The Basic Mathematics of Automatic Speech Recognition

The Source-Channel Model
Hidden Markov Models
N-gram Models
Viterbi Search

"*Recognize Speech*" v/s "Wreck a Nice Beach"

The Basic Mathematics of Automatic Speech Recognition

Sanjeev Khudanpur

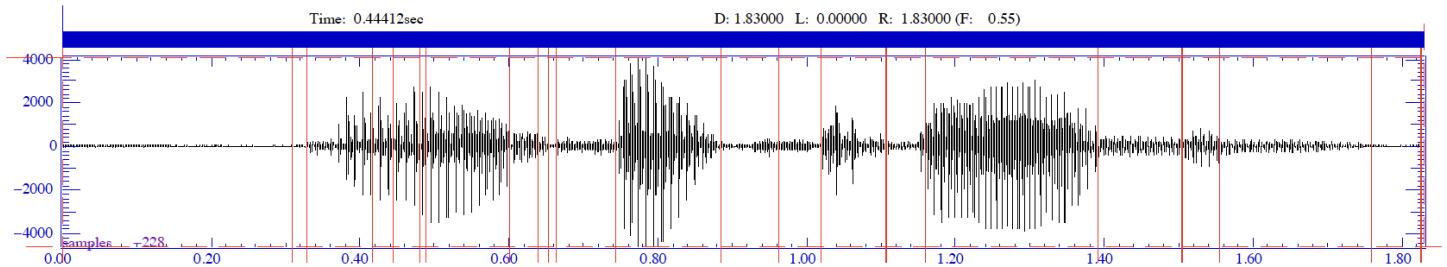
October 13, 2000

Center for Language and Speech Processing
The Johns Hopkins University



Recognition of Spoken Language

Convert an acoustic signal captured by a microphone or telephone to a sequence of words.

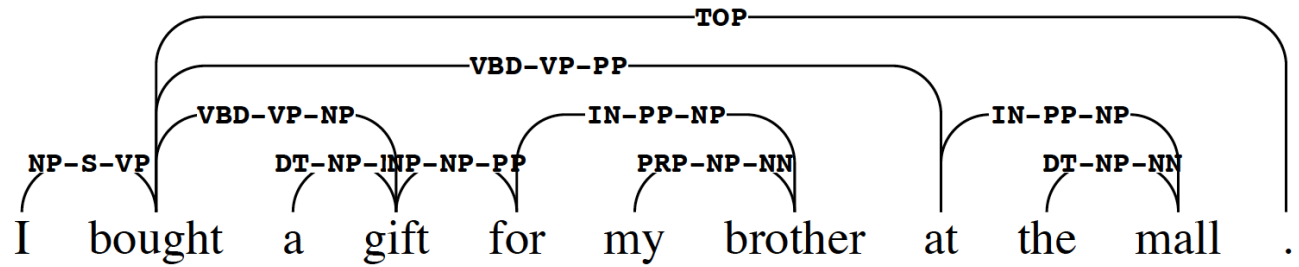


"OUR OWN FOLKS AT HOME AND ..."

Center for Language and Speech Processing
The Johns Hopkins University

Prior Knowledge — Linguistics

Grammar: Thoughts are organized as words, phrases, clauses,...



Phonology: Words have canonical pronunciations.

OWN /ow/ /n/, FOLKS /f/ /ow/ /k/ /s/

Phonetics: Phones have characteristic sounds.

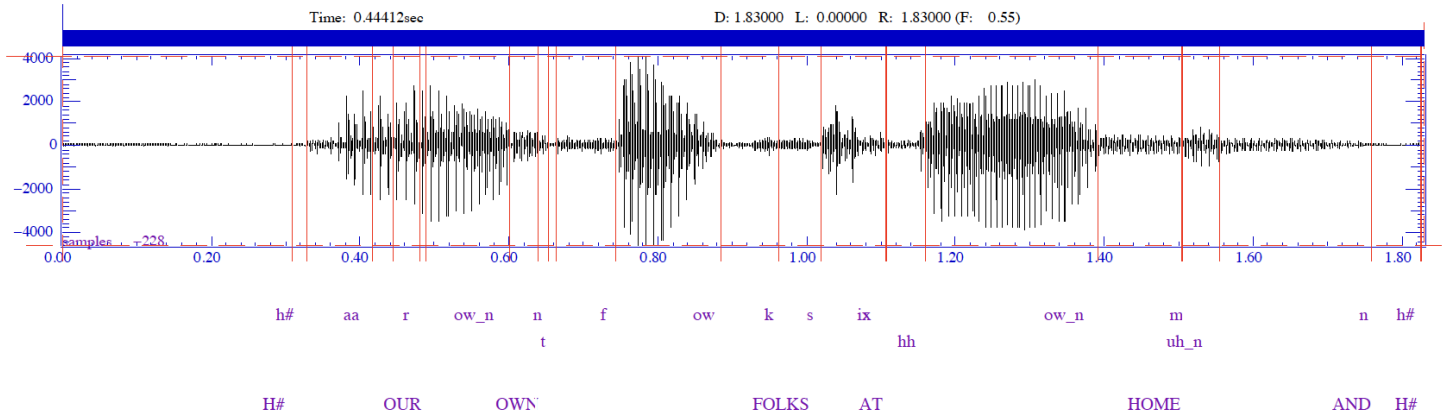
OWN [ow] [n], FOLKS [f] [ow] [k] [s]

