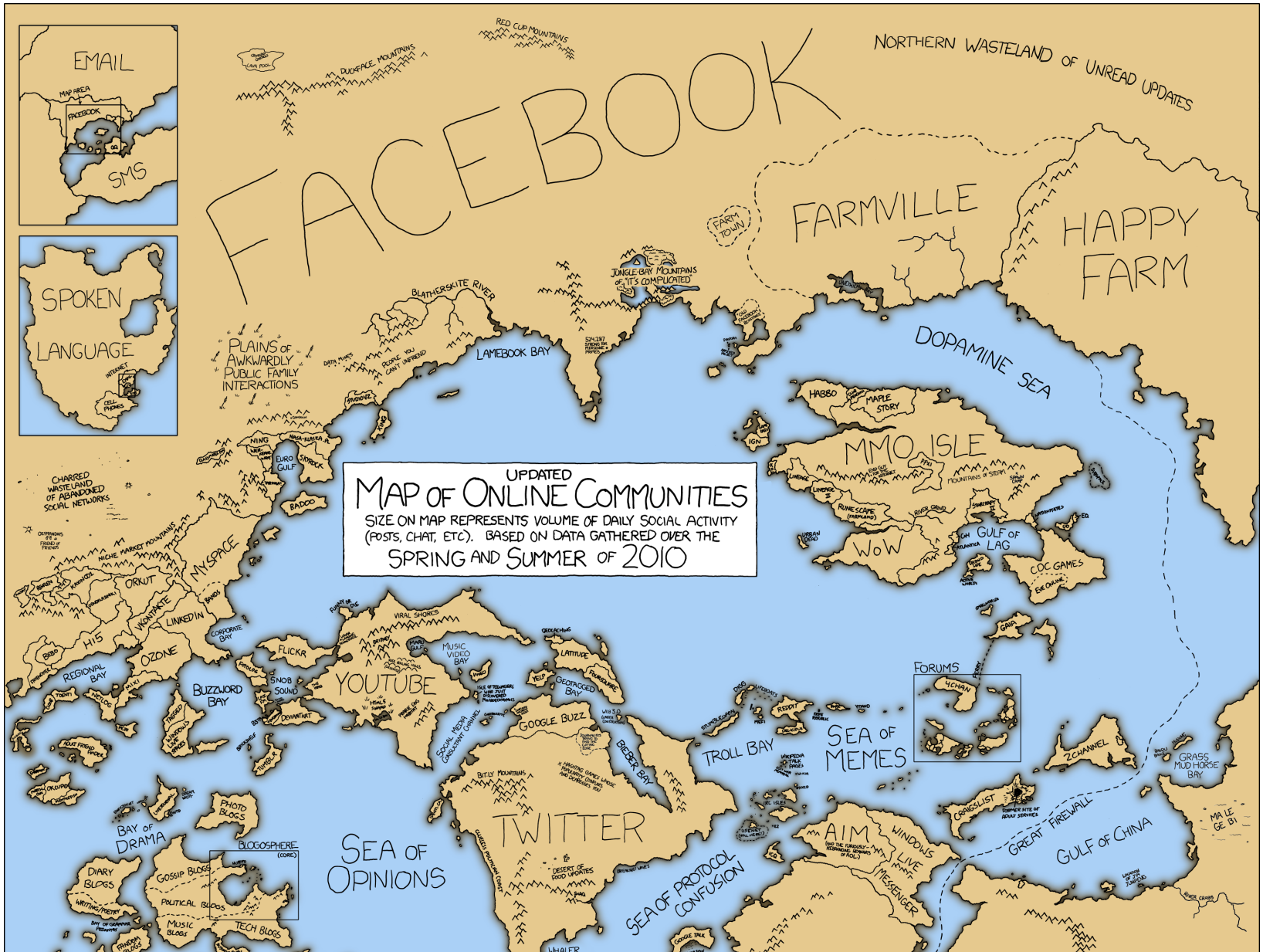
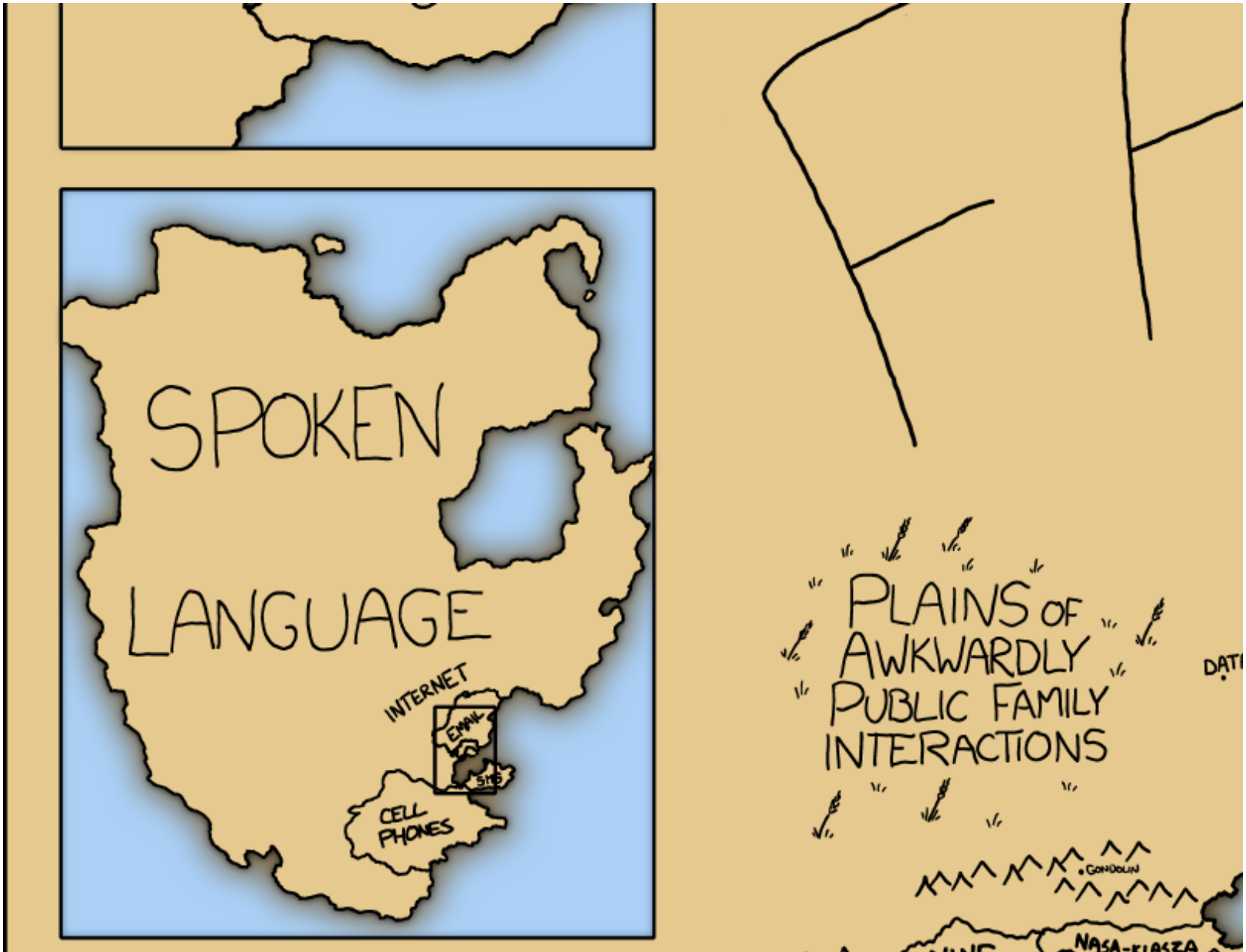
How To Wreck a Nice Beach

