Aspects of the Statistical Approach to Speech Recognition (isitttalk.tex)

Frederick Jelinek
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD

June 29, 2001

1 Terminology

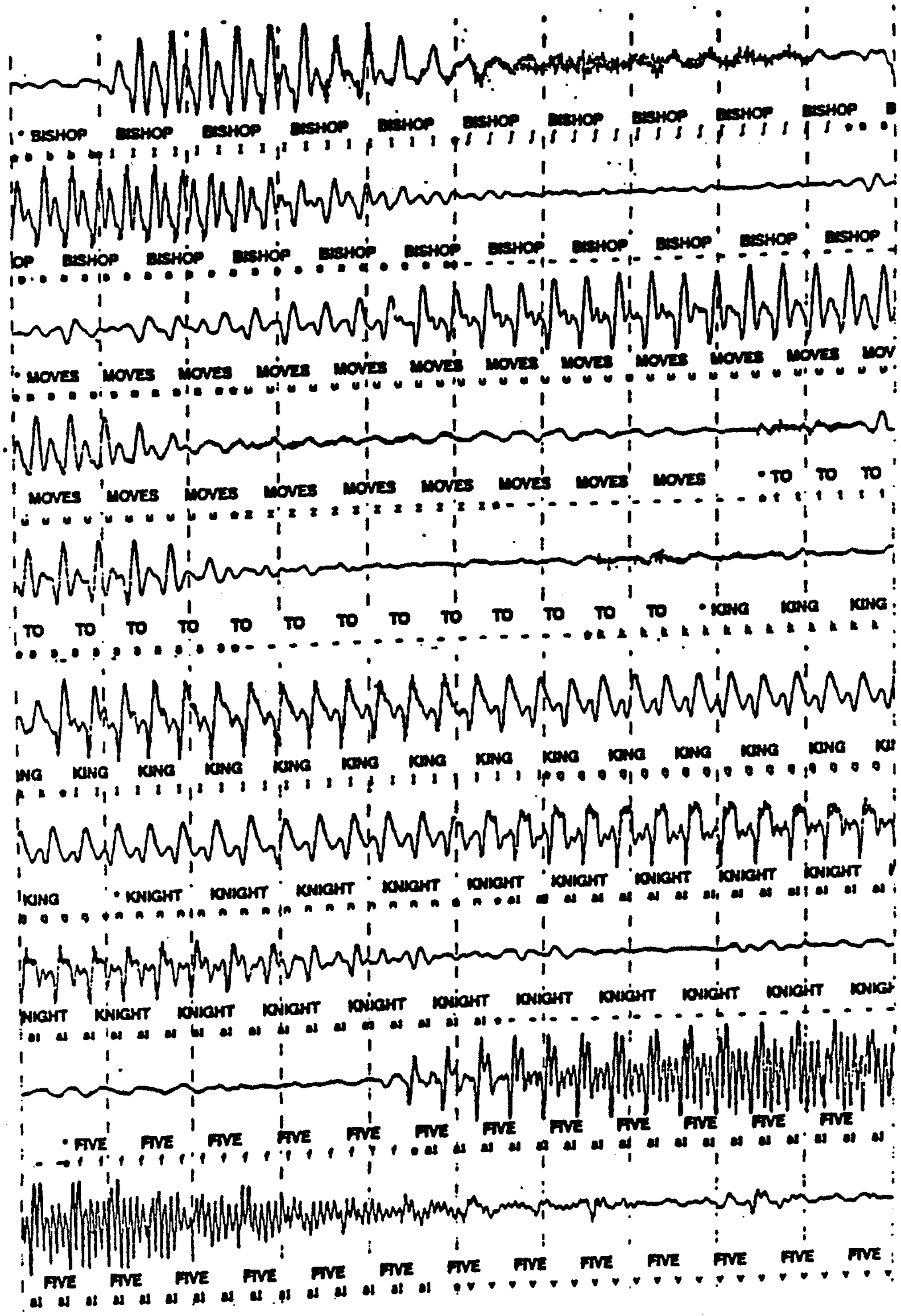
Speech recognition:

Automatic transcription of the sound of speech into text

Speech understanding:

Determination of intended meaning of observed speech

Bishop Moves to King Knight Five



3 Recognition Approach Based on "available" Expert Knowledge

1. A basic pronunciation of words is known (see any dictionary). It can be expressed as strings from the international phonetic alphabet:

$$H \quad AE \; N \quad D \quad \# \quad L \quad A \quad B \quad E \quad L \quad I \quad NG$$

2. Fast speech and co-articulation rules are known from phonology, so transformation *by rule* to surface form is possible:

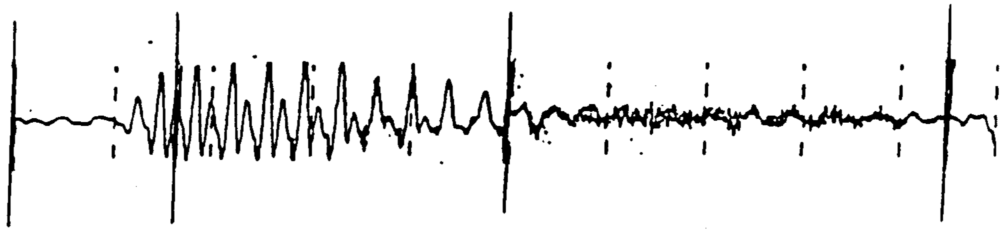
$$H \quad AE \; N \quad L \quad A \quad B \quad L \quad I \quad NG$$

3. Confusion between phones can be estimated from place of articulation studies and psycho-acoustics

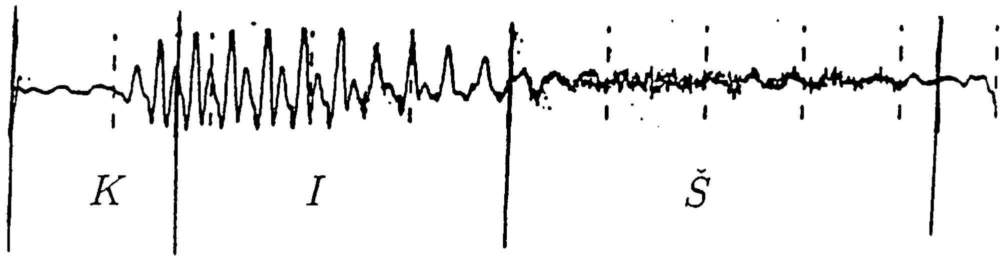
   - I.e., we can find out how frequently s changes into z, how frequently into f, into e, etc.

4 The "expert" procedure

1. Segment speech into successive phones



2. Perform pattern match on the segments: extract a string of most likely phones from the speech signal



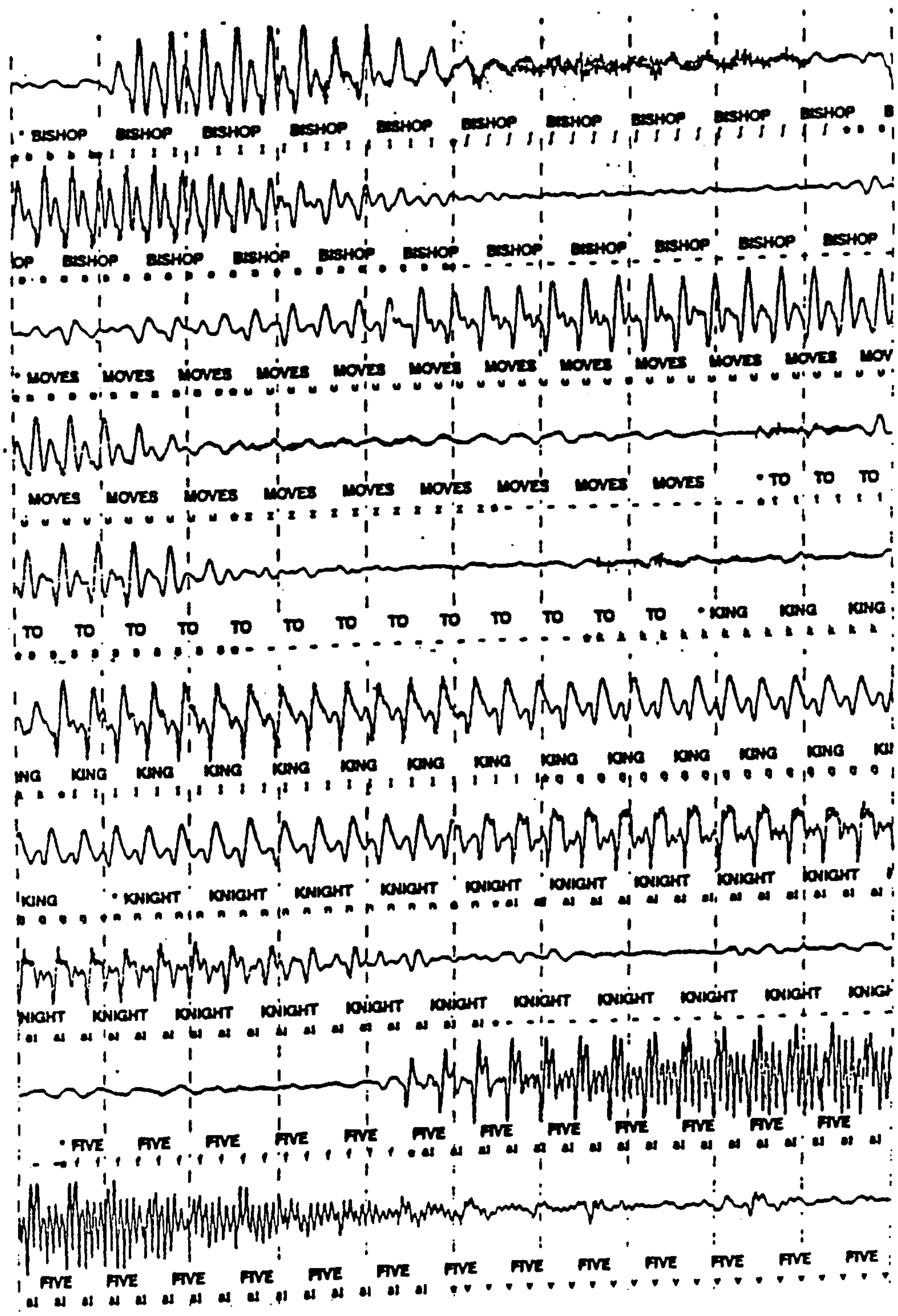
5 The "expert" procedure (cont.)