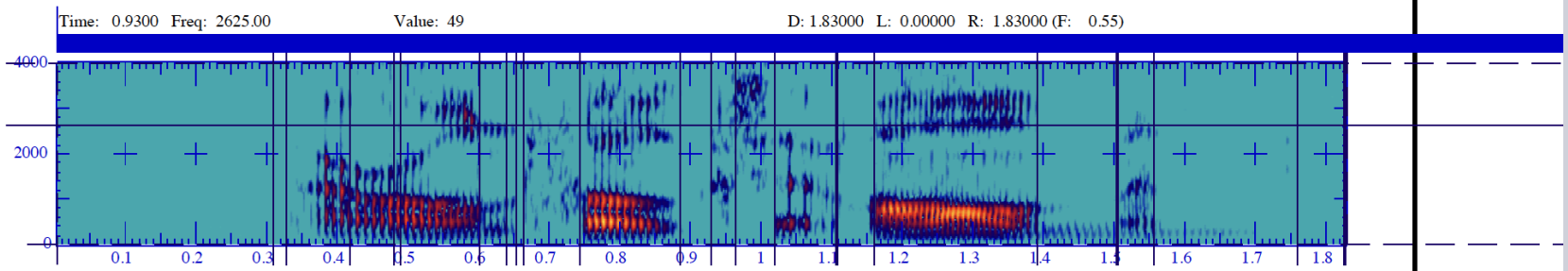
Center for Language and Speech Processing
The Johns Hopkins University



89.1%

The Acoustic Signal





Center for Language and
Speech Processing
The Johns Hopkins University

89.1%

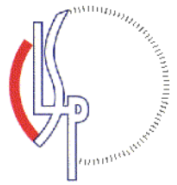
Signal Representation and the Acoustic Processor

* Psychoacoustic studies of the human auditory system – pitch, formants, vowels, consonants;
* "Tuning" or nerve cells in the ear at different frequencies – the MEL cepstrum;

  Get Fourier transform of a window of speech, *bin* the magnitude spectrum into a small number (8-15) of coefficients, get logarithm of the coefficients, decorrelate the feature vector via a Cosine transform.
* Sensitivity of nerve cells to spectral "slope" – the Δ-cepstrum.

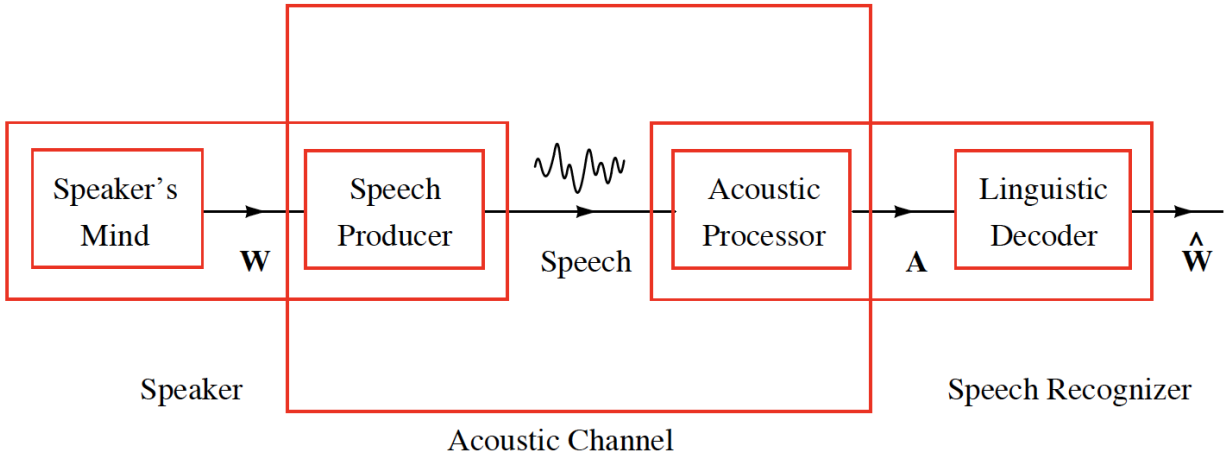
• Feature frames produced at periodic intervals – 10 ms.