Linguistics 409 · Computational Linguistics

Relevant XKCD (802)



Relevant XKCD (802)



What is ASR?

Speech-to-Text Transcription

- Transform recorded audio into a sequence of words
- Just the words, no meaning...

What is ASR?

Speech-to-Text Transcription

- Transform recorded audio into a sequence of words
- Just the words, no meaning...
- But: “Will the new display recognise speech?” vs “Will the nudist play wreck a nice beach?”
- Paralinguistic aspects: how did they say it? (timing, intonation, voice quality)

What isn't ASR?

Speech-to-Text Transcription

- Speaker diarization: Who spoke when?
- Speech recognition: what did they say?
- Language identification: what language are they speaking?
- Automatic Speech Understanding

What is ASR?

- How would automated speech recognition be useful?
- What applications have you used?
- What are some potential applications?

What is ASR?

- How would automated speech recognition be useful?
- What applications have you used?
- What are some potential applications?

What is ASR?

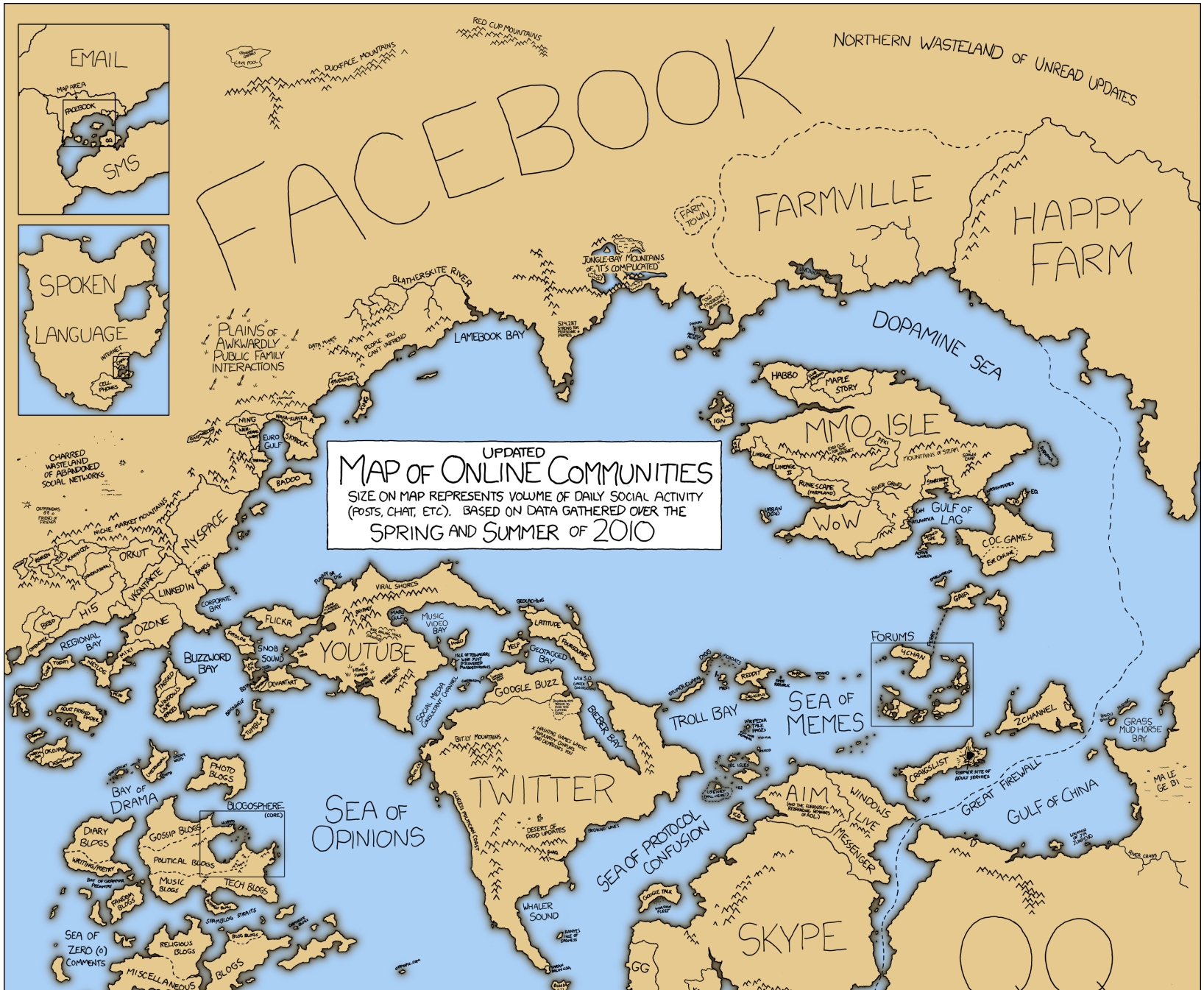
- How would automated speech recognition be useful?
- What applications have you used?
- What are some potential applications?

