# Phonetic Modeling in ASR

Chuck Wooters

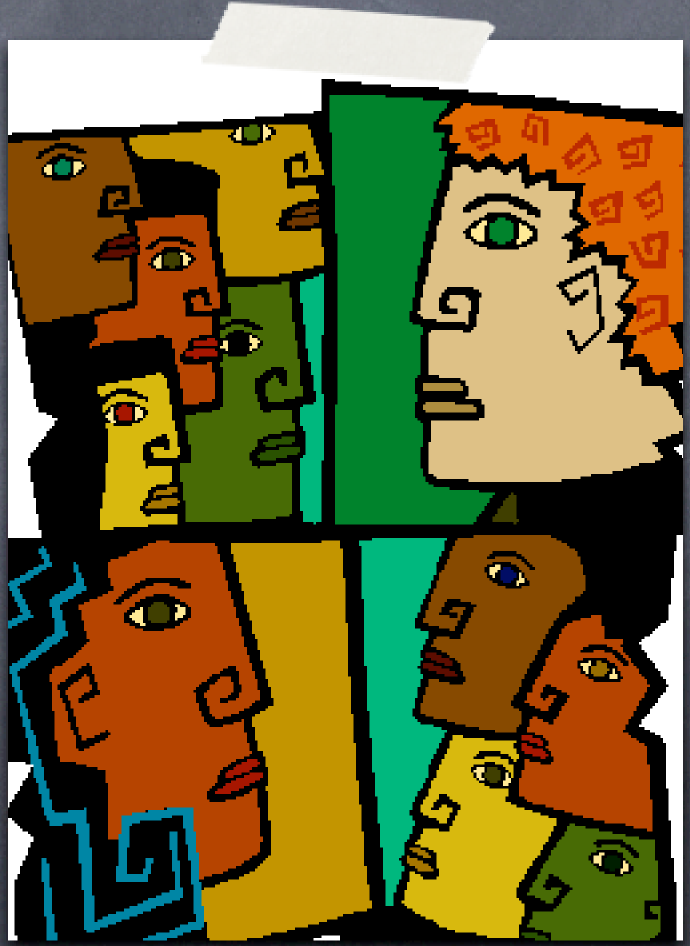
3/16/05

EECS 225d

# Introduction

## VARIATION

- The central issue in Automatic Speech Recognition



# Many Types of Variation

- channel/microphone type
- environmental noise
- speaking style
- vocal anatomy
- gender
- accent
- health
- etc.

# Focus Today

"You say pot[ey]to, I say pot[a]to..."

- How can we model variation in pronunciation?

# Pronunciation Variation

- A careful transcription of conversational speech by trained linguists has revealed...