Center for Language and Speech Processing
The Johns Hopkins University



89.1%

The Source-Channel Model



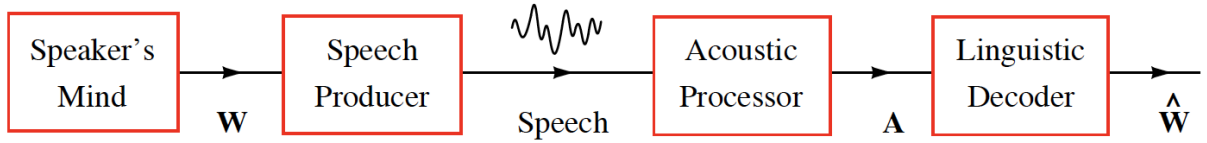
$$\begin{aligned}\widehat{\mathbf{W}} &= \arg\max_{\mathbf{W}} P(\mathbf{W}|\mathbf{A}) \\ &= \arg\max_{\mathbf{W}} \frac{P(\mathbf{A}|\mathbf{W})P(\mathbf{W})}{P(\mathbf{A})} \\ &= \arg\max_{\mathbf{W}} P_A(\mathbf{A}|\mathbf{W})P_L(\mathbf{W})\end{aligned}$$

Center for Language and Speech Processing
The Johns Hopkins University



89.1%

The Components Problems of Speech Recognition



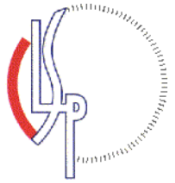
Speaker

Speech Recognizer

Acoustic Channel

$$\begin{aligned}\widehat{\mathbf{W}} &= \arg\max_{\mathbf{W}} P(\mathbf{W}|\mathbf{A}) \\ &= \arg\max_{\mathbf{W}} \frac{P(\mathbf{A}|\mathbf{W})P(\mathbf{W})}{P(\mathbf{A})} \\ &= \underbrace{\arg\max_{\mathbf{W}}}_{3} \underbrace{P_A(\mathbf{A}|\mathbf{W})}_{2} \underbrace{P_L(\mathbf{W})}_{1}\end{aligned}$$

Center for Language and Speech Processing
The Johns Hopkins University



89.1%

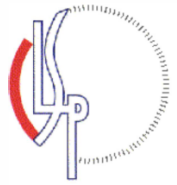
The Component Problems of Speech Recognition

1. **Language Modeling:** Assign probabilities to sequences of words in the language. Construct $P_L(\mathbf{W})$ from generic text or from transcriptions of task-specific dialogues.

2. **Acoustic Modeling:** Assign probabilities to acoustic realizations of a sequence of words. Construct $P_A(\mathbf{A}|\mathbf{W})$ form matched samples of acoustic signals and words.

3. **Hypothesis Search:** Find the word sequence with the maximum *a posteriori* probability. Search through the huge multitude of plausible word sequences for $\arg\max_{\mathbf{W}}$.

Center for Language and
Speech Processing
The Johns Hopkins University

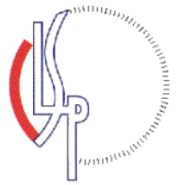


89.1%

The Acoustic Model

- A separate model $P_A(\mathbf{A}|\mathbf{W})$ for every $\mathbf{W}$ requires many samples of acoustic data for every sequence of words $\mathbf{W}$!
- Even if we chop up the acoustic signal to recognize individual words, $P_A(\mathbf{a}_t|w_t)$ still requires many samples of every word.
- There are only ~ 50 phonemes in English – resort to acoustic-phonetic models.
- Construct the model for a word by *stringing together* the models for the constituent phones, and models for word-sequences by *stringing together* models of words, *etc.*

