) IARPA-babel202b-v1.0d; Tagalog (106) IARPA-babel106-v0.2g; Tok Pisin (207) IARPA-babel207b-v1.0a; Zulu (206) IARPA-babel206b-v0.1e; Kazakh (302) IARPA-babel302b-v1.0a; Pashto (104) IARPA-babel104b-v0.4bY; Tamil (204) IARPA-babel204b-v1.1b; Telugu (303) IARPA-babel303b-v1.0a;Dholuo(403) IARPA-babel403b-v1.0b;Guarani(305) IARPA-babel305b-v1.0b;Igbo(306) IARPA-babel306b-v2.0c;Javanese(402) IARPA-babel402b-v1.0a;Amharic(307) IARPA-babel307b-v1.0b;Mongolian(401) IARPA-babel401b-v2.0b;

ACKNOWLEDGEMENT

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U. S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0012. 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.