Why is ASR so difficult?

- What makes ASR so difficult?
  - Fish tale
  - Sinewave speech
  - Indexical/socioindexical variation
  - Intra-speaker variation
  - Context-appropriate variation
  - Allophonic variation
  - but there is something deeper that makes ASR a strange task...

Why is ASR so difficult?

- What makes ASR so difficult?
  - Fish tale
  - Sinewave speech
  - Indexical/socioindexical variation
  - Intra-speaker variation
  - Context-appropriate variation
  - Allophonic variation
  - but there is something deeper that makes ASR a strange task...


EMAIL
MAP AREA
FACEBOOK
SMS
SPOKEN LANGUAGE
INTERNET
CELL PHONES
FACEBOOK
NORTHERN WASTELAND OF UNREAD UPDATES
FARMVILLE
FARM TOWN
HAPPY FARM
DUCKFACE MOUNTAINS
RED CUP MOUNTAINS
JUNGLE-BAY MOUNTAINS OF "IT'S COMPLICATED"
BLATHERSKITE RIVER
LAMEBOOK BAY
PLAINS OF AWKWARDLY PUBLIC FAMILY INTERACTIONS
DOPAMINE SEA
UPDATED
MAP OF ONLINE COMMUNITIES
SIZE ON MAP REPRESENTS VOLUME OF DAILY SOCIAL ACTIVITY (POSTS, CHAT, ETC). BASED ON DATA GATHERED OVER THE SPRING AND SUMMER OF 2010
MMO ISLE
HABBO
MAPLE STORY
LINEAGE
LINEAGE II
RUNESCAPE (FREELAND)
WoW
GULF OF LAG
CHARRED WASTELAND OF ABANDONED SOCIAL NETWORKS
NICHE MARKET MOUNTAINS
NING
NINJA-KUNERA
EURO GULF
SPYBUCK
BADOO
MYSPACE
ORKUT
HI5
XANGA
LINKEDIN
CORPORATE BAY
REGIONAL BAY
OZONE
BUZZWORD BAY
FLICKR
SNOB SOUND
DEVIANTART
YOUTUBE
VIRAL SHORES
MUSIC VIDEO BAY
GEOTAGGED BAY
LATITUDE
YELP
FOURSQUARE
GOOGLE BUZZ
BIEBER BAY
TWITTER
BITLY MOUNTAINS
DESERT OF FOOD UPDATES
WHALER SOUND
SEA OF OPINIONS
BAY OF DRAMA
PHOTO BLOGS
BLOGOSPHERE (CORE)
DIARY BLOGS
GOSSIP BLOGS
POLITICAL BLOGS
TECH BLOGS
MUSIC BLOGS
WRITING/POETRY
SPAM BLOGS
SPAM BLOG STRAITS
RELIGIOUS BLOGS
MISCELLANEOUS BLOGS
SEA OF ZERO (0) COMMENTS
FORUMS
4CHAN
REDDIT
TROLL BAY
SEA OF MEMES
Z-CHANNEL
GRASS MUD HORSE BAY
CRAIGSLIST
AIM
WINDOWS LIVE MESSENGER
GREAT FIREWALL
GULF OF CHINA
SEA OF PROTOCOL CONFUSION
SKYPE
GG
COC GAMES
GAIA
MA LE GE BI

W

Why is ASR so difficult?

Sources generally recognized within CS/NLP:

Size Number of word types in vocabulary, perplexity

Speaker Tuned for a particular speaker, or speaker-independent? Adaptation to F0, formants, indexical properties, socioindexical properties, but also unigram, bigram and trigram probabilities, habitual prosodic patterns, etc.

Environment Acoustic environment includes noise, competing talkers, and so-called channel conditions (microphone, phone connection, room acoustics, etc.)

Style Continuously spoken or isolated? Planned monologue or spontaneous conversation? Human to human speech or Human to (idiot) computer speech?

