lex4all: A language-independent tool for building and evaluating pronunciation lexicons for small-vocabulary speech recognition

Anjana Vakil, Max Paulus, Alexis Palmer, and Michaela Regneri
Saarland University, Computational Linguistics Dept. [anjanav,mpaulus,apalmer,regneri]@coli.uni-saarland.de

INTRODUCTION

lex4all allows non-experts to quickly and easily create a pronunciation lexicon for terms in any low-resource language (LRL), using:

- a small number of audio recordings
- a pre-existing recognition engine in a high-resource language (HRL)

The pronunciation lexicon

- maps terms in the target (LRL) vocabulary to sequences of phonemes in the source language (HRL)
- can be used to add small-vocabulary speech recognition functionality to applications in the LRL.

lex4all helps small-scale developers create speech interfaces in LRLs, without much data or speech technology expertise. Such interfaces can be very beneficial in areas where literacy rates are low or where PCs/internet connections are not always available [1, 2].

The application is open-source and freely available at:
`http://lex4all.github.io/lex4all`

ALGORITHM

We use the Salaam method [1, 2] for the automatic discovery of the best pronunciation sequence for each word in the target vocabulary.

The Salaam method [1, 2]:

- *"Super-wildcard" grammar:*

  Allows recognizer to treat each audio sample as "phrase" of 0-10 "words" with each "word" a sequence of 1-3 source-language phonemes, i.e.:

  $$\{*|**|***\}_0^{10}$$

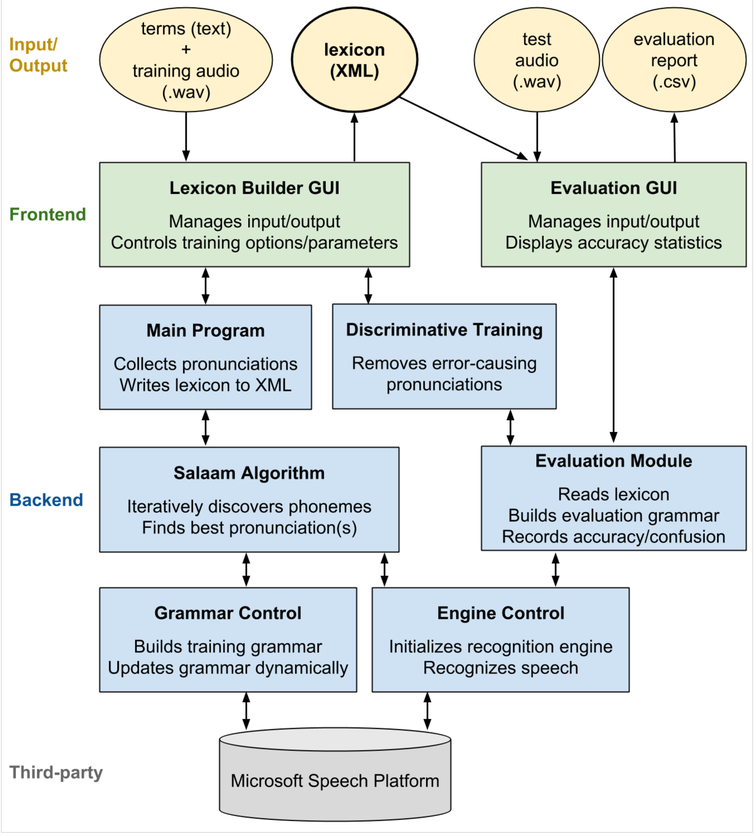
  where * represents a single phoneme of the source language.
- *Iterative training algorithm*

  Uses the grammar and HRL recognizer to discover the best pronunciation sequence(s) for each word in the target vocabulary, one phoneme at a time
- Yields better recognition than expert-written pronunciations [1]

SYSTEM OVERVIEW

lex4all is a desktop application for Windows, based on

- Microsoft Speech Platform [4]
- Salaam method for pronunciation mapping [1, 2] (see "Algorithm")



ADDITIONAL FEATURES

- **Discriminative training** [2]

  An additional training step removes pronunciations in the lexicon that may reduce recognition accuracy by matching multiple words in the vocabulary
- **Evaluation module**

  Facilitates research by automatically simulating recognition on a test set of audio samples. Reports recognition accuracy rates and confusion matrix.
- **Built-in audio recorder**

  Enables audio sample collection. Built using NAudio library [5].

CHALLENGE: RUNNING TIME

The main challenge we faced in engineering a user-friendly application based on Salaam (see "Algorithm") was the long training time due to the large "super-wildcard" grammar required by the algorithm.

- Original backend: 1-3 phonemes per sub-word
  - 40 phonemes (English) → 65,000+ possible combinations
  - Training time (25 words, 5 samples/word): approx. 60-120 minutes
- New backend: only 1 phoneme per sub-word
  - Training time (25 words, 5 samples/word): approx. 2-5 min. (≈20x faster)
- Evaluation
  - Tested on Yoruba data (25 words, 2 speakers, 5 samples/word/speaker)
  - Result: no significant drop in recognition accuracy (see Figure 1)

Figure 1. Evaluation of Word Recognition Accuracy

CONCLUSION & FUTURE WORK

lex4all enables rapid, automatic creation of pronunciation lexicons in any LRL, using an out-of-the-box recognizer for a HRL [4] and an existing algorithm for cross-language pronunciation mapping [1, 2].

We hope that this tool will help developers create speech interfaces for applications in LRLs, and facilitate research in small-vocabulary speech recognition for such languages.

Possible future extensions of the project include:

- **Online lexicon repository**

  Allowing users to upload created lexicons to an online repository would allow sharing and re-use of lexicons across languages/language families.
- **Additional source-language (HRL) recognizers**

  Microsoft offers recognizers in 20+ languages [4]. Using a source language that is more similar to the target LRL could improve recognition accuracy [3].

References

[1] F. Qiao, J. Sherwani, and R. Rosenfeld. 2010. "Small-vocabulary speech recognition for resource- scarce languages," *ACM DEV '10.*
[2] H.Y. Chan and R. Rosenfeld. 2012. "Discriminative pronunciation learning for speech recognition for resource scarce languages," *ACM DEV '12.*
[3] A. Vakil and A. Palmer. 2014. "Cross-language mapping for small-vocabulary ASR for low-resource languages: Investigating the impact of source language choice," *SLTU '14.*
[4] http://msdn.microsoft.com/en-us/library/dd266409
[5] http://naudio.codeplex.com/

Acknowledgments

Many thanks to Roni Rosenfeld, Hao Yee Chan, and Mark Qiao for generously sharing their data and providing valuable advice on implementing the Salaam method.