A Review of Summer Workshop on Innovative Techniques for LVCSR

Aravind Ganapathiraju

Institute for Signal & Information Processing
Mississippi State University
Mississippi State, MS 39762
ganapath@isip.msstate.edu

ISIP Weekly Seminar Series
Fall 1997
September 4, 1997

INTRODUCTION

- Areas of Research
  - Acoustic processing
  - Syllable-based speech recognition
  - Pronunciation modeling
  - Discourse language modeling
- Research at the previous workshops
  - 1995 - Language Modeling workshop
  - 1996 - LVCSR workshop
    - — Speech data modeling (ANN, Multi-band, large context)
    - — Automatic learning of word pronunciations
    - — Hidden speaking mode

ACOUSTIC MODELING

- Goal: Investigate methods that integrate information extracted from various time-scales into the acoustic models.
- Techniques experimented on:
  - Linear discriminant analysis (LDA), Heteroscedastic discriminant analysis (HDA)
  - filtering trajectories of acoustic features
  - investigate different warping functions

FEATURE TRANSFORMATIONS

- LDA - incorrectly assumes equal variances classes, simple Eigen analysis
- HDA - takes care of unequal variance in classes, requires non-linear optimization
- Methods
  - collect class statistics (means and variances of monophones)
  - find feature transformation (LDA or HDA)
  - apply transformation to all data
  - train recognizer with new features
  - a modified EM algorithm used for training

CONCLUSIONS

- LDA - worsened performance by 1%
- HDA - improved performance by 1%, need for a more intelligent training algorithm
- Filtering at different time scales helped on small set of studio quality data, but has not been tested on Switchboard
- “mel” warping seems to a reasonable warping function

PRONUNCIATION MODELING

- ❑ Goal: Model pronunciation variation found in the SWITCHBOARD corpus to improve speech recognition performance
- ❑ Methods
  - ❍ Use hand-labeled phonetic transcriptions as target of modeling
  - ❍ Use dictionary pronunciation, lexical stress and other linguistic information as source of modeling
  - ❍ Use statistical methods to learn the mapping from base forms to the surface forms
  - ❍ Create pronunciation networks to be used as the recognizer's dictionary

MODEL ESTIMATION

- Decision Trees
  - predict phone realizations based on questions concerning baseform context
- Multi-words
  - predict phone realizations based on their frequency of occurrence in pairings with their baseform context
- Unsupervised Learning
  - bootstrap by clustering automatic phone recognition of high frequency words

TRAINING and TEST ISSUES

- Pronunciation Model:
  - cross-word or word-internal
  - should it generalize to unseen contexts
  - should it be word specific
  - should training be on hand-labeled or automatically transcribed data
- Acoustic Model:
  - training on a standard dictionary
  - training on pronunciation realization model

UNSOLVED/FUTURE WORK

- Tree based models
  - effective acoustic retraining
  - improved crossword modeling
- Multi-word models:
  - Derive new multi-words from data
  - Generalize to unseen contexts
- Dynamic pronunciation modeling - use of rate/duration information

DISCOURSE LANGUAGE MODELING

- Goal: Better use of discourse knowledge to improve recognition accuracy
- Understanding spontaneous dialog
  - need to know who said what to whom
- Better human-computer dialog
  - agent needs to know whether you asked it a question or ordered to do something
- First step towards speech understanding
- Can discourse knowledge help improve recognition performance

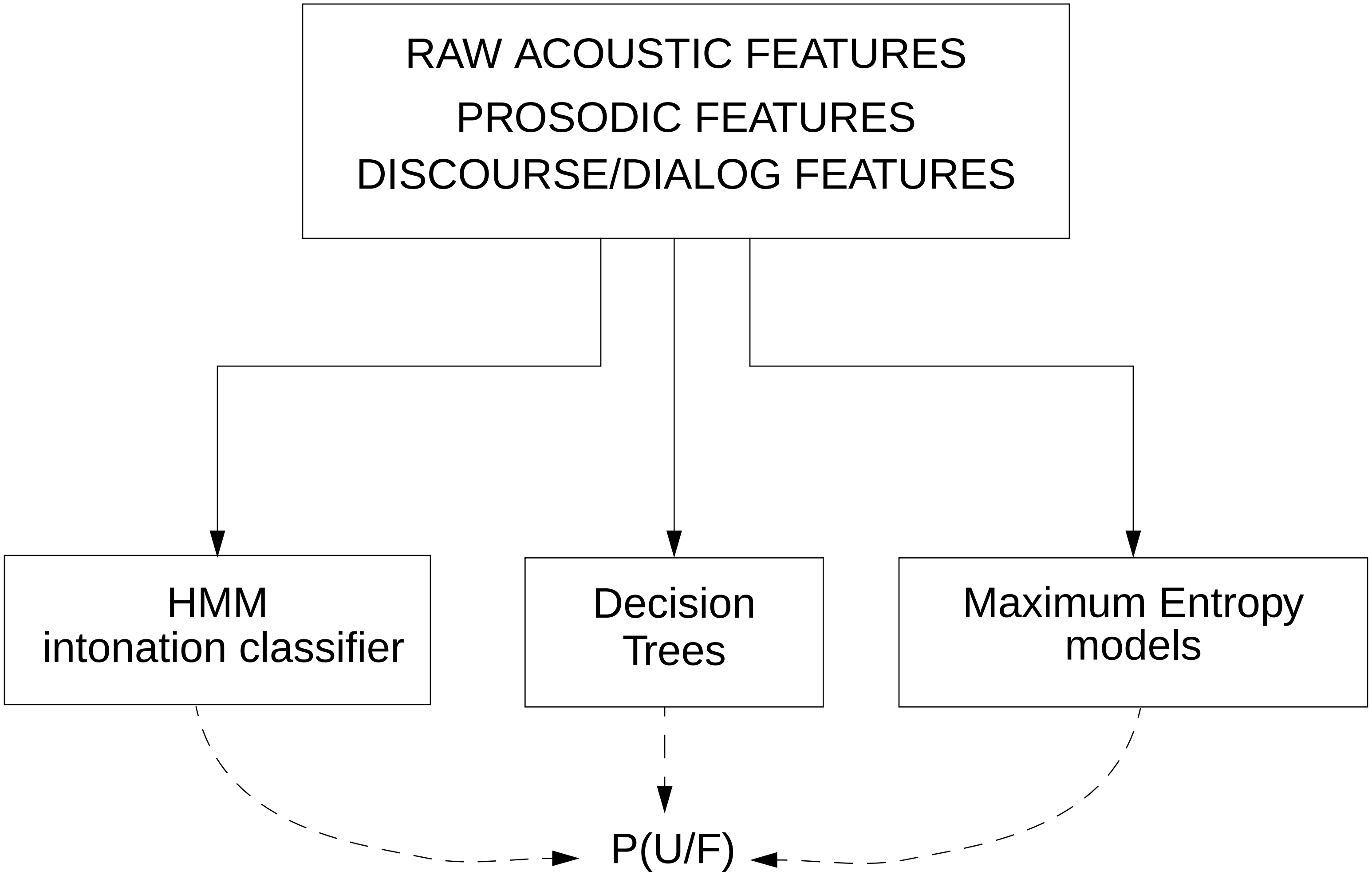
WHY DISCOURSE KNOWLEDGE?

