University of California
Berkeley

College of Engineering
Department of Electrical Engineering
and Computer Sciences

Professors : N.Morgan / B.Gold

EE225D Spring,1999

Human Speech Recognition

Lecture 19

Human Speech Recognition

- Experiments dating from 1918 dealing wth noise and reduced BW
- Statistics of CVC perception
- Comparisons between humans and today's best systems
- Brief peak at human-inspired ASR research

Assessing Recognition Accuracy

- Articulation
- Intelligibility

Fletcher Experiments

- CVC, VC,CV nonsense syllables
- 74% of syllables used
- Tests over different SNR, bands

Articulation Test Record

Date 3-16-28
Title of test Practice tests
Test no. 10
List Nos. 5-9-37

Syllable articulation 0.515 = S
Condition tested 1500~ low pass filter
Observer W.H.S.
Caller E.B.

No.		Observed		Called	Observed		Called	Observed		Called
1	The first group is	ma'v		na'v	po'z		po't'h	kōb	√	kōb
2	Can you hear	pōch	√	pōch	nēz		nēzh	shēt'h		siz
3	I will now say	seng	√	seng	jo'ch	√	jo'ch	fūch	√	fūch
4	As the fourth write	chūd	√	chūd	t'ha'm	√	t'ha'm	thōl	√	thōl
5	Write down	run	√	run	hab	√	hab	pot'h	√	pot'h
6	Did you understand	chis		kiz	def		doth	wa'm	√	wa'm
7	I continue with	fos		fosh	chech		chej	gūm		gūn
8	These sounds are	lo'l	√	lo'l	lun		lon	nāsh		nāth
9	Try the combination	jās		zhāth	shāl	√	shāl	vo'g	√	vo'g
10	Please record	t'ha'th		t'ha'sh	mus	√	mus	lung		long
11	Write the following	wūr	√	wūr	lēd		bēd	dis		dizh
12	Now try	yāp	√	yāp	wif	√	wif	kak		tak
13	Thirteen will be	mad		maj	gōst	√	gōst	t'ha'r		zha'r
14	You should observe	bēch		bēk	thāv		sāv	must	√	must
15	Write clearly	gēm		dēm	kōf	√	kōf	yo'd	√	yo'd
16	Number 16 is	t'heb		veb	ra'g	√	ra'g	jet	√	jet
17	You may perceive	jok		jost	thip	√	thip	rēp		rēj
18	I am about to say	gaf	√	gaf	yar	√	yar	t'hēp		hēp
19	Try to hear	hus	√	hus	zhūt	√	zhūt	—		chuv
20	Please write	hiv		thit'h	kūk		tūk	t'hef		t'hesh
21	Listen carefully to	tōg	√	tōg	fung	√	fung	bās	√	bās
22	The last group is	sho't	√	sho't	t'hev		vesh	t'ho'f		shaf