3. Estimating confusion penalties compute the "distance" between hypothesized words and the extracted phones

K	I	$\check{S}$	recognized string
B	I	$\check{S}$	hypothesized string
-3	0	0	penalties

4. Find what was said with the help a heuristic search based on
   - a grammar of English
     – and
   - an error - scoring system determined by experts

Bishop Moves to King Knight Five



7 The accepted statistical decision criterion

- Denote word sequences by

$$\mathbf{W} \doteq w_1 w_2 ... w_n \qquad \text{(the spoken sentence)}$$

$$\widehat{\mathbf{W}} \doteq \hat{w}_1 \hat{w}_2 ... \hat{w}_n \qquad \text{(the transcribed speech)}$$

where we *ignore* the possibility that the number of transcribed and uttered words *may be different.*

- "Obviously", the recognizer should decide for that word sequence $\widehat{\mathbf{W}}$ which occurred most frequently when the acoustics $\mathbf{A}$ were observed. Or, written mathematically,

$$\widehat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{W}|\mathbf{A}) = \arg\max_{\mathbf{W}} P(\mathbf{A}|\mathbf{W}) \times P(\mathbf{W})$$

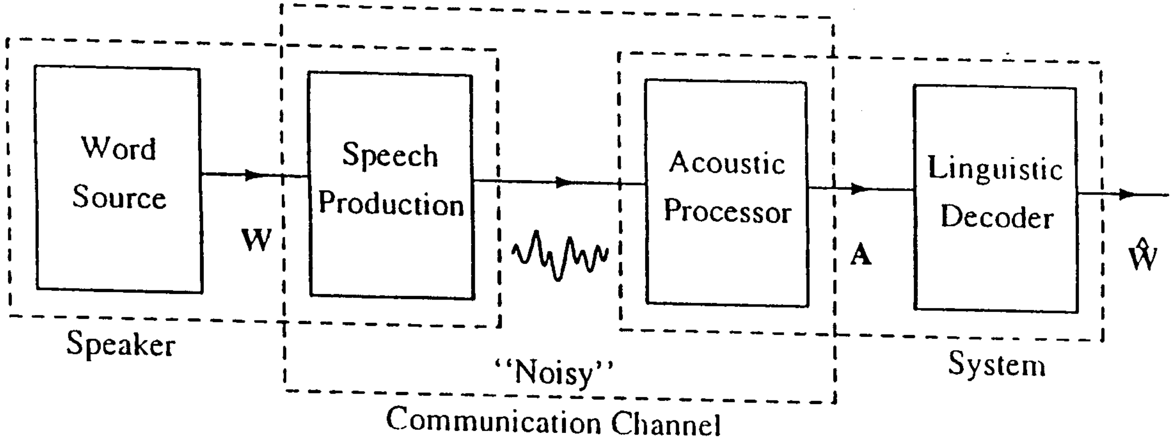
- **Remark:** In cases of *research* interest, it is too much to hope to recover sentences perfectly. Hence we should want to *minimize* the word error rate. The criterion does that only approximately.

8 Consequent component modules of a recognizer

- We conclude that a recognizer must have
  - An **acoustic processor** which transforms the observed speech into a string of symbols $\mathbf{A}$ to be handled by the computer
  - A **hypothesis search module** which seeks the word string $\widehat{\mathbf{W}}$ that attains the maximum of $P(\mathbf{A}|\mathbf{W}) \times P(\mathbf{W})$.

- The search is based on **models** of the speech processes:
  - **Acoustic model:** to compute the probability $P(\mathbf{A}|\mathbf{W})$ that when speaker utters $\mathbf{W}$ the speech will be transformed by the acoustic processor into the string $\mathbf{A}$.
  - **Language model:** to compute the probability $P(\mathbf{W})$ that the speaker will wish to utter the words $\mathbf{W}$

9 The Communication Theory Approach to Speech Recognition



10 The basic pronunciation model

- The system contains a finite pronunciation lexicon specifying a correspondence between each word and its *baseform* expressed as a phonetic sequence

$$\text{chair} \quad \leftrightarrow \quad \text{ČÉR}$$

- An utterance $\mathbf{W}$ is transformed into a phonetic string by replacing each of its words by its baseform followed by a delimiter

