Collecting and evaluating speech recognition corpora for nine Southern Bantu languages

Jaco Badenhorst, Charl van Heerden, Marelie Davel and Etienne Barnard

March 31, 2009

meraka
INSTITUTE
African Advanced Institute for Information
& Communications Technology

Outline

- Introduction
- Background:
  - ASR corpus design
  - The Lwazi ASR corpus
- Computational analysis
  - Approach
  - Analysis of phoneme variability
- Conclusion

Introduction

- Information flow in developing countries
  - Availability of alternate information sources is low in developing countries
  - Telephone networks (cellular) are spreading rapidly
- Spoken dialog systems (SDSs)
  - Widespread belief that impact can be significant
  - Speech-based access can empower semi-literate people
- Applications of SDSs
  - Education (Speech-enabled learning)
  - Agriculture
  - Health care
  - Government services

Introduction

- To implement SDSs: ASR and TTS systems are needed
- Main linguistic resources needed for telephone-based ASR systems:
  - Electronic pronunciation dictionaries
  - Annotated audio corpora
  - Recognition grammars
- Challenges:
  - ASR only available for handful of African languages
  - Lack of linguistic resources for African languages
  - Lack of relevant audio for specific application (language used, profile of speakers, speaking style, etc.)

ASR audio corpus

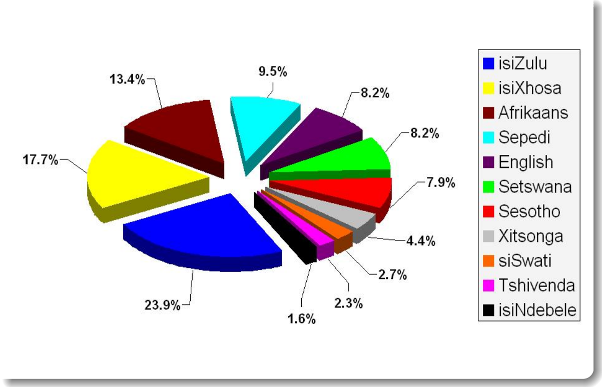
- Resource intensive process
- Factors that add to complexity:
  - Recordings of multiple speakers
  - Matching channel and style
  - Careful orthographic transcription
  - Markers required to indicate important events (eg. non-speech)
- Size of corpora:
  - Corpora of resource-scarce languages tend to be very small (1-10 hours of audio)
  - Contrasts with speech corpora used to build commercial systems (hundreds to thousands of hours)

Project Lwazi

- Three year (2006-2009) project commissioned by the South African Department of Arts and Culture
- Development of core speech technology resources and components (ASR, TTS, SDS, etc.)
- National pilot demonstrating potential impact of speech based systems in South Africa
- All 11 official languages of South Africa

Project Lwazi: Languages

- Distribution of home languages for South African population:
  - 9 Southern Bantu languages, 2 Germanic languages

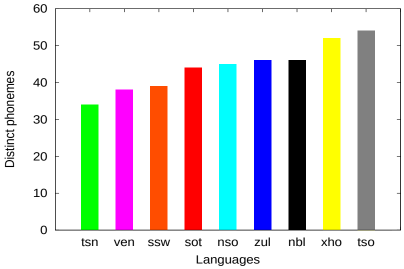


Project Lwazi

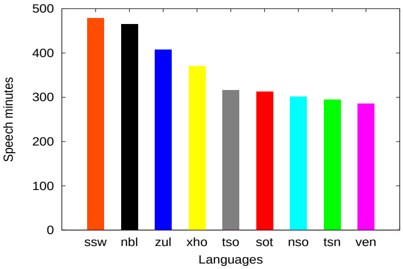
- ASR corpus:
  - Approximately 200 speakers per language
  - Speaker population selected to provide a balanced profile with regard to age, gender and type of telephone (cellphone/landline)
  - Read and elicited speech recorded over telephone channel
  - 30 Utterances/speaker:
    - 16 Randomly selected from phonetically balanced corpus
    - 14 Short words and phrases

Project Lwazi: Southern Bantu languages

Distinct phonemes per language



Speech minutes per language



Setswana	🟩
Tshivenda	🟪
siSwati	🟧
Sesotho	🟥
Sepedi	🟦
isiZulu	🟦
isiNdebele	⬛
isiXhosa	🟨
Xitsonga	⬜

- Amount of data within Lwazi ASR corpus

Computational analysis

- Goal:
  - Understand data requirements to develop a minimal system that is practically usable
  - Use as seed ASR system to collect additional resources
  - Implications of additional speakers and utterances
  - Develop tools:
    - Provide indication of data sufficiency
    - Potential for cross-language sharing

Computational analysis

- Approach:
  - Measure acoustic variance in terms of the separability between probability densities by modelling specific phonemes
  - Statistical measure provides an indication of the effect that additional training data will have on recognition accuracy
  - Utilise the same measure as indication of acoustic similarity across languages

Computational analysis

- Mainly focus on four languages here:
  - isiNdebele (nbl)
  - siSwati (ssw)
  - isiZulu (zul)
  - Tshivenda (ven)
- We report only on single-mixture context-independent models (similar trends observed for more complex models)
- Report on examples from several broad categories of phonemes (SAMPA) which occur most in target languages:
  - /a/ (vowels)
  - /m/ (nasals)
  - /b/ and /g/ (voiced plosives)
  - /s/ (unvoiced fricatives)