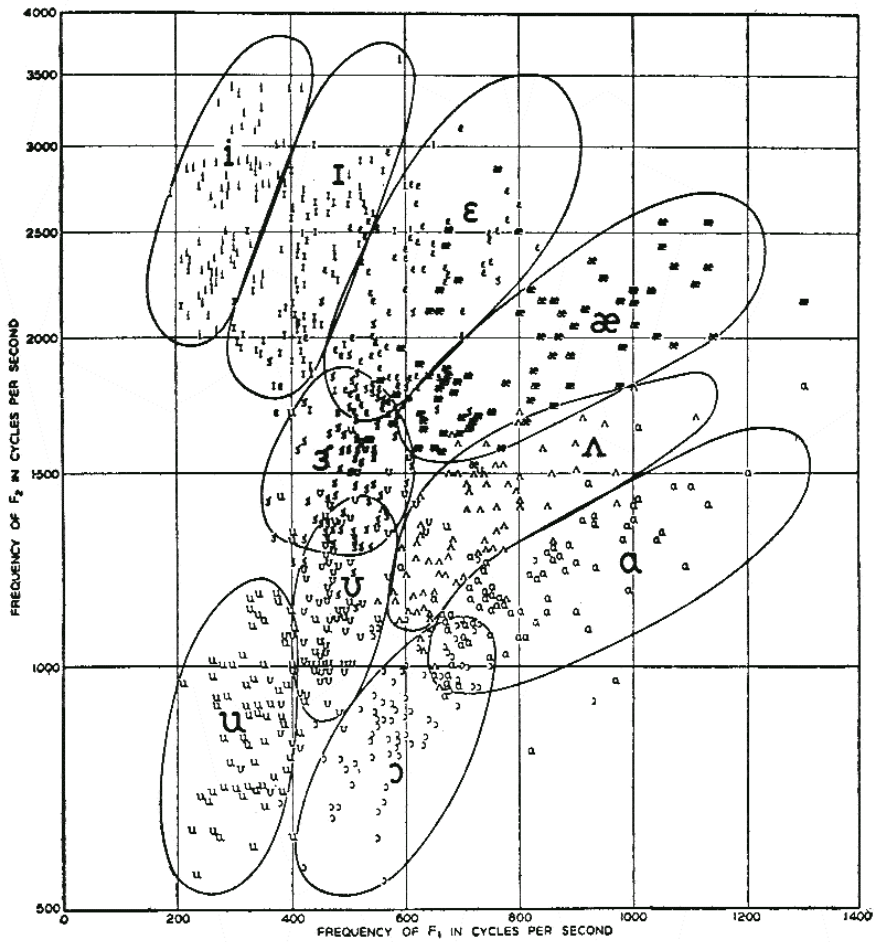
Theory follows engineering here...

- Speech perception is one of the few fields in mainstream linguistics that has really embraced the findings of computational linguistics.
- ASR researchers realized long ago that intense effort is needed to derive and encode linguistic rules that accurately model speech
- And even then the model is not very good...

Famous Quote attributed to Fred Jelinek circa 1988

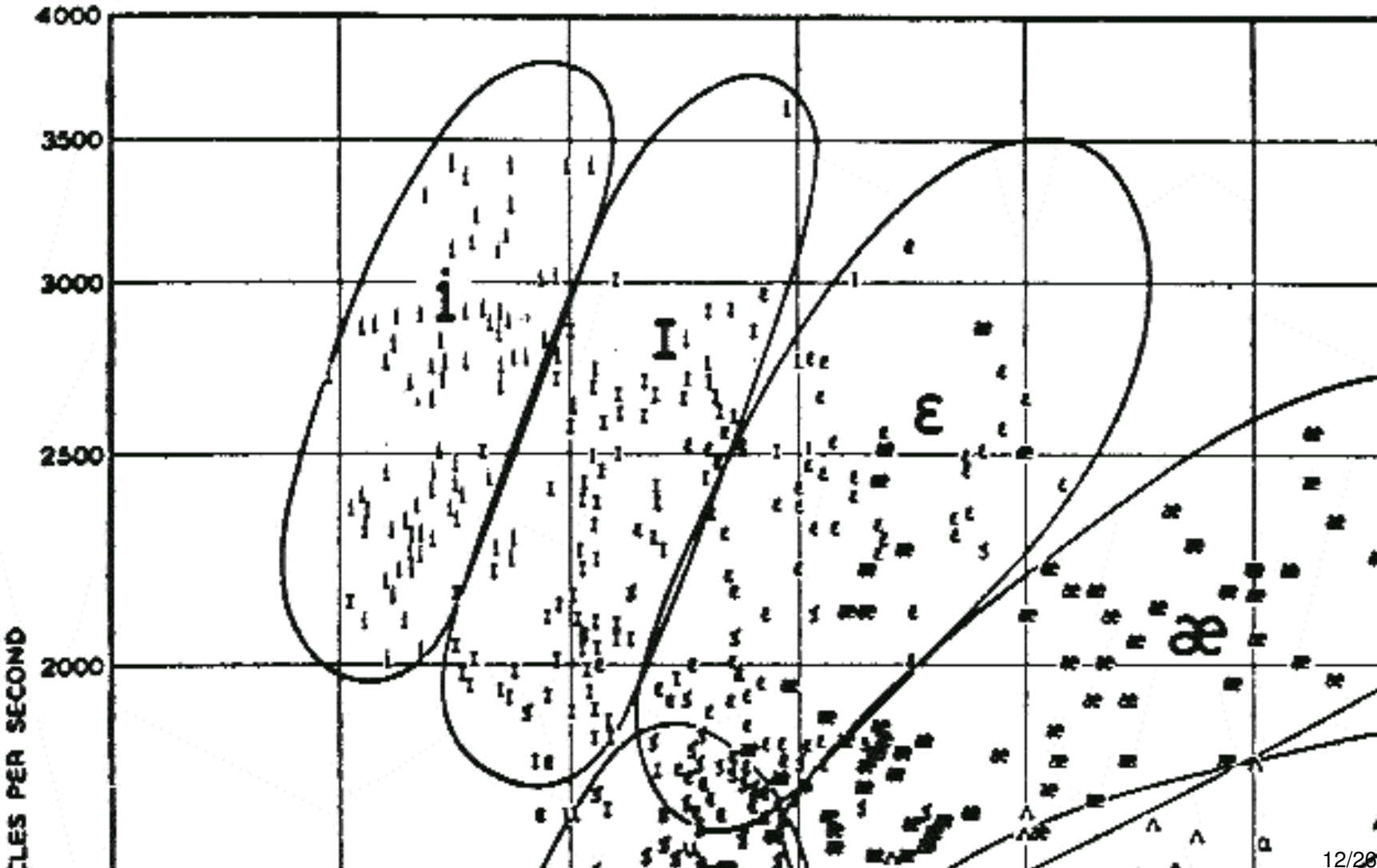
"Every time I fire a linguist,
the performance of the speech recognizer goes up"

