Cross-language mapping for small-vocabulary ASR in under-resourced languages: Investigating the impact of source language choice

Anjana Vakil and Alexis Palmer

UNIVERSITÄT DES SAARLANDES

Department of Computational Linguistics and Phonetics
University of Saarland, Saarbrücken, Germany

SLTU'14, St. Petersburg
15 May 2014

Outline

UNIVERSITÄT DES SAARLANDES

Small-vocabulary recognition: Why & how

Cross-language pronunciation mapping

- The Salaam method (Qiao et al. 2010)

Our contribution: Impact of source language choice

- Data & method
- Experimental results
- Conclusions

Ongoing & future work

Small-vocabulary recognition: Why & how

Goal: Enable non-experts to quickly develop basic speech-driven applications in any Under-Resourced Language (URL)

- Training/adapting recognizer takes data, expertise
- Many applications use $\leq$100 terms (e.g. Bali et al. 2013)

Strategy: Use **existing** HRL recognizer for small-vocab recognition in URLs (Sherwani 2009; Qiao et al. 2010)

Key: Mapped pronunciation lexicon

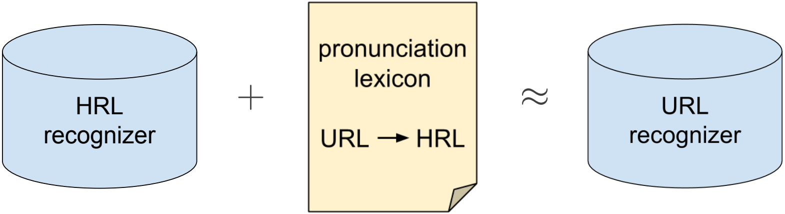
Terms in target lg. (URL) → Pronunciations in source lg. (HRL)

Yoruba			English
igba	i͡gba	→	igbə | ibə | ...?

Key: Mapped pronunciation lexicon

Terms in target lg. (URL) → Pronunciations in source lg. (HRL)

Yoruba			English
igba	ig͡ba	→	igbə | ibə | ...?



Cross-language pronunciation mapping

UNIVERSITÄT DES SAARLANDES

The Salaam Method (Qiao et al. 2010)

- Requires $\geq$1 sample per term (a few minutes of audio)
- Mimics phone decoding
- "Super-wildcard" recognition grammar:

$$\textit{term} \rightarrow \{*| **| ***\}_0^{10}$$

(∗ = any source-language phoneme)

- Iterative training algorithm finds confidence-ranked matches

igba → ibæə, ibɹə, ibɛə, ...

- Accuracy: ≈80-98% for $\leq$50 terms

Impact of source language choice

Hypothesis

More phoneme overlap between source/target languages $\rightarrow$ Easier pronunciation-mapping $\rightarrow$ Higher recognition accuracy

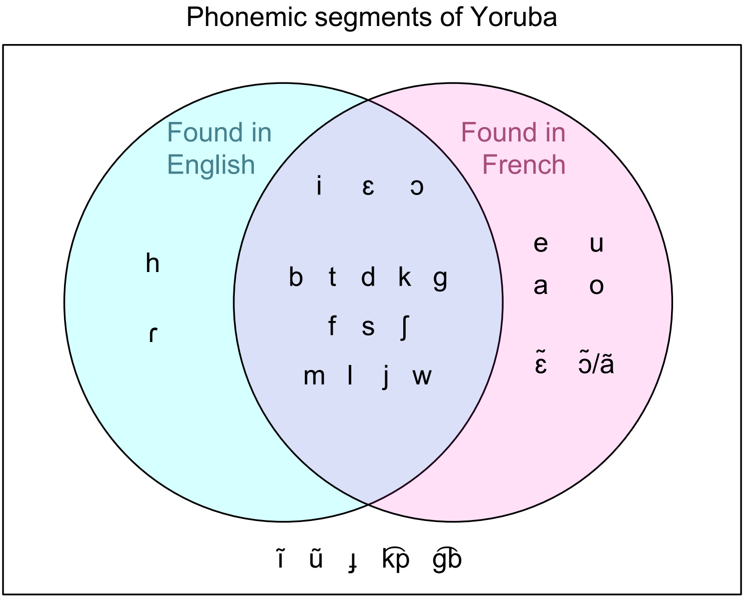
Impact of source language choice

Hypothesis

More phoneme overlap between source/target languages $\rightarrow$ Easier pronunciation-mapping $\rightarrow$ Higher recognition accuracy

