Analysis of Keyword Spotting Performance across IARPA Babel Languages

William Hartmann, Damianos Karakos, Roger Hsiao, Le Zhang, Tanel Alumäe, Stavros Tsakalidis, Richard Schwartz

Raytheon BBN Technologies, Cambridge, MA, USA {william.hartmann, damianos.karakos, le.zhang, stavros.tsakalidis, rich.schwartz}@raytheon.com

Introduction

- We analyze our keyword spotting performance across 16 languages from the IARPA Babel program.
- An open question from the Babel program is why so much variation exists between the performance of different languages.
- We demonstrate that features of the keywords explain much of the variation in performance within a language.
- This keyword-dependent variation must be taken into account when analyzing cross-language performance.
- The IARPA T&E team also provided inter-annotator agreement for four of the Babel languages.
- The inter-annotator agreement shows a remarkable correlation with ATWV, suggesting that the factors that make it difficult for a native speaker to consistently transcribe speech also impact ASR systems.

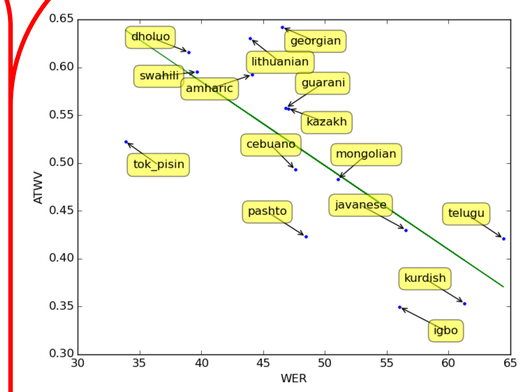
Experimental Setup

- We use the Sage speech recognition toolkit.
- Sage combines BBN's Byblos with open source toolkits such as Kaldi and CNTK.
- Sage also includes a cross-toolkit FST recognizer that supports models built using the various component technologies.
- All models are baseline monolingual DNN systems trained on 40 hours of transcribed speech.
- Keyword spotting is performed using both whole word and fuzzy phonetic search.

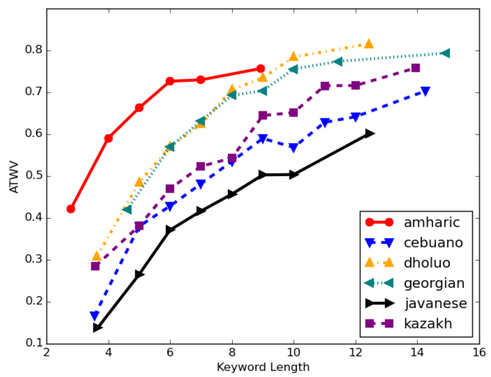
IARPA Babel Data

- We use 16 FLP language packs: Amharic, Cebuano, Dholuo, Georgian, Guarani, Igbo, Javanese, Kazakh, Kurdish, Lithuanian, Mongolian, Pashto, Swahili, Tamil, Telugu, and Tok Pisin*
- Each language contains about 40 hours of transcribed data.
- Lexicons are built using simple G2P rules.
- Trigram language models are built using only the available transcribed training data.

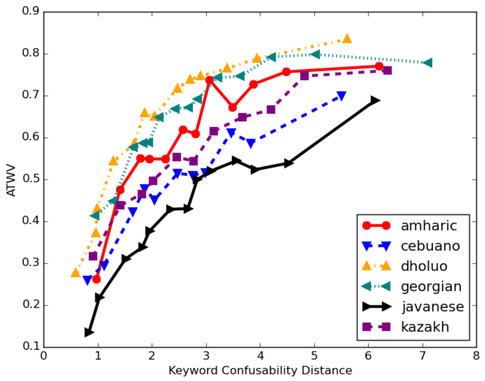
Bottleneck Feature Network Types



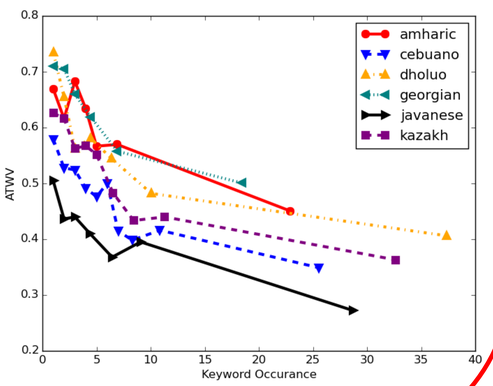
- While languages with lower WER tend to have higher ATWV, the relationship is not strong.
- Pashto and Georgian have similar WER, but their ATWV is 20 points apart.