Peterson & Barney 1952



Frequency of second formant *versus* frequency of first formant for ten vowels by 76 speakers.

Peterson & Barney 1952



Theory follows engineering here...

- It is very difficult to take account of the variability of spoken language with, say, derivational phonology (which is not to say that people didn't try!)
- Data-driven machine learning approach: Construct simple models of speech which can be learned from large amounts of data (thousands of hours of speech recordings)
- See, for example, exemplar theories of speech perception.

Fundamental equations of Statistical Speech Recognition

If $\mathbf{X}$ is the sequence of acoustic feature vectors (observations) and $\mathbf{W}$ denotes a word sequence, the most likely word sequence $\mathbf{W}^*$ is given by

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} P(\mathbf{W} \mid \mathbf{X})$$

Applying Bayes' Theorem:

$$\begin{aligned} P(\mathbf{W} \mid \mathbf{X}) &= \frac{p(\mathbf{X} \mid \mathbf{W}) P(\mathbf{W})}{p(\mathbf{X})} \\ &\propto p(\mathbf{X} \mid \mathbf{W}) P(\mathbf{W}) \\ \mathbf{W}^* &= \arg\max_{\mathbf{W}} \underbrace{p(\mathbf{X} \mid \mathbf{W})}_{\substack{\text{Acoustic} \\ \text{model}}} \quad \underbrace{P(\mathbf{W})}_{\substack{\text{Language} \\ \text{model}}} \end{aligned}$$