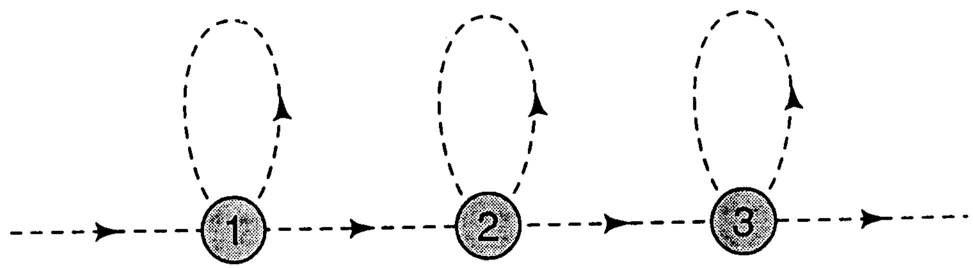
$$\text{blue chair} \quad \leftrightarrow \quad |\ \text{B L Ú}\ |\ \text{Č É R}\ |$$

- Phones are pronounced according to their immediate context: the acoustic model of a phone is a *tri-phone*

$$|-\text{B}+\text{l},\ \ \text{b}-\text{L}+\text{ú},\ \ \text{l}-\text{Ú}+|,\ \ \text{ú}-|+\text{č},\ \ |-\text{Č}+\text{é},\ \ \text{č}-\text{É}+\text{r},\ \ \text{é}-\text{R}+|$$

11 The basic acoustic model

- The microphone input is transformed by a signal processor into a sequence $a_1 a_2 \ldots a_k \cdots$ of vectors of *cepstral* coefficients
  - Vectors are generated 100 times a second
- Each tri-phone corresponds to a hidden Markov model (HMM) of the same structure:



- Transitions take place once every centi-second. States generate normally distributed vectors.
  - Tri-phones differ in that their statistical parameters have different values.
  - The parameter values are estimated from transcribed speech data by the EM algorithm.

12 The basic language model

- Almost universally a *trigram* language model is used

$$P(\mathbf{W}) = \prod_{i=1}^{n} P(w_i|w_{i-2}, w_{i-1})$$

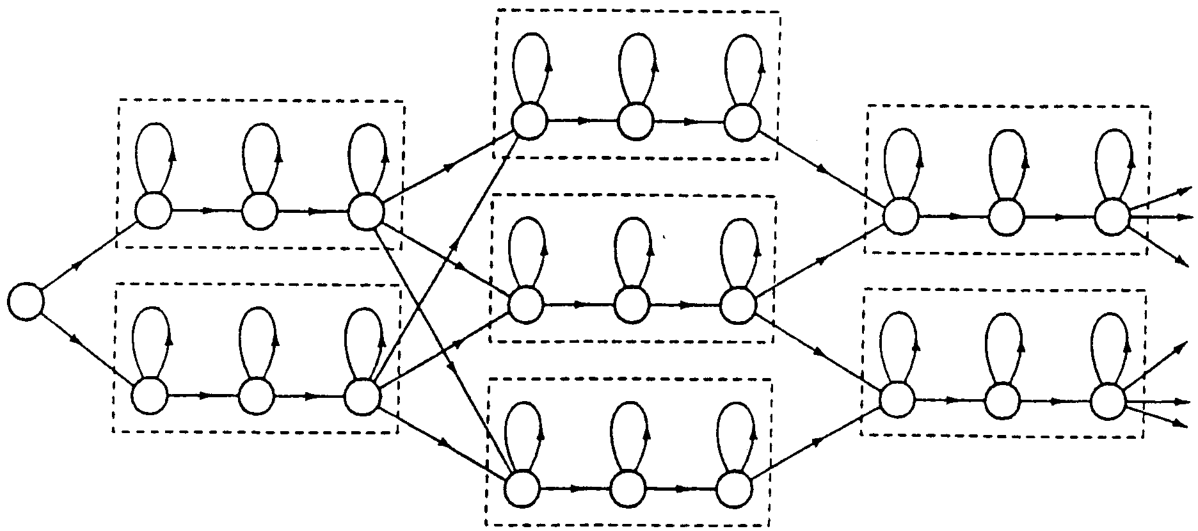
- The probabilities are estimated from trigram counts collected from text data. Smoothing is required:

$$P(c|a,b) = \begin{cases} f(c|a,b) & \text{if } C(a,b,c) \geq 6 \\ \alpha\, Q_T(c|a,b) & \text{if } \ 6 > C(a,b,c) \geq 1 \\ \beta(a,b)\, P(c|b) & \text{otherwise} \end{cases}$$

- The above is *back-off* smoothing. $Q_T(c|a,b)$ is a *Good - Turing* estimate based on the counts $C(a,b,c)$

13 Advantages of the HMM formulation

- Simple and uniform structure
- The complete model for the task (including the language model) is one large composite HMM:
  - the transitions between words are *ordinary* HMM transitions between the *final* state of the previous word and the *initial* state of the next word