s = 0.909
c = 0.735
s = 0.793

cvc = 0.491
s^3 = 0.499

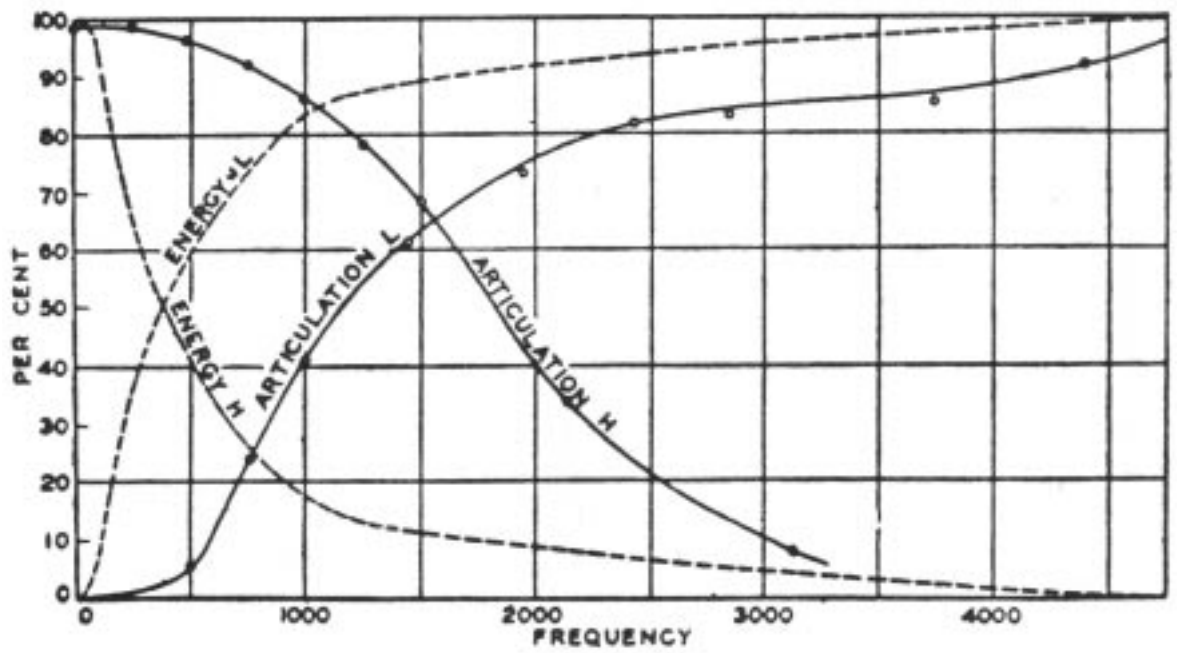


Figure 4 : This figure is reproduced from [9] and p.280 of [10].

Speech was low- and high-pass filtered with very sharp filters having a cutoff frequency defined by the abscissa. Two things were measured for each filtered speech sound, the RMS level and the articulation. The speech energy for the two filter outputs is shown by the dashed lines and the articulations are shown by the solid lines. The curve labeled "Articulation H" is the same as s_H, and the curve labeled "Articulation L" is the same as s_L. Note how the energy curves cross at the 50% point, as they should for two sharp filters. Note how the articulation curves do not cross at 50% but at 60%. Also, the frequency of the crossover is very different for energy and articulation. The equal energy point is at 450 Hz, while the equal articulation point is at 1550 Hz.

Articulation Results

- $S = \upsilon c^2$

- Error independence between bands

Articulation Index (AI)

- $(1 - s(a, c)) = (1 - s(a, b))(1 - s(b, c))$ (1)
- $\log_{10}(1 - s(a, c)) = \log_{10}(1 - s(a, b)) + \log_{10}(1 - s(b, c))$ (2)
- $AI(s) = \dfrac{\log_{10}(1 - s)}{\log_{10}(1 - s_{max})}$ (3)
- $AI(s(a, c)) = AI(s(a, b)) + AI(s(b, c)) \qquad f_a \leq f_b \leq f_c$ (4)

Underlying Density

- $AI(s(0, f_c)) = \int_0^{I_c} D(f)df$ (5)

and

- $D(f) = \frac{\partial}{\partial(f_c)} AI(s(0, f_c))$ (6)

Finally, for each of K bands,

- $D_k = \int_{f_k}^{I_{k+1}} D(f)df$ (7)

where limits chosen so all D_k are equal.