Center for Language and Speech Processing
The Johns Hopkins University

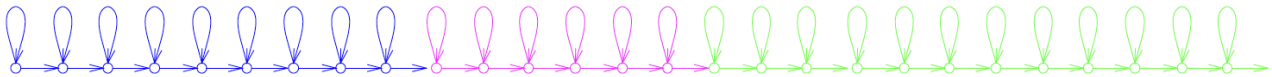


89.1%

Acoustic Phonetic Models

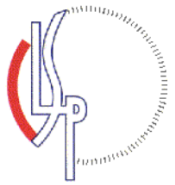
W AND OUR ONLY

B [ae] [n] [d] [aw] [r] [ow] [n] [l] [iy]



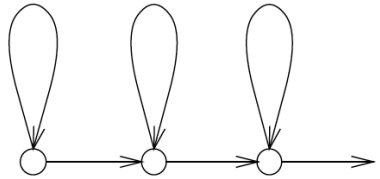
$P_A(\mathbf{A}|\mathbf{W})$ is a string of hidden Markov models (HMMs).

Center for Language and Speech Processing
The Johns Hopkins University



89.1%

Hidden Markov Models

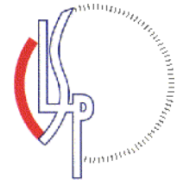


A hidden Markov model has a discrete Markov chain s_t, called the state process, which is *not* observed, and at discrete time instants, an output symbol a_t is emitted with a conditional probability $P_E(a_t|s_t)$.

$$P_A(\mathbf{A},\mathbf{S}) = \prod_{t=1}^{N} P_E(a_t|s_t)P_T(s_t|s_{t-1})$$

Three-state HMMs are commonly used to model phonetic segments – the onset, the nearly stationary middle, and the decay.

Center for Language and Speech Processing
The Johns Hopkins University



89.1%

Computational Issues with HMMs (1)

To compute the (marginal) probability of a acoustic signal $\mathbf{A}$, for a given $\mathbf{W}$, we must sum over all possible state sequences.

$$\begin{aligned} P_A(\mathbf{A}|\mathbf{W}) &= \sum_{\mathbf{S}\in\mathcal{S}^N_{\mathbf{W}}} P_A(\mathbf{A},\mathbf{S}|\mathbf{W}) \\ &= \sum_{\mathbf{S}\in\mathcal{S}^N_{\mathbf{W}}} \prod_{t=1}^{N} P_E(a_t|s_t)P_T(s_t|s_{t-1}) \end{aligned}$$

- There are exponentially many state sequences. Typically, $|\mathcal{S}| \sim 100$ and $N \sim 500$, $\Longrightarrow \left|\mathcal{S}^N\right| \sim 10^{1000}$.
- Efficient (Forward-Backward) algorithm for computation in essentially $2 \times |\mathcal{S}| \times N$ operations.

Center for Language and Speech Processing
The Johns Hopkins University



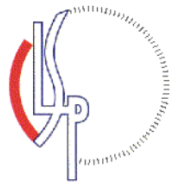
89.1%

Computational Issues with HMMs (2)

Estimating the component models $P_E(a_t|s_t)$ and $P_T(s_t|s_{t-1})$ would be easy if $\mathbf{S}$ were available along with $\mathbf{A}$ for every training sample. But $\mathbf{S}$ is missing.

- Efficient (Baum-Welch) algorithm for iterative update of model parameters.
- An instance of the Expectation-Maximization technique for maximum likelihood estimation from incomplete data.

Center for Language and Speech Processing
The Johns Hopkins University



89.1%

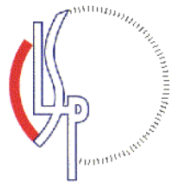
Computational Issues with HMMs (3)

To compute the most likely state sequence (among sequences corresponding to several word hypotheses $\mathbf{W}$) for an acoustic signal $\mathbf{A}$, for a given $\mathbf{W}$, we must search over all possible state sequences.

$$\begin{aligned}
\arg\max_{\mathbf{W}} P_A(\mathbf{A}|\mathbf{W}) &= \arg\max_{\mathbf{W}} \sum_{\mathbf{S}\in\mathcal{S}_{\mathbf{W}}^N} P_E(\mathbf{A}|\mathbf{S})P_T(\mathbf{S}) \\
&\approx \arg\max_{\mathbf{W}} \max_{\mathbf{S}\in\mathcal{S}_{\mathbf{W}}^N} P_E(\mathbf{A}|\mathbf{S})P_T(\mathbf{S}) \\
&= \arg\max_{\mathbf{S}\to\mathbf{W}} P_A(\mathbf{S}|\mathbf{A})
\end{aligned}$$

- Efficient algorithm discovered (Viterbi) for computation in essentially $|\mathcal{S}| \times N$ operations.