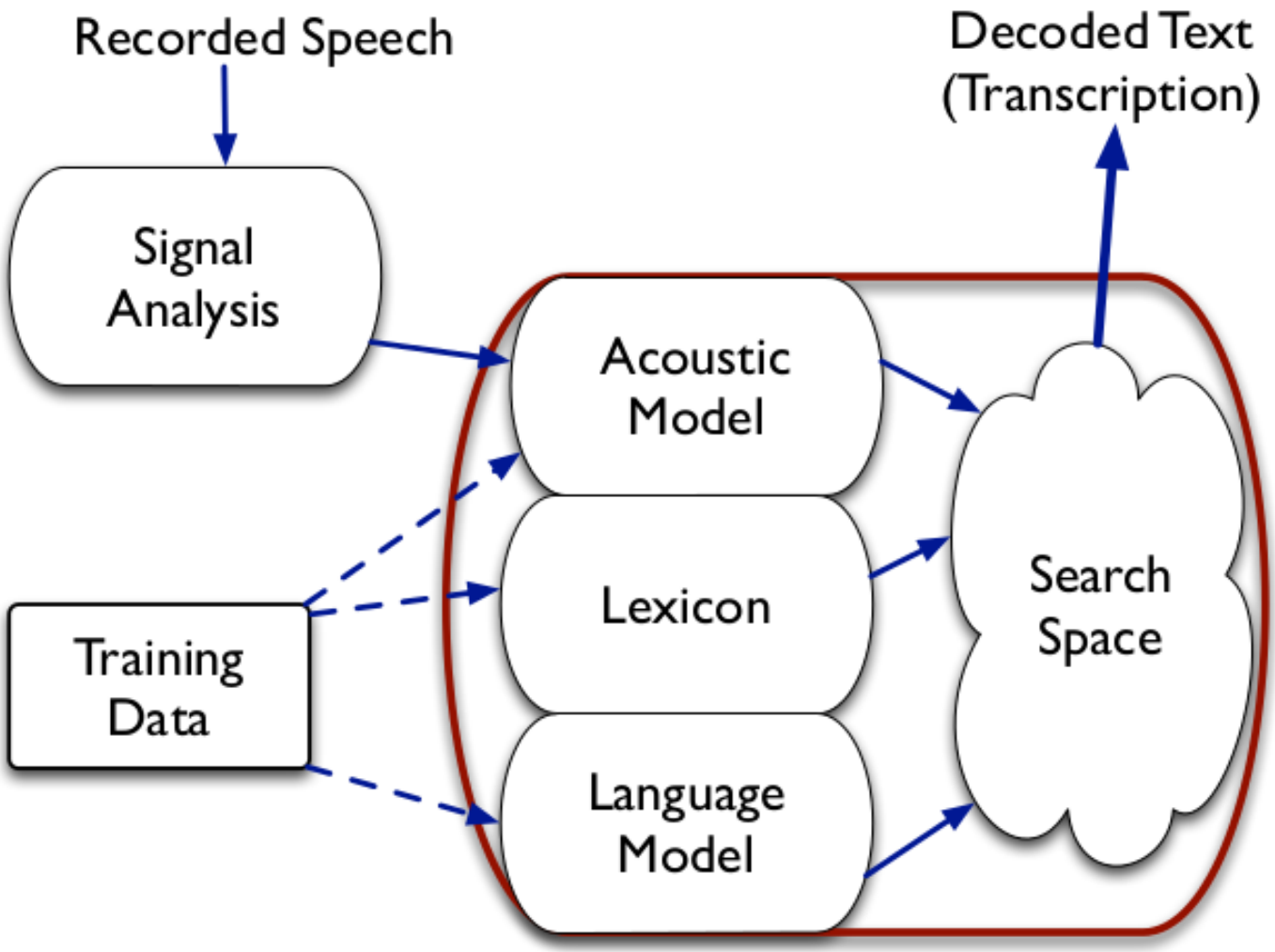
Automatic speech recognition

From the license agreement that comes with Nuance's Dragon Dictate:

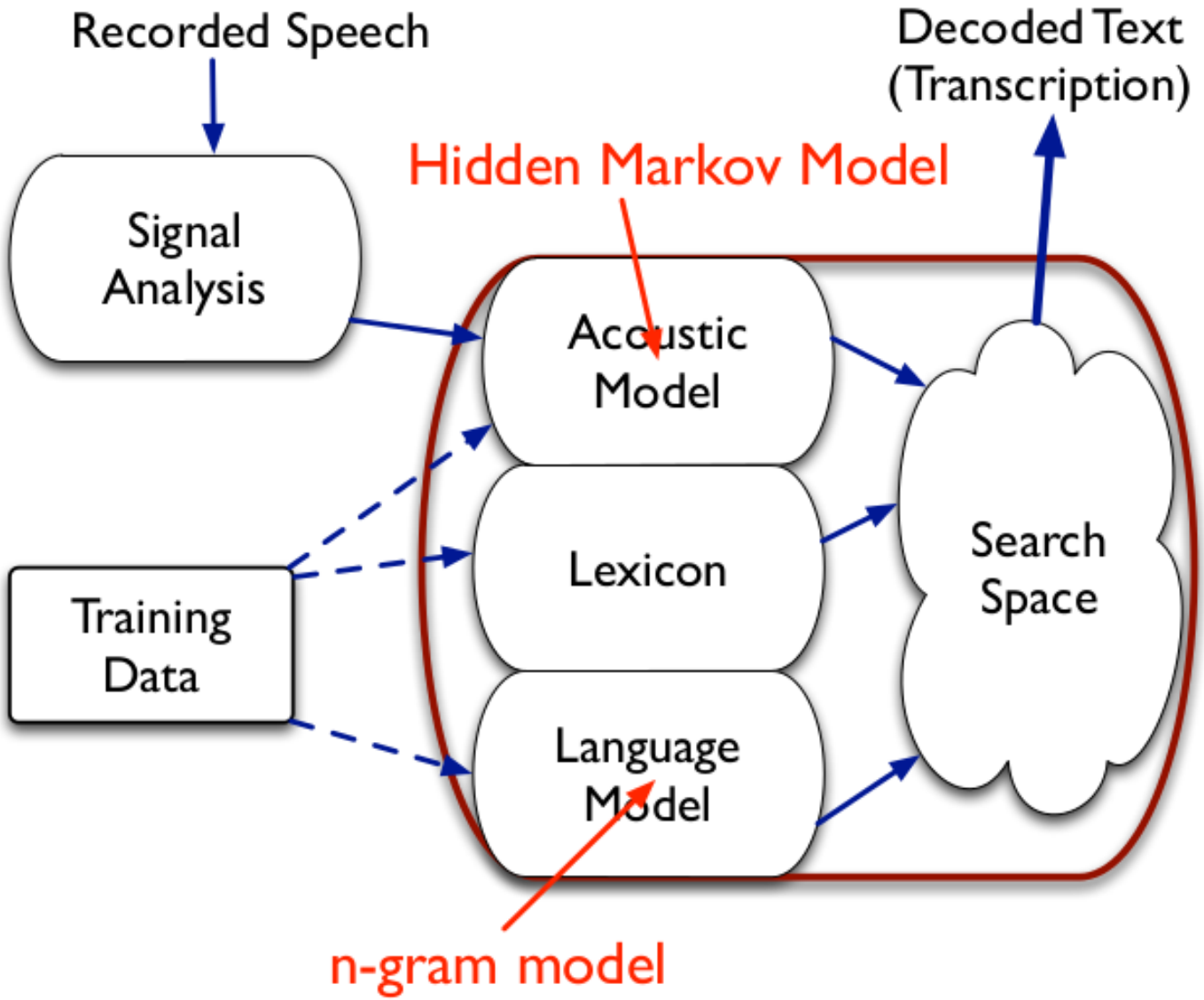
9. Limitation of Liability

... LICENSEE UNDERSTANDS THAT SPEECH RECOGNITION IS A **STATISTICAL PROCESS AND THAT RECOGNITION ERRORS ARE INHERENT IN THE PROCESS**. LICENSEE ACKNOWLEDGES THAT IT IS **LICENSEE'S RESPONSIBILITY TO CORRECT RECOGNITION ERRORS BEFORE USING THE RESULTS**. OF THE RECOGNITION.