Experiment

- Target language: Yoruba
- Source languages: English (US), French (France)


Phonemic segments of Yoruba
Found in English
Found in French
h
ɾ
i ɛ ɔ
b t d k g
f s ʃ
m l j w
e u
a o
ɛ̃ ɔ̃/ã
ĩ ũ ɟ k͡p g͡b

UNIVERSITÄT DES SAARLANDES

Data

- 25 Yoruba terms (subset of Qiao et al. 2010 dataset)
- 5 samples/term from 2 speakers (1 male, 1 female)
- Telephone quality (8 kHz)

Data & method

Data

- 25 Yoruba terms (subset of Qiao et al. 2010 dataset)
- 5 samples/term from 2 speakers (1 male, 1 female)
- Telephone quality (8 kHz)

Method

- Generate Fr./En. lexicons with Salaam (Qiao et al. 2010)
  - Microsoft Speech Platform (msdn.microsoft.com/library/hh361572)
  - 1, 3, and 5 pronunciations per term

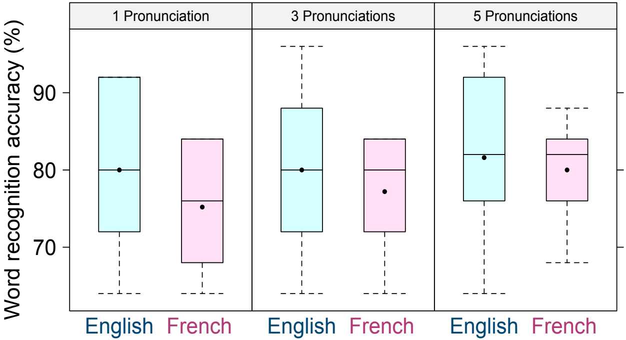
Data

- 25 Yoruba terms (subset of Qiao et al. 2010 dataset)
- 5 samples/term from 2 speakers (1 male, 1 female)
- Telephone quality (8 kHz)

Method

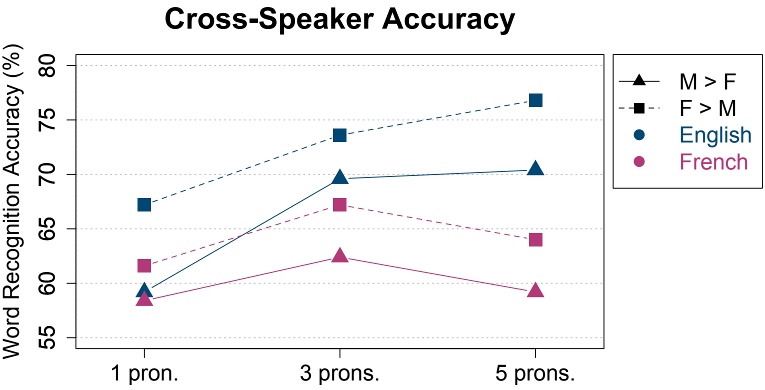
- Generate Fr./En. lexicons with Salaam (Qiao et al. 2010)
  - Microsoft Speech Platform (msdn.microsoft.com/library/hh361572)
  - 1, 3, and 5 pronunciations per term
- Compare mean word recognition accuracy
  - Same-speaker: Leave-one-out
  - Cross-speaker: Train M > Test F; F>M
  - *t*-tests for significance ($\alpha = 0.05$)