14 Advantages of the HMM formulation (cont.)

- The search for the best word sequence $\widehat{\mathbf{W}}$ is just a search for the best path through the **trellis** of the **composite HMM.**
  - The hypothesis search is mostly based on the Viterbi algorithm and sometimes on the stack algorithm.
- We can determine the values of the model parameters directly from speech data:
  - Using the special case (forward - backward) of the EM algorithm
    * This is a maximum likelihood approach (maximum mutual information also possible)
  - Applies to **all languages**

(12)

1	The	are	**to**	know	the	issues	necessary
2	This	will		have	this	problems	data
3	One	the		understand	these	**the**	information
4	Two	would		do	problems		above
5	A	also		get	any		other
6	Three	do		the	a		time
7	Please	**need**		use	problem		people
8	In			provide	them		operators
9	**We**			insert	**all**		tools
•				•			•
•				•			•
•				•			•
93				request			factors
94				respond			facts
95				supply			I
96				write			jobs
97				me			MVS
98				**resolve**			old
•							•
•							•
•							•
1636							mailroom
1637							marketplace
1638							provision
1639							reception
1640							shop
1641							**important**

1	role	and	**the**	**next**	be	meeting	of
2	thing	from			**two**	months	•
3	that	in				years	
4	to	to				meetings	
5	contact	are				to	
6	parts	with				week	
7	point	were				**days**	
8	for	requiring					
9	**issues**	still					
•		•					
•		•					
•		•					
61		being					
62		during					
63		I					
64		involved					
65		would					
66		**within**					

16 Another example of the power of trigrams

Reconstruction of a short sentence from a bag of words:

- Scramble words of a sentence

- Use trigram language model to find most probable word order, i.e.,
  - From set $\{v_1, v_2, ..., v_n\}$ find the sequence

$$w_1 = v_{i_1}, w_2 = v_{i_2}, ..., w_n = v_{i_n}$$

  that will maximize the value of

$$P(w_1 w_2 ... w_n) \doteq P(w_1)P(w_2|w_1)P(w_3|w_1 w_2)...P(w_n|w_{n-2}w_{n-1})$$

17 Sentence reconstruction results

- 38 randomly selected sentences of $n \leq 10$ words

- 24 sentences reconstructed exactly (63%).

- 9 more reconstructions have same meaning as originals (24%)

- Reconstruction error only 13%.

18 Reconstruction examples

- Meaning preserved:

would	*I*	*report*	*directly*	*to*	*you*	?
I	*would*	*report*	*directly*	*to*	*you*	?

now	*let*	*me*	*mention*	*some*	*of*	*the*	*disadvantages*	.
let	*me*	*mention*	*some*	*of*	*the*	*disadvantages*	*now*	.

he	*did*	*this*	*several*	*hours*	*later*	.
this	*he*	*did*	*several*	*hours*	*later*	.

- Meaning destroyed:

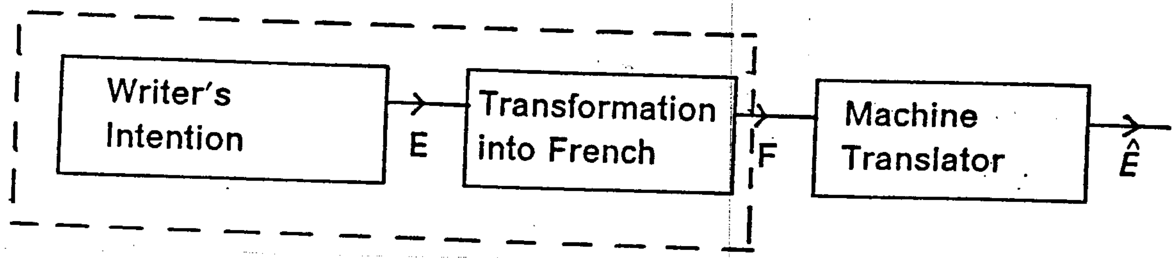
in	*our*	*organization*	*research*	*has*	*two*	*missions*	.
in	*our*	*missions*	*research*	*organization*	*has*	*two*	.

exactly	*how*	*this*	*might*	*be*	*done*	*is*	*not*	*clear*	.
clear	*is*	*not*	*exactly*	*how*	*this*	*might*	*be*	*done*	.

19 Natural Language Processing (NLP) tasks

- Under the influence of success in automatic speech recognition (ASR), the communication theory formulation is being applied also to the following problems:
  - Tagging: part-of-speech assignment to text (POS)
  - Machine translation (MT)
  - Text parsing

20 Communication theory formulation of the MT problem



21 A mathematical formulation of MT

- Example: French–to–English translation:
  - $\mathbf{F}$ is observed French word string

$$\mathbf{F} = f_1, f_2, ..., f_K \qquad f_i \in \mathcal{F}$$

  - $\mathbf{E}$ is hypothesized underlying English text

$$\mathbf{E} = e_1, e_2, ..., e_n \qquad e_i \in \mathcal{E}$$

- The translation machine seeks

$$\widehat{\mathbf{E}} = \arg\max_{\mathbf{E}} P(\mathbf{E})\, P(\mathbf{F} \mid \mathbf{E}, K)$$

- $P(\mathbf{E})$ is provided by an English language model (LM). The main problem is designing a model of the *transformation* process $P(\mathbf{F} \mid \mathbf{E}, K)$.