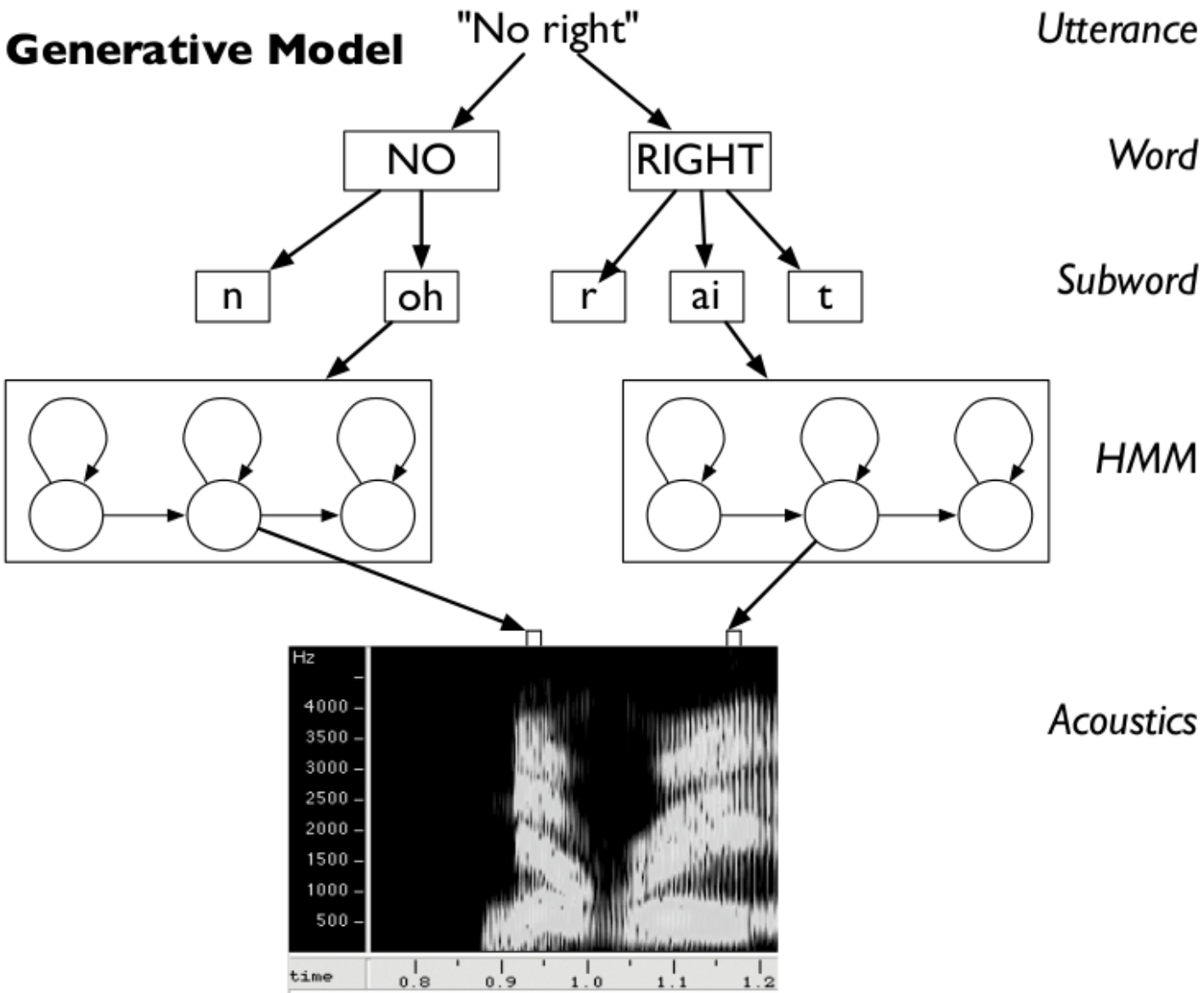
ASR Schematic Model



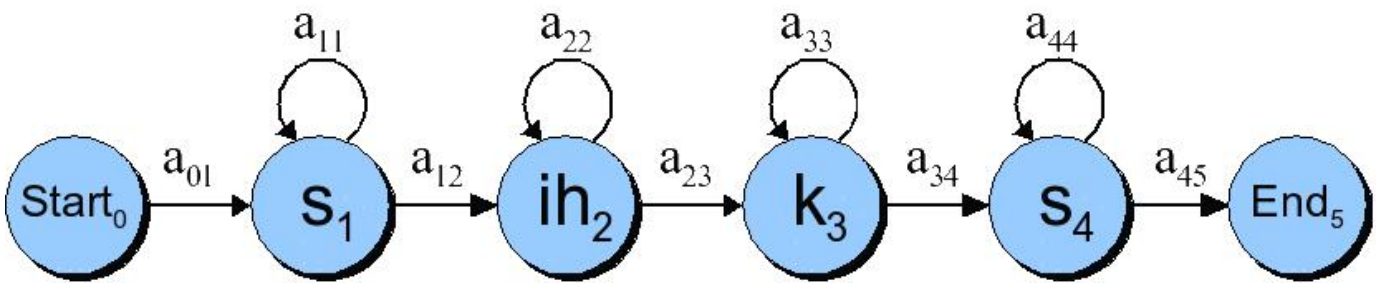
ASR Schematic Model



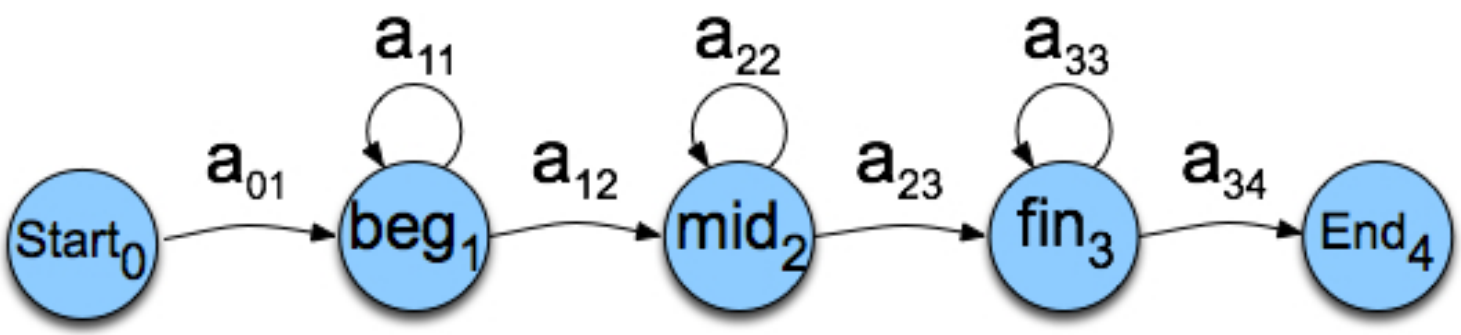
ASR Schematic Model



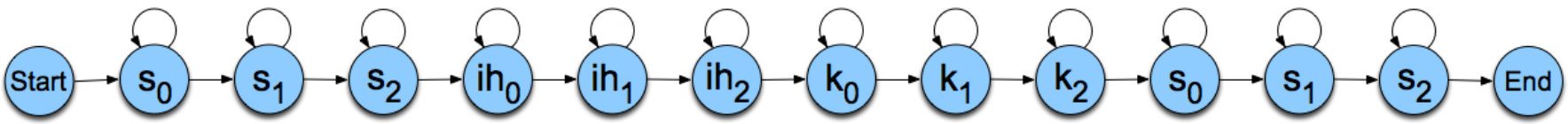
HMM for ‘six’



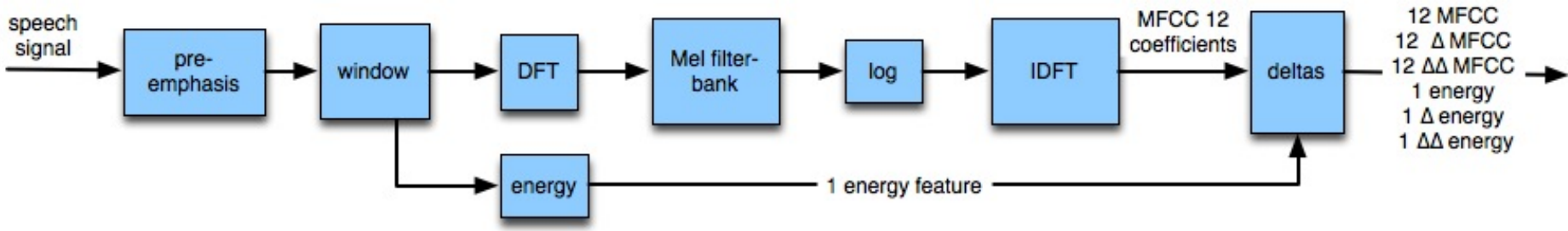
Each phone has 3 states



Each phone has 3 states: HMM for ‘six’



MFCC: Mel Frequency Cepstral Coefficients



The Cepstrum

One way to think about the cepstrum:

- separating the source and filter
- Speech waveform is created by
- A glottal source waveform
- Passes through a vocal tract which, because of its shape, has a particular filtering characteristic

Articulatory facts:

- The vocal cord vibrations create harmonics
- The mouth is an amplifier
- Depending on shape of oral cavity, some harmonics are amplified, some are dampened

Data!

- ASR is another example of learning from data
- Standard corpora with agreed evaluation protocols are very important for the development and evaluation of ASR
- TIMIT corpus (1986) –first widely used corpus, still in use
  - Utterances from 630 North American speakers
  - Phonetically transcribed, time-aligned
  - Standard training and test sets, agreed evaluation metric (phone error rate)
- Many standard corpora released since TIMIT: DARPA Resource Management, read newspaper text (e.g, Wall St Journal), human-computer dialogues (e.g, ATIS), broadcast news (e.g, Hub4), conversational telephone speech (e.g. Switchboard), multiparty meetings (eg AMI)
- Standard corpora have most value when closely linked to evaluation benchmark tests (with new test data from the same domain)

Evaluation

- How accurate is a speech recognizer?
- Use dynamic programming to align the ASR system's output with a reference transcription (a gold standard)
- Three types of error: insertion, deletion, and substitution
- Word error rate (**WER**), sums the three types of error. If there are N words in the reference transcript, and the ASR output has S substitutions, D deletions and I insertions, then:

$$WER = 100 \cdot \frac{I + D + S}{N}\%$$

- Accuracy = 100 - WER%
- Speech recognition evaluations: common training and development data, release of new test sets on which different systems may be evaluated using word error rate

For next time:

For next time:

1. Next time we'll talk in more detail about cepstral coefficients and HMMs for ASR.
2. Keep reading chapter 9 of Jurafsky and Martin
3. Midterm due Monday