

# Term Projects:

11-751 Speech Recognition

09-17-2008

# Outline

- General Introduction
- Term Project Requirements
  - Proposal
  - Presentation
  - Final report
  - Deadlines
- Examples of Previous Term Projects
  - <http://www.is.cs.cmu.edu/11-751/wiki/Projects>
- Our Ideas for possible Term Projects

# Subject

- Goal of the speech term project:  
*Develop hand-on experience on speech research*
- What you are expected to do:
  - Propose a small research project
    - Find you own or
    - Accept one of our crazy ideas ...
  - Define the problem (do some literature study)
  - Design (and implement) a solution to this problem
  - Evaluate your approach(es)
  - Present the results (10-15 min slide presentation)
  - Write and hand in a report (5-10 pages)

# Teaming Up

- Building a Team to work together on a common project is fine with us
- If you build Teams:
  - The Proposal must be divided in sub-tasks, evenly distributed among the team members
  - The Proposal should clearly state which team member is doing which sub-task
  - Each member contributes to the project
  - Each member contributes to the presentation
  - Each member contributes to the report

# Suggestions

- Your own ideas are highly welcome
- ... but, hey, our ideas are as good as yours 😊
- Be realistic!
  - Don't put too much on your plate
  - Plan approximately 2 weeks full time
- Decide on one project and stick with it
- Start early
- Start early
- Start early

# Your Project Proposal

- Send a **1 page proposal** for your planned term project
  1. Describe the problem you want to address
  2. Describe how you want to solve the problem
  3. Formulate your expectations
  4. Give a timeline with milestones
- The purpose of this proposal is:
  - We get an impression of your plan
  - We can identify whether it is appropriate for a term project (too much, too few, too hard, too easy ...)
  - You think about what to do in advance
  - You and us have a timeline to monitor progress

# Your Presentation

- 10- 15 minutes talk, about:
  - Introduction to the problem you solved
  - The approach(es) you used
  - Experimental results
  - Analysis of the results
  - Possibly remodeled approach & Improvements
  - A demo system, if applicable
  - Lessons Learnt
  - Conclusion
  - Future work (if you want to continue)

# Your Final Report

Detailed report, **up to 10 pages**, including:

- {0.5 pages} State the problem you tackled
- {1 page} Literature survey, relevant work
- {1.5 page} Describe the approach(es) you used
- {2 pages} Present Experimental results
- {3 pages} Analyze the Results ...
  - ... Success?
  - Reasons for success or failure
  - Limitations of the approach, solution
  - What can be improved?
- {1 page} Lessons Learnt
  - Give the original timeline and milestones
  - Compare with the actual timeline and goals reached
- {1 page} Conclusions

# Important Deadlines

- Proposal DRAFT: **Wednesday, Oct 8**
  - Proposal submission : **Wednesday, Oct 15**
    - Due midnight by email to Wilson AND Ian, any data format
  - Progress Report:
    - Intermediate Report: **Wednesday, Nov 12**
- All reports are due midnight by email to Wilson AND Ian, and describe the project status in 2 pages:
- (1) Status: What has been done, what lies ahead,
  - (2) Progress: Compare your timeline with the reality and
  - (3) Adjustments: comment on differences (if any), reconsider
- Presentations: **Dec 3-8**
  - Final reports: **Monday, Dec 15**

**<http://www.speech.cs.cmu.edu/inner/conferences/>**

# Project Grading

- Task difficulty
- Amount of Support
- Quantity and quality of work
- Quality of results
- Presentation
- Quality of report

# List of Previous Projects

<http://www.is.cs.cmu.edu/11-751/wiki/Projects>

- Recognizing Words You've Never Heard      Nguyen Bach
- Speech Interface for SPICE      Sameer Badaskar
- Hardware Auditory Nerve Model      Jeffrey Johnston
- Children's Speech Recognition      Rohit Kumar
- Drunk Speech Classification      Joe Laws
- Alignment of Tamil Singing Speech      Udhay Nallasamy
- Speaker Recognition      Aaron Phillips
- Adaptive Language Modeling for CALO Meetings      Yitao Sun
- Finding a Minimal Training Set      Yi Wu

# In case ...

- Everybody is encouraged to come up with own ideas, visions, lifelong dreams ...
- If you don't have those, we will provide you with some of our (crazy) ideas :-
- *Don't panic*, we do not ask you to
  - Reduce WER to 0%  
(hmm, sounds like a good idea ...)
  - Beat the best commercial systems
  - Develop a SR system that works for me