- Word “DO” has an error rate of 72%
- “DO” present in almost every yes-no-question
- If we detect a yes-no-question we could increase P(DO)
- yes-no-question easily detected by rising intonation

UTTERANCE TYPE DETECTION

- Words and word grammar
  - pick the most likely utterance type (UT) given the word string
- Discourse grammar
  - pick the most likely UT given the surrounding utterance types
- Prosodic information
  - pitch contour
  - energy/SNR
  - speaking rate

UTTERANCE TYPE DETECTION



WHAT DID WE LEARN?

- Successful utterance type detection
- First step towards automatic discourse understanding
- Prosodic information is useful for discourse processing
- Only marginal recognition win, why?
    - with complete knowledge of utterance type gain of only 2% over baseline recognizer
    - maximum win in question detection but database primarily statement oriented

SYLLABLE-BASED SPEECH RECOGNITION

- All state-of-the-art LVCSR systems have been predominantly phone based
- Phone is not a very flexible unit for spontaneous speech
- Cannot exploit temporal dependencies when modeling unit's of very short duration
- Syllable is a reasonable alternate
  - Longer time window to better capture contextual effects
  - can be viewed as a stochastic model on top of a collection of phones, thus inherently modeling more variations

SYLLABLES OFFER MORE!

- Stability of a syllable as a recognition unit
    - Insertion and deletion rate of syllable is as low as 1% as compared to 12% for phones
    - Clearly syllable is much more stable
- Longer duration makes it easier to exploit temporal and spectral variations simultaneously (Parameter trajectories, Multi-path HMMs)
- Possibility of compact coverage

WHAT DOES A SYLLABLE SYSTEM COMPARE WITH?

- Only context independent syllables were used
  - context independent phone system is a reasonable lower bound for performance (62.3% WER)
- Comparing with cross-word context dependent phone system not correct since cross-word modeling for syllables not done
- A better upper bound is a word-internal context dependent phone system (49.8% WER)

BASELINE SYLLABLE SYSTEM

- A syllabified lexicon used for syllable definitions
- 9023 syllable seeded for complete coverage of training data
- Syllable durations found from forced alignment
- Number of states in HMM proportional to syllable duration
- Due to under trained models, used only 800 syllables for testing
- Monophones used to fill up the test lexicon
- Performance - 55.1% WER

HYBRID SYLLABLE SYSTEM

- Error analysis of baseline system:
  - errors on words with mixed or all phone representation high
- Suggests mismatch at syllable phone junctions
- 800 syllables and monophones trained together
- Performance - 51.7% WER

OTHER IMPORTANT EXPERIMENTS

- Finite duration modeling
  - long tails for some of the syllable model duration histograms.
  - high word deletion rate
  - both these suggest need for durational constraints on models
  - number of states in model proportional to expected stay
  - performance - 49.9% WER
- Monosyllabic word modeling
  - 75% of training word tokens are monosyllabic
  - 200 monosyllabic words cover 71%
  - monosyllabic words account for 70% of error
  - created separate models for monosyllabic words
  - performance - 49.3%, with finite duration 49.1

MAJOR CONCLUSIONS

- Ofcourse, we proved that syllable models work as well as triphone models, if not better
- Lexical issues need to be addressed
  - a quick post workshop experiment showed a gain of 1% by looking at one particular issue (ambisyllabics)
- We have not explicitly exploited temporal characteristics of syllables
  - parameter trajectories and multi-path HMMs need to be tested
- Context dependent syllable modeling and state tying
  - will involve decision tree clustering

WORKSHOP CONCLUSIONS

- Not much gain in terms of reduction in word error rate
- Pronunciation modeling has been repeatedly shown to be useful
- Generalized discriminant analysis shows promise
- Discourse level information is not explicitly beneficial in improving recognition accuracy
- Decision trees are used successfully in all aspects of speech recognition
- Overall it is sad that there was no breakthrough
- Isn’t that good for us? More things to solve and more time to get there to the top!

WHY WAIT? LETS DO IT FOLKS!!!!