22 The basic transformation model $P(\mathbf{F} \mid \mathbf{E}, K)$.

- Based on *alignment* $\mathbf{L}$ between $\mathbf{F}$ and $\mathbf{E}$ stating which subset of words $\{f_{i,1}, ..., f_{i,m_i}\}$ of $\mathbf{F}$ have their "origin" in particular words e_i of $\mathbf{E}$.
  - The alignment is specified by *labeling* each word f_j of $\mathbf{F}, 1 \leq j \leq K$ by a label $l \in \{1, 2, ..., n\}$
- We then get

$$P(\mathbf{F} \mid \mathbf{E}, K) = \sum_{\mathbf{L}} P(\mathbf{F}, \mathbf{L} \mid \mathbf{E}, K)$$

where

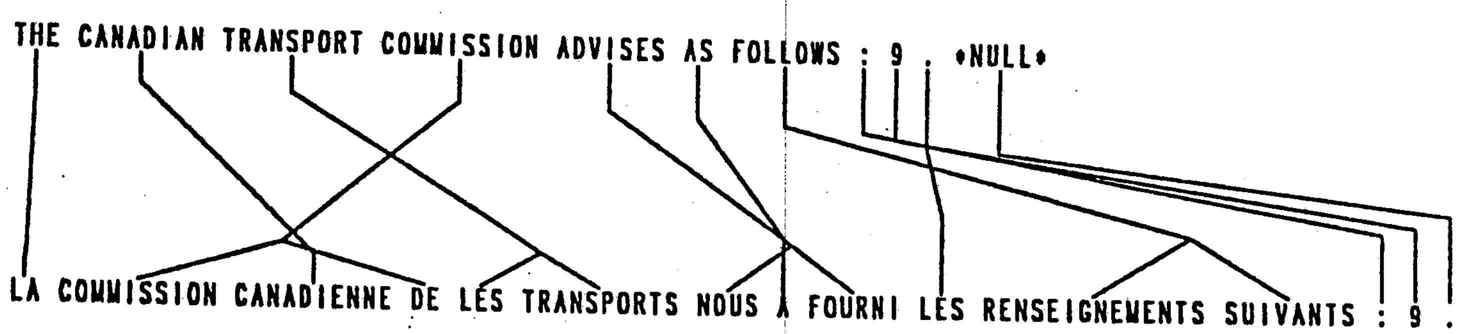
$$P(\mathbf{F}, \mathbf{L} \mid \mathbf{E}, K) = P(\mathbf{F} \mid \mathbf{L}, \mathbf{E}, K)\, P(\mathbf{L} \mid \mathbf{E}, K)$$

and

$$P(\mathbf{F} \mid \mathbf{L}, \mathbf{E}, K) = \prod_{i=1}^{n} Q(m_i|e_i) \prod_{j=1}^{m_i} P(f_{i,j}|e_i)$$

- $P(\mathbf{L} \mid \mathbf{E}, K)$ is a rather complex *permutation* probability that is independent of $\mathbf{F}$ itself, and is made up of factors that depend on the words e_i.
- The transformation process $P(\mathbf{F} \mid \mathbf{E}, K)$ is **hidden**, and is not sequential!
  - The parameters can be extracted by an iterative re-estimation process from a database of mutual bi-lingual translations.