# Project Ideas

- Language ID for multi-party S2S translation
- Disfluency recognition
- Recognition of OOV words
- Multimodal input of new-words
- wFST-based ASR for portable devices
- Audio classification of non-speech events
  - (Bowlingual, animal/bird identification)
- Spoken dialog systems
  - Voice-based control of your Roomba

# Language ID for multi-party S2S translation

- S2S translation typically assumes two participants and two languages
  - Primary user who holds the device ( $L_a$ )
  - Participant who speaks  $L_b$
- What if a third person is involved in the discourse
  - What language are they speaking?
  - Who is currently talking?

**Problem:** How to rapidly identify the current language being spoken?

**Given:** Baseline ASR systems, evaluation corpora

**Required:** Compare various approaches (**JANUS**)

# Disfluency Recognition

- For many speech-based applications (S2S translation) we want a cleaned transcript as input to down stream processes
  - Hesitations, repetitions, partial words
- Cleaning the ASR hypothesis not sufficient
  - Regions of disfluencies contain recognition errors

**Given:** Baseline ASR system, evaluation corpora

**Required:** Compare various approaches to recognize disfluencies during ASR decoding (**JANUS**)

# Recognition of OOV words

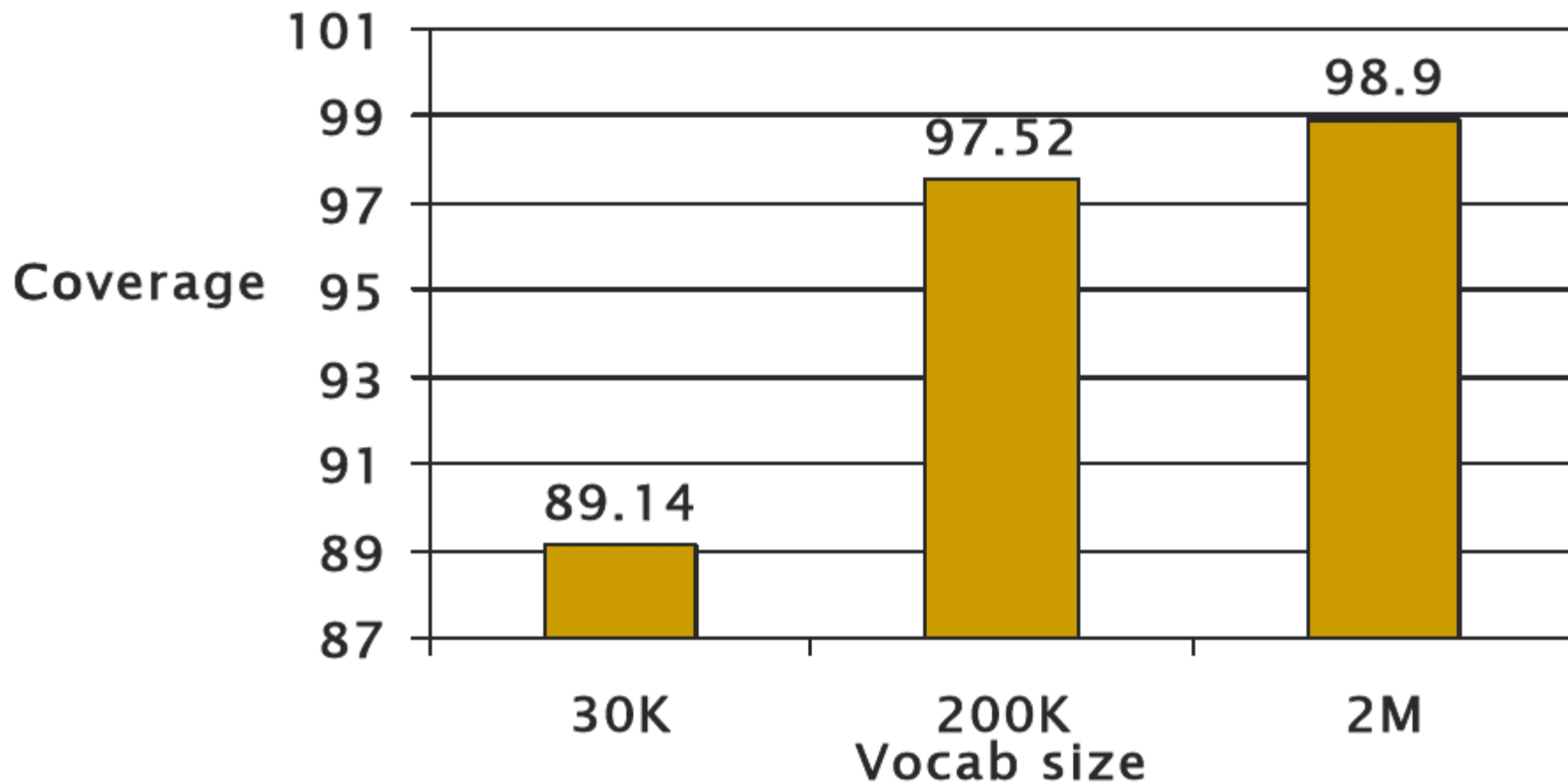
- Current ASR systems can only recognize vocabulary that occurs in the recognition lexicon
- Infrequent words are typically removed from lexicon
- Often these are semantically important or important for the application task; person, place and organization names