- ATWV performance within a language can vary more than performance across languages depending on the keywords chosen.
- For all languages, as the length of the keywords increase, so does the ATWV.
- Even accounting for these keyword features, large gaps in performance between languages are still seen.

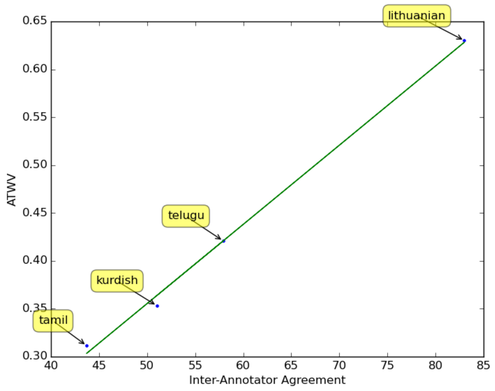


- A similar relationship is seen with keyword confusability distance.
- Keyword confusability distance is the average minimum Levenshtein distance for a keyword in each utterance, like a weighted keyword length.



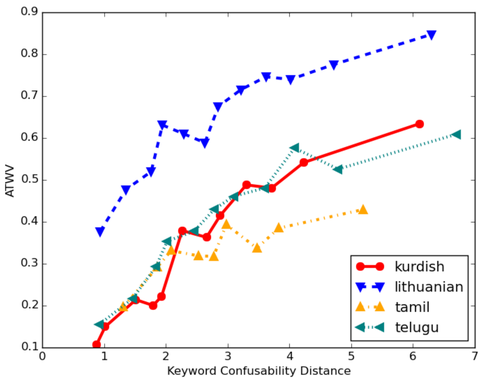
- ATWV is inversely proportional the number of occurrences of each keyword.
- This is partially due to the definition of ATWV—detections of rare words are worth more than common words.
- Not only do these keyword features correlate with ATWV, but they all of the various features correlate with each other as well.

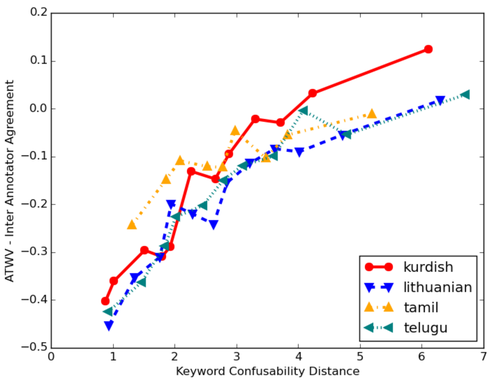
Inter-Annotator Agreement



- The IARPA Babel T&E team computed inter-annotator agreement for four languages.
- The agreement correlates strongly with ATWV for our baseline systems.
- The low-level agreement for some of the languages highlights the overall difficulty of the task.

- As in the previous figures, there is a strong relationship between keyword confusability and ATWV for the four languages.
- Overall, Lithuanian performs much better than the other languages.
- Lithuanian is also the language with the highest inter-annotator agreement.





- We can normalize the ATWV by the inter-annotator agreement by subtracting the agreement from the ATWV.
- The motivation is that performance can roughly only be as good as the level of inter-annotator agreement.
- This brings the results much closer together, with no more than 10 points of variation between languages.

- We have shown how the features of keywords can greatly impact the overall performance.
- The variation in performance across languages also correlates with the level of inter-annotator agreement.
- The factors that make it difficult for a native speaker to consistently transcribe speech also impact ASR systems.

This research was 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.

*Language Pack IDs: IARPA-babel{307b-v1.0b, 301b-v2.0b, 403b-v1.0b, 404b-v1.0a, 305b-v1.0a, 306b-v2.0c, 402b-v1.0b, 302b-v1.0a, 205b-v1.0a, 304b-v1.0b, 401b-v2.0b, 104b-v0.4bY, 202b-v1.0d, 303b-v1.0a, 207b-v1.0b, 204b-v1.1b}