23 Aligning French and English words

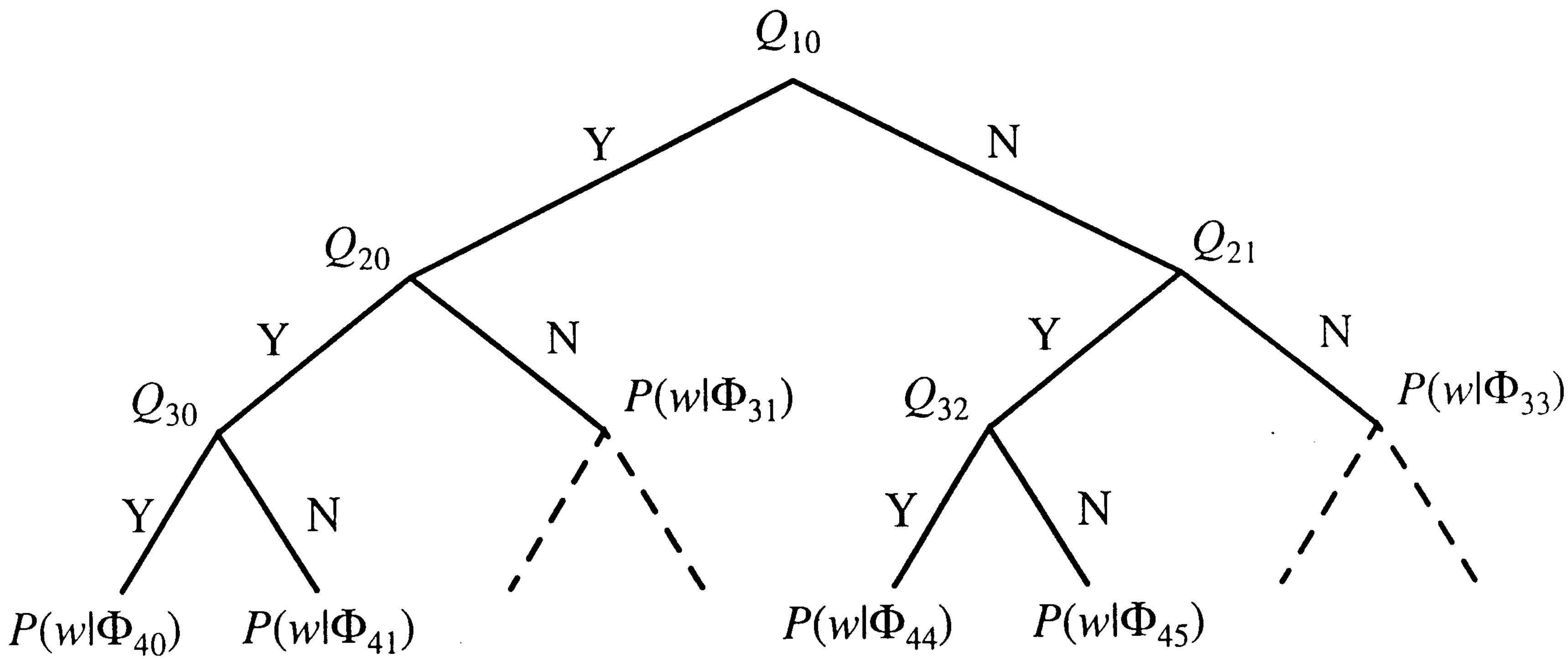


24 Essentials of the statistical approach

- Choice of structure of the parametric model
  - Based on intuitive understanding of the process
  - A compromise between precision and feasibility of parameter estimation
- Automatic estimation of statistical parameter values from data based on a clearly defined criterion
  - Data needs human annotation
  - **Data is invariably sparse**
- Overcoming data sparseness
  - Equivalence classification of states in parameter space
  - Smoothing of estimated parameter statistics
- Objective evaluation of performance based on annotated data
  - Possible for ASR, POS, and parsing
  - Human judgement required for MT

25 Overcoming sparseness: equivalence classification

- In no NLP area is the available data sufficient for direct estimation of probabilities of the distinguishing phenomena. E.g., in ASR:
  - Tri-phone HMM building blocks: ~75 phones $\Longrightarrow$ $421,875$ tri-phones$\Longrightarrow$ $1,265,625$ states
  - Tri-gram language model: ~$60K$ words $\Longrightarrow$ 2.16×10^{14} parameters
- Phenomena must be put into *equivalence* classes to which statistical parameters will correspond
- Automatic class selection on the basis of relevant training data:
  - top–down: decision trees
  - bottom–up: agglomerative clustering
- Both methods have *cross-entropy* as criterion



Q_{10}
Y
N
Q_{20}
Q_{21}
Y
N
Y
N
Q_{30}
$P(w|\Phi_{31})$
Q_{32}
$P(w|\Phi_{33})$
Y
N
Y
N
$P(w|\Phi_{40})$
$P(w|\Phi_{41})$
$P(w|\Phi_{44})$
$P(w|\Phi_{45})$

Figure 10.1

26 Example: tri-phone clustering

1. Train-up monophone model and use it to segment transcribed speech

2. Divide the speech into a training and check portion

   - In each portion, for each monophone and state (3 states per phone) obtain a collection of speech vector sequences:
     $\{z_{11}z_{12}...z_{1k_1}\}$, $\{z_{21}z_{22}...z_{2k_2}\}$, ... , $\{z_{n1}z_{n2}...z_{nk_n}\}$, each sequence corresponding to the specific tri-phone context.

3. Questions concerning *phonetic* context (e.g., in the speech transcription, is the preceding phone a fricative?) divide both collections into two subsets

   - Create two states, fitting the *training* statistics of each to the corresponding data subset
   - Test the likelihood of generating the check sequences from the corresponding two states