**Given:** Baseline ASR system, evaluation corpora

**Required:** Compare various approaches to recognize OOV words during (or after) ASR decoding (**JANUS**)

# Recognition of OOV words

## Vocabulary Coverage



# Multimodal input of new-words

- In S2S systems, users can add new words to the system, by entering a dialog and inputting the word, and then verifying the pronunciation of the word via TTS
- Can we improve the efficiency of this interface, by using a speech+text interface
  - Type in the word and speak it
  - Estimate best phone sequence by fusing both sources of information

# Project Ideas

- Language ID for multi-party S2S translation
- Disfluency recognition
- Recognition of OOV words
- Multimodal input of new-words
- **wFST-based ASR for portable devices**
- **Audio classification of non-speech events**
  - (Bowlingual, animal/bird identification)
- **Spoken dialog systems**
  - Voice-based control of your Roomba

---

## Course project ideas related to speaker recognition

1. Normalization methods for speaker verification
2. Improving BIC-based speaker clustering

Interested students please contact:

Qin Jin

[qjin@cs.cmu.edu](mailto:qjin@cs.cmu.edu)

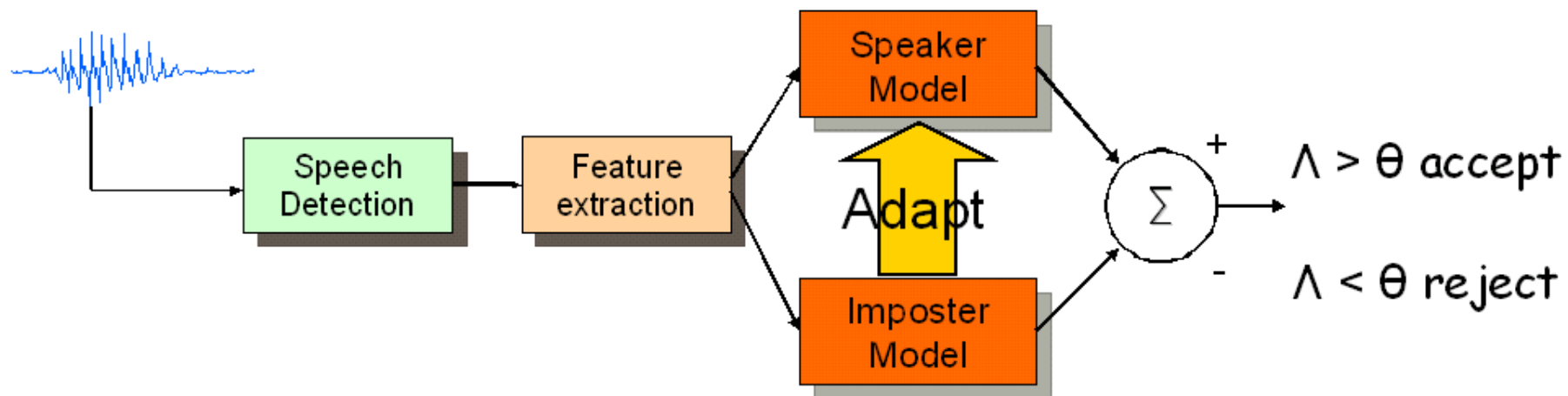
412-268-5477

InterACT, 407 S. Craig Street, Room 213

---

# Background – Speaker Verification

---



# Normalization Methods for Speaker Verification

---

## ➤ Feature Normalization:

- Feature warping: mapping the observed cepstral feature distribution to a normal distribution over a sliding window