Center for Language and Speech Processing
The Johns Hopkins University



89.1%

The Language Model

Grammatical theory, formal languages, Noam Chomsky, ...
Little success in application to speech recognition.

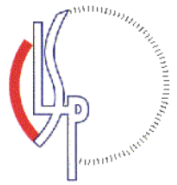
Simple statistical methods work well – Markov models.

$$\begin{aligned} P_L(\mathbf{W}) &= P_L(w_1, w_2, \ldots, w_k) \\ &= \prod_{t=1}^{k} P_L(w_t|w_{t-1}, w_{t-2}), \end{aligned}$$

where $P_L(w_t|w_{t-1}, w_{t-2})$ is estimated by simply counting up the relative frequencies f of tuples in a corpus of text.

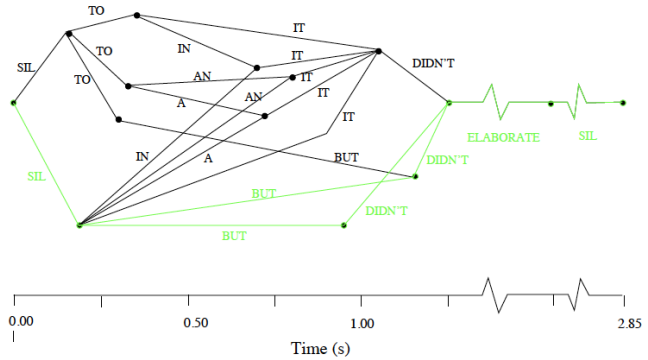
$$P_L(w_t|w_{t-1}, w_{t-2}) = \lambda_0 \frac{1}{|\mathcal{V}|} + \lambda_1 f(w_t) + \lambda_2 f(w_t|w_{t-1}) + \lambda_3 f(w_t|w_{t-1}, w_{t-2})$$

Center for Language and Speech Processing
The Johns Hopkins University



The Search Problem

Even for Viterbi search, the list of word hypotheses is very large.



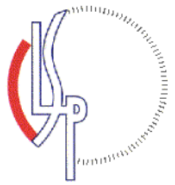
- Use graph minimization from Finite State Automata Theory to further reduce computational burden in the search.
- Use heuristics to prune unlikely hypotheses from the search.
- Important research area for real-time applications.

Center for Language and Speech Processing
The Johns Hopkins University

The State-of-the-Art

Application	Accuracy
Replacing Touch-Tone Menus	99.5%
Call Classification and Routing	95%*
Interactive Voice Response Systems	90%
Desk-top Dictation (with enrollment)	95+%
Transcription of Broadcast News	80-85%
Conversational Telephone Speech	65%
Universal voice interface	??.

Center for Language and Speech Processing
The Johns Hopkins University



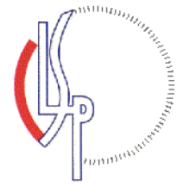
Pronunciation Variability in Conversational Speech

Pronunciations vary due to

- coarticulation: *have to* → *hafto*;
- reductions: *going to* → *gonna*;
- fast speech: *probably* → *probly*;
- Dialect: *native speakers*.
- Accent: *nonnative speakers* .

Conversational speech has remarkable prosodic variability as well.

Center for Language and Speech Processing
The Johns Hopkins University



89.1%

Pronunciation Variability in Conversational Speech (Audio Examples)

- Read v/s Play-Acting v/s Truly Spontaneous speech
  - Courtesy SRI
- Phonetic transcription of spontaneous speech.
  - Courtesy ICSI

Center for Language and Speech Processing
The Johns Hopkins University



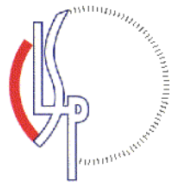
89.1%

Dialect Diversity (Audio Examples)

Courtesy Prof. W. Labov, UPenn

one block	wear socks
the top	the locks?
of red	looking red
her head	I said
are steady	men's clothing

Center for Language and Speech Processing
The Johns Hopkins University



Conclusion

100% accuracy is not always needed – much can work with less.

Applications must be designed with capabilities in mind.

Robust statistical estimation methods must be developed.

Current techniques require large databases matched to target applications.

Mathematical models for language understanding, dialogue and discourse are needed.

Center for Language and Speech Processing
The Johns Hopkins University

