



# Speech Processing 15-492/18-492

---

Multilinguality

# Dealing with *\*all\** Languages

- ◆ *Over 6000 Languages*
  - *Maybe not all commercially interesting ... now*
- ◆ *Major languages (economic)*
  - *Cell phone manufacturers list 46 languages*
  - *But even those not all covered*

# What you need

## ◆ ASR

- *Acoustic model (lots of speakers)*
- *Pronunciation Lexicon*
- *Language model*

## ◆ TTS

- *Acoustic model (one speaker)*
- *Pronunciation Lexicon*
- *Text analysis*

# Writing Systems

- ◆ *Romanized writing systems*
  - *Latin-1 (iso-8599-1)*
  - *Covers many Western Europeans languages*
- ◆ *Cyrillic*
  - *Covers many Eastern European Languages*
- ◆ *Arabic Scripts*
  - *Arabic(s), Farsi, Urdu, etc*
- ◆ *Devenagari*
  - *Covers many Northern India Languages*
- ◆ *Chinese Hanzi*
  - *Covers some Chinese dialects but different versions*
- ◆ *Many other scripts some non-standard*

# Writing Systems

- ◆ *Letter based*
  - *Latin, Cyrillic*
- ◆ *Consonant based*
  - *Arabic, Hebrew*
- ◆ *Mora based*
  - *Half syllable or syllable*
  - *Indian scripts, Japanese native scripts*
- ◆ *Syllable based*
  - *Hangul, Chinese*

# Standards

- ◆ *Writing standards*
  - *Taught at schools, newspapers, computer support*
  - *Typically standardized spelling*
- ◆ *May be mostly spoken*
  - *Occasionally written*

# Language Specific Issues

- ◆ *No explicit markings*
  - *Stress, accent, tones*
- ◆ *No word boundaries*
  - *Chinese, Thai*
- ◆ *No (short) vowels*
  - *Arabic, Hebrew*
- ◆ *Rich morphology*
  - *Many different words in the languages*
  - *Finnish, Turkish, Greenlandic*

# Genre Specific Issues

- ◆ *No capitals, punctuations*
- ◆ *Unpunctuated*
- ◆ *Plain vs polite form*
- ◆ *Speech vs text form*
- ◆ *Many foreign phrases*
  - *(technology directed genre's)*
- ◆ *Many new abbreviations*
  - *E.g. SMS messages*

# Character Encoding

- ◆ *Unicode vs utf8 vs latin*
  - *Documents mix them*
- ◆ *Sometime accent omitted*
  - *For ease of typing*
- ◆ *Lots of standards*
  - *Unicode, EUC, BIG5, TIS42, ...*
  - *Everyone has their own standard*
- ◆ *Some create their own standards*
- ◆ *Mixed character sets*

# Phoneme Sets

- ◆ *Hard to find consensus for new languages*
  - *Typically lots of different dialects*
- ◆ *What level of distinction?*
  - *Some good for speech but not really phonetic*
  - */t/ vs /dx/ in “water”*
- ◆ *Often doesn't include foreign phones*
  - */w/ in German is common for younger people*

# Words

- ◆ *May be hard to define*
  - *No word boundaries*
- ◆ *Rich morphology*
  - *Words have many variations of compounds*
  - *Yomenakatta -> could not read*
  - *Yomemasendeshita -> could not read (polite)*
- ◆ *Gender specific speech*
  - *Boku vs atashi*
- ◆ *Language mixtures*

# Pronunciation lexicons

- ◆ *“proper” speech vs “actual” speech*
- ◆ *Hard to generalize*
  - *Chinese*
- ◆ *Cross lingual pronunciations*
  - *“Human” (English/German)*

# “Industry” way

- ◆ *Collect at least 100 hours of spoken speech*
  - *At least 20 different speakers*
  - *Mixture of gender, age, etc*
  - *Through desired channel (phone/desktop)*
- ◆ *Collect at least 5 hours from one speaker*
  - *High quality recording studio*
- ◆ *Data should be targeted to application*
- ◆ *Build pronunciation lexicon*
  - *Expert phonologist*

# Industry way

- ◆ *Probably 3-6 months*
  - *Lead developer*
  - *Local language expert*
  - *Lots of human transcribers*
- ◆ *Costs?*
  - *Many hundreds of thousands*

# Or cheaper (?) ...

- ◆ *Find existing data*
  - *Linguistic Data Consortium (UPenn)*
  - *ELRA (European equivalent)*
  - *Appen, Australia*
  - *Find local people who have collected data*
- ◆ *Found data might be in wrong format*
  - *Data cleaning is often the most expensive*

# Actual way

- ◆ *Often mixture*
  - *Found data for initial model*
  - *Collect data with actual/initial application*

# Multilingual Systems

- ◆ *Support lots of different languages*
  - *Press 1 for Spanish*
  - *Press 2 for Gujarati ...*
- ◆ *Automatically detect language*
- ◆ *Mixed language*

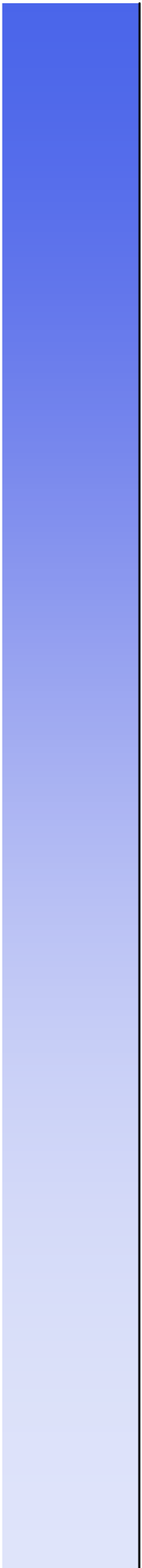
# Multilingual (Menu)

- ◆ *Speak in your language*
  - *Eki-mai no tsugi no bus no ha?*
  - *When is the next bus to the station*
- ◆ *Need multiple recognizers*
  - *Run in parallel and take best result*
- ◆ *Or shared acoustic models*
  - *Recognizing both languages at once (mix)*

# Multilingual (in line)

- ◆ *Code switching*
  - *European, India, Bilingual areas*
  - *Hinglish, Spanglish*
- ◆ *Borrowed words and phrases*
  - *Dad, time kyu hua hai*
  - *One lakh*
  - *Computer walla*
  - *numbers*
- ◆ *Can be inflected*
  - *Was updated -> up gedaten*

Lilac



# HW2: TTS

- ◆ *Due 3:30pm Monday October 20<sup>th</sup>*
- ◆ *Install Festival and Festvox*
- ◆ *Find 10 errors in each of two different synthesizers*
- ◆ *Build a voice*
  - *A Talking Clock*
  - *A general voice*
  - *(or both)*