## ➤ Score Normalization:

- normalizing the distribution of the likelihood scores
- Z-norm: normalizes the score distribution using target-specific statistics

$$S_{znorm}(X, \lambda) = \frac{S(X, \lambda) - \mu_\lambda}{\sigma_\lambda}$$

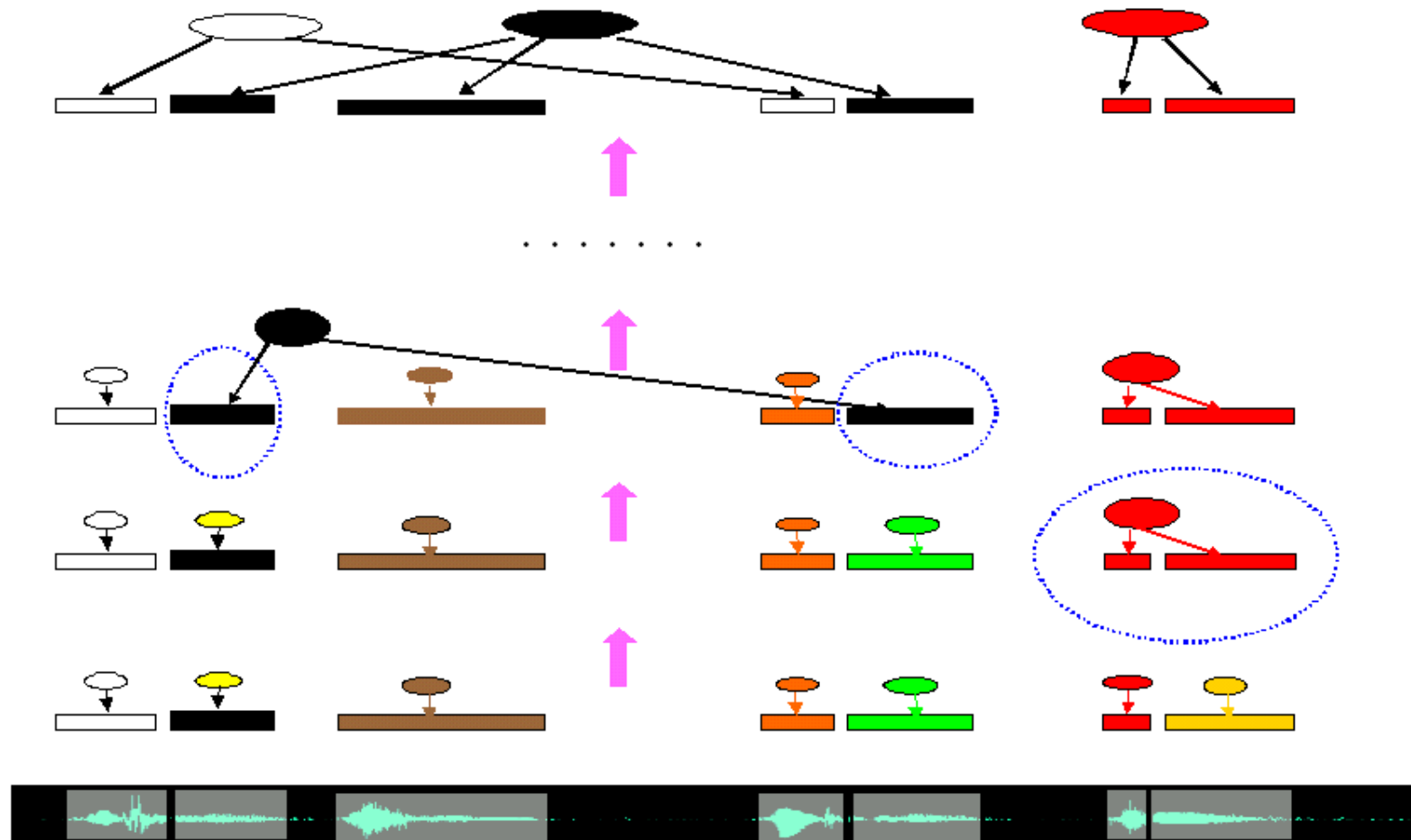
- T-norm: normalizes the score according to the mean and standard deviation of a set of imposter scores

$$S_{tnorm}(X, \lambda) = \frac{\log f'(X|\lambda) - \mu_X}{\sigma_X}$$

- The project work includes review and evaluate the feature warping and T-norm, Z-norm for text-independent speaker verification on NIST evaluation data