# 80 Ways To Say “and”

| N | phonetic transcription | | | | | N | Phonetic Transcription | | | | | N | Phonetic Transcription | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 82 | ae | n | | | | 2 | ax | nx | | | | 1 | ix | dcl | d | | | |
| 63 | eh | n | | | | 2 | q | ae | ae | n | d | 1 | ae | eh | n | | | |
| 45 | ix | n | | | | 2 | q | ix | n | | | 1 | hh | n | | | | |
| 35 | ax | n | | | | 2 | ix | n | dcl | d | | 1 | ix | n | t | | | |
| 34 | en | | | | | 2 | ih | | | | | 1 | ae | ax | n | dcl | d | |
| 30 | n | | | | | 2 | eh | eh | n | | | 1 | iy | eh | n | | | |
| 20 | ae | n | dcl | d | | 2 | q | eh | nx | | | 1 | m | | | | | |
| 17 | ih | n | | | | 2 | ix | d | n | | | 1 | ae | ae | n | d | | |
| 17 | q | ae | n | | | 1 | eh | m | | | | 1 | nx | | | | | |
| 11 | ae | n | d | | | 1 | ax | n | dcl | d | | 1 | q | ae | ae | n | | |
| 7 | q | eh | n | | | 1 | aw | n | | | | 1 | q | ae | ae | n | dcl | d |
| 7 | ae | nx | | | | 1 | ae | q | | | | 1 | q | ae | eh | n | dcl | d |
| 6 | ae | ae | n | | | 1 | eh | dcl | | | | 1 | q | ae | ih | n | | |
| 6 | ah | n | | | | 1 | ah | nx | | | | 1 | aa | n | | | | |
| 5 | eh | nx | | | | 1 | ae | n | t | | | 1 | q | ae | n | d | | |
| 4 | uh | n | | | | 1 | eh | d | | | | 1 | ? | nx | | | | |
| 4 | ix | nx | | | | 1 | ah | n | dcl | d | | 1 | q | ae | n | q | | |
| 4 | q | ae | n | dcl | d | 1 | ey | ih | n | dcl | d | 1 | eh | n | m | | | |
| 3 | eh | n | d | | | 1 | ae | ix | n | | | 1 | q | eh | en | dcl | | |
| 3 | q | ae | nx | | | 1 | ae | nx | ax | | | 1 | eh | ng | | | | |
| 3 | eh | | | | | 1 | ax | ng | | | | 1 | q | eh | n | q | | |
| 2 | ae | n | dcl | | | 1 | ay | n | | | | 1 | em | | | | | |
| 2 | ae | | | | | 1 | ih | ah | n | d | | 1 | q | eh | ow | m | | |
| 2 | ax | m | | | | 1 | ae | hh | | | | 1 | q | ih | n | | | |
| 2 | ax | n | d | | | 1 | ih | ng | | | | 1 | q | ix | en | | | |
| 2 | ae | eh | n | dcl | d | 1 | ix | | | | | 1 | er | | | | | |
| 2 | eh | n | dcl | d | | 1 | ae | n | d | dcl | | | | | | | | |

From “SPEAKING IN SHORTHAND - A SYLLABLE-CENTRIC PERSPECTIVE FOR UNDERSTANDING PRONUNCIATION VARIATION” by Steve Greenberg

# Outline

- Phonetic Modeling
- Sub-Word models
  - Phones (mono-, bi-, di- and triphones)
  - Syllables
  - Data-driven units
  - Cross-word modeling
- Whole-word models
- Lexicons (Dictionaries) for ASR

ICSI

# Phonetic Modeling

# Phonetic Modeling

- How do we select the basic units for recognition?
  - Units should be <u>accurate</u>
  - Units should be <u>trainable</u>
  - Units should be <u>generalizable</u>
- We often have to balance these against each other.

# Sub-Word Models

# Sub-Word Models

- Phones
  - Context Independent
  - Context Dependent
- Syllables
- Data-driven units
- Cross-word modeling

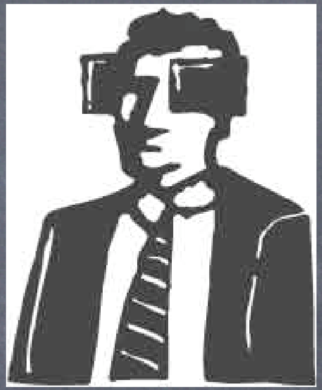
# Phones

# Phones

- Note: “phones” != “phonemes” (see G&M pg. 310)
- E.g.:

| Phoneme | Phone |
|---|---|
| Ascii-65 | AAAAA |

# “Flavors” of Phones



- Context Independent:
  - Monophones

---

- Context Dependent:
  - Biphones
  - Diphones
  - Triphones

# Context Independent Phones

# Context Independent "Monophones"

"cat" = [k ae t]

- Easy to train:
  - only about 40 monophones for English
- The basis of other sub-word units
- Easy to add new pronunciations to lexicon

# Typical English Phone Set

| Phone | Example | Phone | Example | Phone | Example |
|---|---|---|---|---|---|
| iy | FEEL | ih | FILL | ae | GAS |
| aa | FATHER | ah | BUD | ao | CAUGHT |
| ay | BITE | ax | COMPLY | ey | DAY |
| eh | TEN | er | TURN | ow | TONE |
| aw | HOW | oy | COIN | uh | BOOK |
| uw | TOOL | b | BIG | p | PIG |
| d | DIG | t | SAT | g | GUT |
| k | CUT | f | FORK | v | VAT |
| s | SIT | z | ZAP | th | THIN |
| dh | THEN | sh | SHE | zh | GENRE |
| l | LID | r | RED | y | YACHT |
| w | WITH | hh | HELP | m | MAT |
| n | NO | ng | SING | ch | CHIN |
| jh | EDGE | | | | |

Adapted from "Spoken Language Processing" by Xuedong Huang, et. al.

