



Speech Processing 15-492/18-492

Speech Recognition

Acoustic modeling

Pronunciation dictionary

Acoustic Modeling

- ◆ *Speech and Signal Variability*
- ◆ *Measuring Error*
- ◆ *Pronunciation lexicons*

Variability in Speech Signal

- ◆ *“Mr Wright should write to Ms Wright right away about his Ford or four door Honda.”*
 - *Homophones: same pronunciation*
 - *“wright” “right” “write” / r ay t /*
 - *“ford or” “four door” / f ao r d ao r /*

Style Variability

- ◆ *Different articulation in different situations*
- ◆ *Clear vs Conversational*
- ◆ *Whisper vs shouting*
- ◆ *Talking to machine, talking to others*
- ◆ *Frustrated speech*

Speaker variability

- ◆ *Gender, age, dialect, health*
- ◆ *Speaker dependent systems*
- ◆ *Speaker independent systems*
- ◆ *Speaker adaptive systems*
 - *Enrolment stage (acoustics and language)*

Environment Variability

- ◆ *Different background noises*
 - *Office vs Outside*
- ◆ *Different applications, different environments*
 - *Desktop dictation, to Warehouse pick*
- ◆ *Single speaker vs Multispeaker*
- ◆ *Background music*

Channel Variability

- ◆ *Telephone vs Desktop*
 - *8KHz vs 16KHz*
- ◆ *PDA vs Desktop*
- ◆ *Close-talking vs far-field*
- ◆ *Cell Phone vs Landline*

Measuring Speech Recognition Error

◆ *Word Error Rate*

- *Substitutions: word is replaced*
- *Deletions: word is missed out*
- *Insertions: word is added*

$$WER = 100\% \times \frac{\text{Subs+Dels+Ins}}{\text{word in correct sentence}}$$

Word Error Rate

- ◆ *WER requires:*
 - *Transcription (the correct word string)*
 - *Alignment between ASR output and Transcript*
 - *Not just left to right matching*
- ◆ *Sometimes Accuracy is given*
 - *100-WER*
 - *NOT number of words correct*

Word Error Rate

- ◆ *Can get > 100%*
 - *But something is very wrong*
- ◆ *Outputting “the” only, ignoring the speech*
 - *Sometimes gives WER < 100%*
- ◆ *All words are treated equal*
 - *“This specimen” vs “The specimen”*
 - *“Is absent” vs “Is present”*

Signal Acquisition

- ◆ *High quality signal quality*
 - *Lower sample rate will increase WER*
 - *8KHz baseline*
 - *16KHz -10%*

End-Point Detection

- ◆ *Long silence will likely increase WER*
 - *It will recognize phantom words*
- ◆ *Need to find the speech in the signal*
 - *VAD (Voice Activity Detection)*
 - *Find beginning and end of speech*
- ◆ *Typically do continuous recognition*
 - *Recognized while listening*
 - *But need end point (have to wait)*

Feature normalization

- ◆ *Sometimes do normalization*
 - *Remove mean from MFCCs*
 - *Can make recognition more reliable in noise*
- ◆ *Often include deltas and delta deltas*
- ◆ *Sometimes to feature reduction*
 - *Principal Component Analysis*

What phones/segments

- ◆ *Need the best set for discrimination*
 - *Not necessary the same as Linguistic Phones*
- ◆ *More phones means more training*
 - *And needs to have consistent Lexicon*
- ◆ *Extra phones*
 - *t vs dx*
 - *t vs nx: /t w eh n t iy/ vs /t w eh nx iy/*
 - *Stops as closures and bursts*
 - *Schwas: ax and ix*
 - *Syllabics: el, em, en*
 - *Accents/Tones: ah1, ah0,*

Context dependency

- ◆ *Care about the contexts of each phone*
 - *Post vocalic /r/ and /n/ /m/ affect vowel*
 - *Utterances start and end affect phonemes*
- ◆ *Need more than simple phone models*

Tri-phone Models

- ◆ *Have models for each phone and context*
 - 43^3 contexts about 80K models
- ◆ *Not all contexts have enough examples*
 - *oy (oy) oy* very rare
 - *sh (ax) n* very common
- ◆ *Merge tri-phones that are similar*
 - *E.g t(ih)n with d(ih)n*

Find phones to merge

- ◆ *Using phonetic features*
 - *Most similar feature, most similar acoustics*
 - *Stops, voicing, vowel type ...*
- ◆ *Usually automatic cluster of triphones*
 - *Using CART trees indexed by phonetic features*

Adaptation

- ◆ *Change behavior after use*
- ◆ *Human adaptation*
 - *They will change how they speak*
- ◆ *Channel adaptation*
 - *Cepstral Normalization*
- ◆ *Model adaptation*
 - *Move the means (or weights on means)*

Adaptation

- ◆ *Assume recognition is correct*
 - *(Maybe with some threshold)*
- ◆ *Modify model to make answer more correct*
 - *Adaptation to speaker characteristics*
 - *Adaptation to speaker style*
 - *Can improve accuracy by a few %*

Pronunciation lexicon

- ◆ *Need list of words and their pronunciation*
 - *Pencil p eh n s ih l*
 - *Two t uw*
 - *Too t uw*
 - ...
- ◆ *Need pronunciation of ALL words*

What's a word

- ◆ *Basic words are clear*
- ◆ *What about morphological variants*
 - *walk, walks, walked, walking*
- ◆ *Multi-word words*
 - *Los Angeles, New York*
- ◆ *Contractions*
 - *Wanna, gonna ...*
- ◆ *Yes ALL words that you will recognize*

Pronunciation variants

◆ *Homographs: (same writing different pronunciation)*

- *bass: / b ae s / (fish) / b ey s / (music)*
- *project: N / p r aa jh eh k t / V /p r ax jh eh k t /*

◆ *Natural variants*

- *route: / r uw t / and / r aw t /*
- *coupon: / k uw p ao n / and / k y uw p ao n /*
- *water: / w ao t er / and / w ao dx er /*

CMU Pronunciation Dict

- ◆ *Free pronunciation lexicon*
- ◆ *American English*
- ◆ *Over 100K words*
- ◆ *Not always consistent*
- ◆ *Words for your application will be missing*
 - *We can never get a complete lexicon*

Pronunciation of Unknown Words

- ◆ *Build statistical model from lexicon*
 - *Predict pronunciation from letters*
 - *(Humans do this when they see a new word)*
- ◆ *Typically about 70-85% correct for new words*
 - *Should always check domain words*

Modeling Variability

- ◆ *In Gaussians (in HMM state)*
 - *Multiple mixtures*
- ◆ *In HMM topology*
 - *Number of states and connectivity*
- ◆ *In State Tying*
 - *Sharing Gaussians between states*
- ◆ *In Phone choice*
 - *More/less phones*
- ◆ *In Lexical Pronunciation*
 - *Multiple lexical entries*

Summary

- ◆ *Acoustic modeling*
- ◆ *Word Error Rate/Accuracy*
- ◆ *Lexical pronunciation*

Reading

- ◆ *Section 8.2 Definition of Hidden Markov Model pp 380-393*
- ◆ *Section 8.4 Practical Issues in using HMMS pp 398-405*
- ◆ *In Huang et al.*
- ◆ *Two page description of the contents emailed to dhuggins@cs.cmu.edu before 3:30pm Monday 15th September*