# Background – Speaker Clustering

- Hierarchical, agglomerative clustering technique with BIC stopping criteria



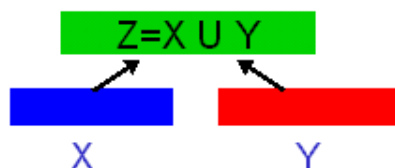
# Improving BIC-based Speaker Clustering

$$BIC(H_0) = \log L(Z, M) - \frac{\lambda}{2} \#(M) \log N$$

$$BIC(H_1) = \log L(X, M_1) + \log L(Y, M_2) - \frac{\lambda}{2} \#(M_1 + M_2) \log N$$

Merge?

$$\Delta BIC = BIC(H_1) - BIC(H_0) = \log \frac{L(X, M_1)L(Y, M_2)}{L(Z, M)} - \frac{\lambda}{2} \#(M) \log N + \frac{\lambda}{2} \#(M_1 + M_2) \log N$$



If  $\Delta BIC > 0$ , Stop clustering

- Lamda usually needs to be tuned in practice
- Remove the tunable Lamda by using different models sizes for modeling

$$\#(M) = \#(M_1 + M_2)$$

$$\begin{aligned} \Delta BIC = BIC(H_1) - BIC(H_0) &= \log \frac{L(X, M_1)L(Y, M_2)}{L(Z, M)} - \frac{\lambda}{2} \#(M) \log N + \frac{\lambda}{2} \#(M_1 + M_2) \log N \\ &= \log \frac{L(X, M_1)L(Y, M_2)}{L(Z, M)} \end{aligned}$$

- The project work includes systematic comparison of standard and improved BIC-based speaker clustering on evaluation dataset

# Other Possible Projects

- Intoxicated Speech
  - (Joe Laws, Yu-Hsiang Chiu & Chanwoo Kim)
- The Dolphins project
  - (Alan Black, Tanja Schultz, Bob Frederking)
- The SPICE Project
  - (Alan Black, Tanja Schultz)

# Towards Communication with Dolphins

- Idea:

The Dolphins project (Alan Black, Tanja Schultz, Bob Frederking) is run in cooperation with the Wild Dolphins Foundation. The long-term goal is to communicate with wild dolphins. Right now we are at the VERY beginning. We have (very noisy) recordings of dolphins but we do not know yet WHAT they say ...

- Wanted:

Project A: Find meaningful 'sound' units

Project B: Indexing the Video Database



# Why communicating with Dolphins?

- Why do we want to talk to Dolphins?
  - They might have a lot to say  
(e.g. Indus river dolphin is similar to species from 20 mio years ago)
  - It is a challenging scientific problem
    - Cross language boundaries → Cross species boundaries
    - Different sound production, perception, frequency range
    - Different medium (water), transmission, omni-directional
    - Nothing is known about dolphins' language
  - It involves spending a lot of time in the Bahamas 😊
- Why do the Dolphins want to talk to us?
  - We don't know ...
  - ... but there is a lot of evidence that they try hard

# Intoxicated Speech

- Question: If and how does speech change if it is spoken under the influence of alcohol
- Given: A database of intoxicated speech, a baseline recognizer for non-intoxicated speech
- Previous Projects: Preliminary work on the database, nicely organized
- Wanted: A speech recognizer (Acoustic Models) trained/adapted on intoxicated speech; comparison to non-intoxicated speech

# The SPICE Project

- SPICE: **S**peech **P**rocessing: **I**nteractive **C**reation and **E**valuation Toolkits for new languages
- To build speech processing systems we need
  - data resources (speech, text, dicts, bilingual corpora)
  - language technology experts
  - 4500-6000 languages
- How to bridge the gap?
- Our Solution: Provide web-based tools that allow native speakers to create and evaluate speech processing components (speech recognition, understanding, translation, synthesis) without language technologies background
- Build on GlobalPhone and FestVox