# Monophones

Major Drawback

- Not very powerful for modeling variation:
  - Example: “key” vs “coo”

# Context Dependent Phones

# Biphones

- Taking into account the context (what sounds are to the right or left) in which the phone occurs.
- Left biphone of [ae] in “cat”: k_ae
- Right biphone of [ae] in “cat”: ae_t

“key” = k_iy iy_#
“coo” = k_uw uw_#

# Biphones

- More difficult to train than monophones:
  - Roughly (40^2 + 40^2) biphones for English
  - If not enough training for a biphone model, can "backoff" to monophone

# Triphones

- Consider the sounds to the left AND right
- Good modeling of variation
- Most widely used in ASR systems

"key" = #_k_iy k_iy_#
"coo" = #_k_uw k_uw_#

# Triphones

- Can be difficult to train:
  - there are LOTS of possible triphones (roughly 40^3)
  - Not all occur
  - If not enough data to train a triphone, typically back-off to left or right biphone

# Triphones

- Don't always capture variation:

  "that rock" vs. "theatrical"

  ae_t_r ae_t_r

- Sometimes helps to cluster similar triphones

# Diphones

- Modeling the transitions between phones
- Extend from middle of one phone to the middle of the next

"key" = #_k k_iy iy_#
"coo" = #_k k_uw uw_#

# Syllables

# Syllables

Syllable

Rime

[Onset] Nucleus [Coda]

str eh ng th s

"Strengths"

# Syllables

- Good modeling of variation
  - Somewhere between triphones and whole-word models
- Can be difficult to train (like triphones)
- Practical experiments have not shown improvements over triphone-based systems.

# Data-driven Sub-Word Units

# Data-driven Sub-Word Units

- Basic Idea:
  - More accurate modeling of acoustic variation
  - Cluster data into homogeneous "groups"
    - sounds with similar acoustics should group together
  - Use these automatically-derived units instead of linguistically-based sub-word units

# Data-driven Sub-Word Units

- Difficulties:
  - Can have problems with training, depending on number of units
  - Real problem: generalizability
    - How do we add words to the system when we don't know what the units "mean"
    - Create a mapping from phones?

# Cross-word Modeling

# Cross-word Modeling

- Co-articulation spans word boundaries:
  - "Did you eat yet?" -> jeatyet
  - "could you" -> couldja
  - "I don't know" -> idunno
- We can achieve better modeling by looking across word boundaries
- More difficult to implement- what would dictionary look like?
  - Usually use lattices when doing cross-word modeling

# Whole-word Models

# Whole-word Models

- In some sense, the most "natural" unit
- Good modeling of coarticulation within the word
- If context dependent, good modeling across words
- Good when vocabulary is small e.g. digits:
  - 10 words
  - Context dependent: 10x10x10 = 1000 models
  - Not a huge problem for training

# Whole-word Models

- Problems:
  - difficult to train: needs lots of examples of *every* word
  - not generalizable: adding new words requires more data collection

# Lexicons

# Lexicons for ASR

- Contains:
  - words
  - pronunciations
  - optionally:
    - alternate pronunciations
    - pronunciation probabilities
- No definitions

cat: k ae t
key: k ey
coo: k uw
the: 0.6 dh iy
     0.4 dh ax

# Lexicon Generation

- Where do lexical entries come from?
  - Hand labeling
  - Rule generated
- Not too bad for English, but can be a big expense when building a recognizer for a new language
- For a small task, may want to consider whole-word models to bypass lexicon gen