27 Example: tri-phone clustering (cont.)

4. Choose that question (and its split) for which the likelihood of the check sequences is highest.

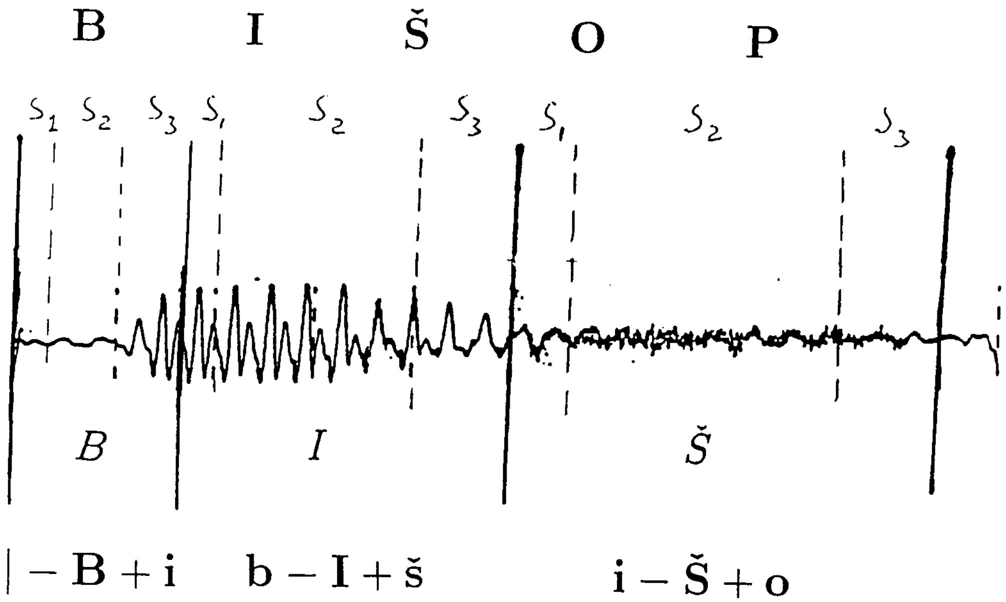
5. Keep splitting until stopping criterion met.

6. Using the obtained state equivalence classification, train-up the total tri-phone acoustic model

7. Iterate

28 Aligning speech with tri-phone states in context

- Baseform:



29 Comments on decision tree method

- Method is greedy
- Choice of the basic set of questions depends on intuition (expert knowledge)
- Chou algorithms generates questions directly from data, but suffers from data sparseness
- Main draw-back of decision trees is *data fragmentation*
  - Paradox: method used to overcome data sparseness suffers itself from the same sparseness

30 Overcoming sparseness: maximum entropy estimation

- Maximum entropy method estimates distributions $P(\mathbf{x})$ by insisting that
  - $P(\mathbf{x})$ satisfy prescribed linear constraints

$$\sum_{\mathbf{x}} P(\mathbf{x})\, k_i(\mathbf{x}) = d(i) \qquad i = 1, 2, ..., M$$

  - given the constraints, the entropy of the chosen distribution should be maximal
- In practice, $k_i(\mathbf{x})$ are indicator functions chosen so that $d(i)$ is its *believed* value.
  - The idea is that while data is sparse, there is enough of it to reliably estimate the marginal.
  - E.g., *voting*: estimate $P(x, y_1, ..., y_k)$ when $P(x, y_i)$, $i = 1, ..., k$ are thought reliably estimated
- $P(\mathbf{x})$ is a product of factors, one for each constraint in which $\mathbf{x}$ participates:

$$P(\mathbf{x}) = \prod_{i=1}^{M} e^{\lambda_i k_i(\mathbf{x})}$$

λ_i's must be determined so resulting $P(\mathbf{x})$ satisfies imposed constraints.

31 Example: Language modeling

- Memory of trigram model is too short
  - A pentagram is better, **however**, the number of its parameters is excessive
- But: we could have an approximation to a pentagram model $P(w_1, w_2, w_3, w_4, w_5)$ by constraining marginals
  - E.g. $P(w_3, w_4, w_5), P(w_2, w_5), P(w_1, w_5)$, etc.
- We can add grammatical constraints, in the form of parts of speech $g \in \{$NOUN, VERB, PREPOSITION, ...$\}$:

$$\sum_{\mathbf{x}} P(w_4, w_5)\, k_g(w_4) = f(g, w_5) \text{ where } k_g(w) = \begin{cases} 1 & \text{if } g(w) = g \\ 0 & \text{otherwise.} \end{cases}$$

32 Comments on maximum entropy method

- Method warps the probability distribution so it satisfies the imposed constraints
- **Advantages:**
  - Natural way of handling simultaneous conditions
- **Draw-backs:**
  - Excessive computational effort (and/or storage requirement) to determine λ_is
  - Difficulty of finding an efficient set of constraint functions $k_i(\mathbf{x})$
  - Necessity of knowing targets $d(i)$ *exactly*
  - Assumption that whatever we don't know exactly, we don't know at all.

33 In conclusion

- The statistical approach provides us with **a unified point of view** applicable to all languages and requiring a minimum of expert preparation
- A clear statement of the problem and of the goal
  - Search for $\widehat{\mathbf{W}}$ maximizing $P(\mathbf{A}|\mathbf{W}) \times P(\mathbf{W})$
- The entire design of the recognizer is based on actual data related to the process:
  - Raw and transcribed speech (for $\mathbf{A}$ and $P(\mathbf{A}|\mathbf{W})$)
  - Training text (for $P(\mathbf{W})$)
- Current speech recognizers (based on a vocabulary of 60 thousand words) are capable of transcribing natural dictated speech with less than a 10% error rate.
- The next challenges:
  - Real speech and text understanding
  - Machine translation