Same-speaker accuracy



1 Pronunciation		3 Pronunciations		5 Pronunciations	
80.0	75.2	80.0	77.2	81.6	80.0
$p = 0.20$		$p = 0.34$		$p = 0.59$	

Results



En. mean	63.2	71.6	73.6
Fr. mean	60.0	64.8	61.6
p (* $\leq$ .05)	0.41	0.04*	0.04*

Results

Accuracy by word type (**nasal**)

	English	French
Best	duro	ogba
	ogba	iba
	shii	mejo
	ogoji	ogoji
	mesan	**lehin**
	beeni	**tunse**
	⋮	⋮
	iba	**mesan**
	igba	**ookan**
	ogorun	**sun**
	meta	meji
	sun	bere
Worst	meji	igba

Conclusions

UNIVERSITÄT DES SAARLANDES

Hypothesis

More phoneme overlap between source/target languages $\rightarrow$ Easier pronunciation-mapping $\rightarrow$ Higher recognition accuracy

Predicted: French accuracy $>$ English accuracy
Observed: French accuracy $\leq$ English accuracy

Conclusions

UNIVERSITÄT DES SAARLANDES

Hypothesis

More phoneme overlap between source/target languages $\rightarrow$ Easier pronunciation-mapping $\rightarrow$ Higher recognition accuracy

Predicted: French accuracy $>$ English accuracy
Observed: French accuracy $\leq$ English accuracy

Possible explanations:

- Source languages may be too similar w.r.t. target language

Conclusions

UNIVERSITÄT DES SAARLANDES

Hypothesis

More phoneme overlap between source/target languages $\rightarrow$ Easier pronunciation-mapping $\rightarrow$ Higher recognition accuracy

Predicted: French accuracy $>$ English accuracy
Observed: French accuracy $\leq$ English accuracy

Possible explanations:

- Source languages may be too similar w.r.t. target language
- Better metric needed for evaluating source-target match

Conclusions

Hypothesis

More phoneme overlap between source/target languages $\rightarrow$
Easier pronunciation-mapping $\rightarrow$ Higher recognition accuracy

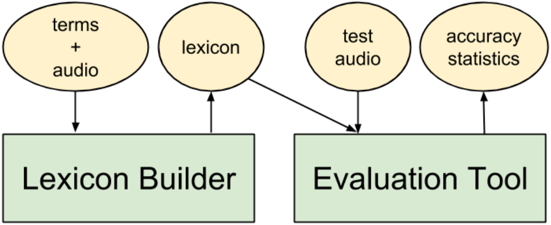
Predicted: French accuracy $>$ English accuracy
Observed: French accuracy $\leq$ English accuracy

Possible explanations:

- Source languages may be too similar w.r.t. target language
- Better metric needed for evaluating source-target match
- Baseline recognizer accuracy may play a role

Ongoing & future work

lex4all: Pronunciation **Lex**icons **for A**ny **L**ow-resource **L**anguage (Vakil et al. 2014)



`http://lex4all.github.io/lex4all`

Planned experiments:

- More source-target language pairs
- Discriminative training (Chan and Rosenfeld 2012)
- Algorithm modifications

References

- K. Bali, S. Sitaram, S. Cuendet, and I. Medhi. "A Hindi speech recognizer for an agricultural video search application". In: *ACM DEV*. 2013.
- H. Y. Chan and R. Rosenfeld. "Discriminative pronunciation learning for speech recognition for resource scarce languages". In: *ACM DEV*. 2012.
- F. Qiao, J. Sherwani, and R. Rosenfeld. "Small-vocabulary speech recognition for resource-scarce languages". In: *ACM DEV*. 2010.
- J. Sherwani. "Speech interfaces for information access by low literate users". PhD thesis. Carnegie Mellon University, 2009.
- A. Vakil, M. Paulus, A. Palmer, and M. Regneri. "lex4all: A language-independent tool for building and evaluating pronunciation lexicons for small-vocabulary speech recognition". In: *ACL 2014: System Demonstrations*. 2014.

Thank you! Thanks also to:

Roni Rosenfeld, Mark Qiao, Hao Yee Chan, Dietrich Klakow, Manfred Pinkal