Analysis of phoneme variability

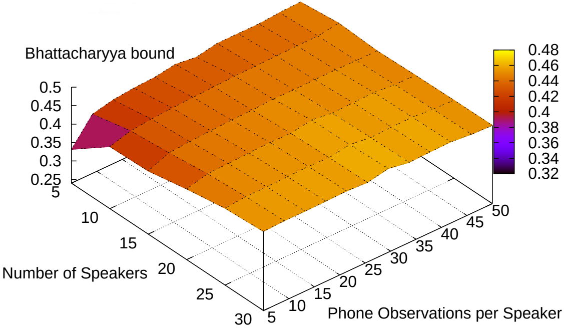


Figure: Speaker-and-utterance three-dimensional plot for the siSwati nasal /m/

Number of phoneme utterances

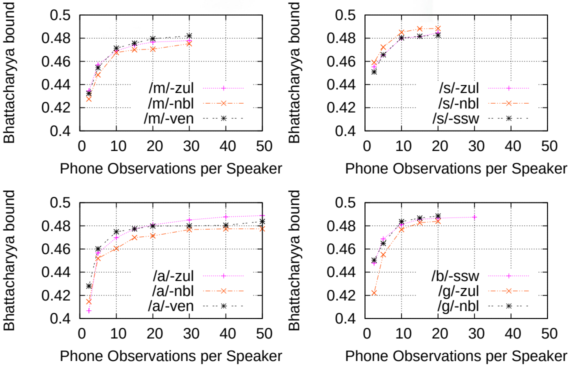


Figure: Effect of number of phoneme utterances per speaker on similarity measure for different phoneme groups using data from 30 speakers

Number of speakers

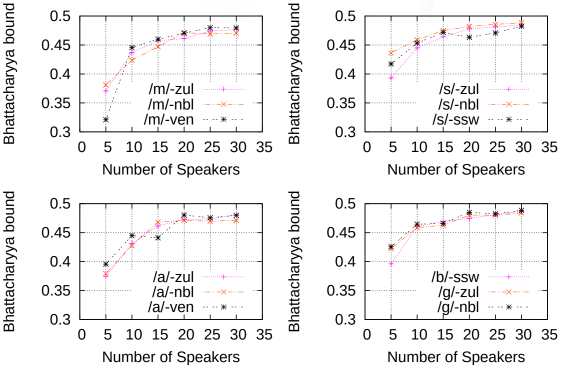
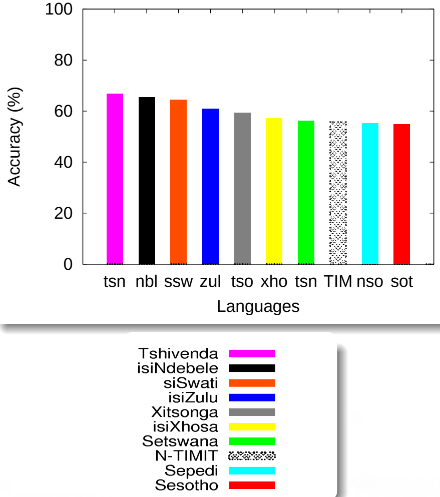


Figure: Effect of number of speakers on similarity measure for different phoneme groups using 20 utterances per speaker

Initial ASR Accuracy

Accuracy of phoneme recognisers



- Developed initial ASR systems for all of the Bantu languages
- Test sets: 30 speakers per language
- ASR system is *phoneme recogniser*, with flat language model
- A rough benchmark of acceptable phoneme accuracy: N-TIMIT

Impact of data reduction

- Division factor of 8:
  - Approximately 20 training speakers
  - Correlate well with the stable phoneme similarity values

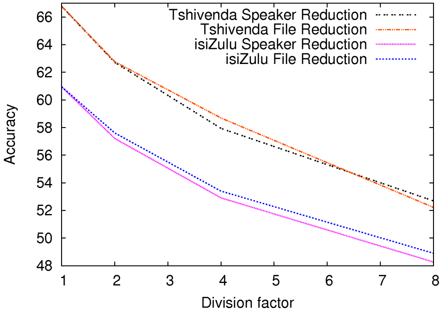


Figure: Reducing the number of speakers has (approximately) the same effect as reducing the amount of speech per speaker

Distances between phonemes

- Based upon proven stability of our phoneme models:
    - Phoneme similarity between phonemes across languages

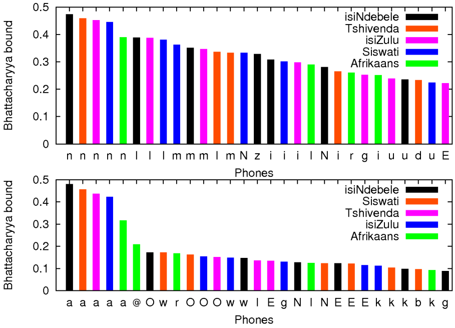


Figure: Effective distances for isiNdebele phonemes /a/ and /n/ and their closest matches.

Conclusion

- New method to determine data sufficiency
- Confirmed that different phoneme classes have different data requirements
- Our results suggest that similar phoneme accuracies may be achievable by using more speech from fewer speakers
- Based upon proven model stability we performed successful measurements of distances between phonemes of different languages

Conclusion

- Project Lwazi website:
  - http://www.meraka.org.za/lwazi
  - More info
  - Download corpora (ASR, TTS)
  - Download tools
  - Contact details