Multi-independent Channel Model

- Fletcher's articulatory band : 1mm along the basilar membrane (20 between 300 and 8000Hz)
- A single zero error band means zero error overall!!
- Robustness to a range of problems

AI and Noise

- Saturating SNR at 0 and 30 dB,

- $D_k = \frac{1}{K} SNR_k / 30$ (8)

and

- $AI = \sum_{k=1}^{K} D_k$ (9)

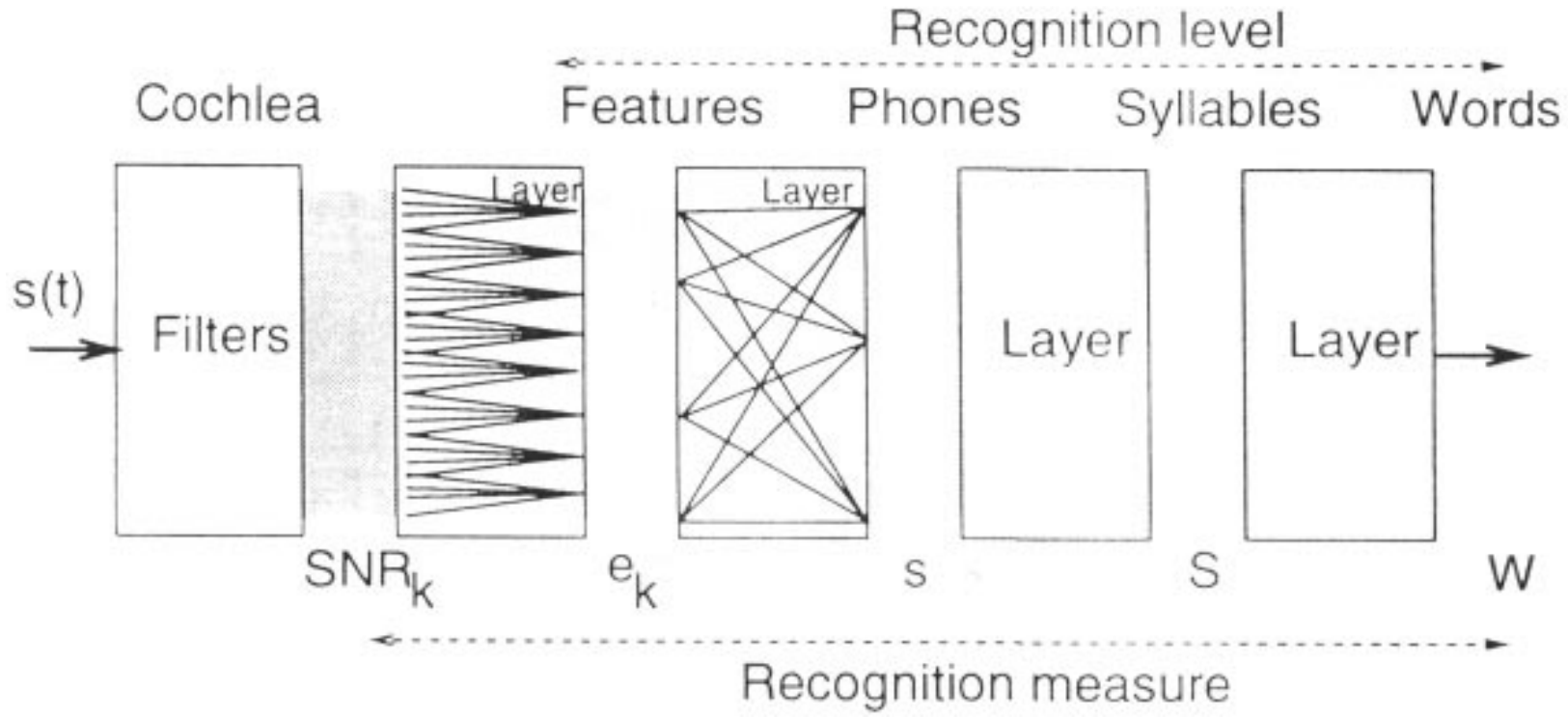


Fig.6: Hypothetical cascade of recognition layers, starting with the cochlea. The articulation measures shown at the bottom are defined in Table II. The words along the top describe the physical correlate of the measure. No feedback is assumed between layers in this oversimplified model of HSR. The first layer, the cochlea, determines the signal-to-noise ratio in about 2800 overlapping critical band channels. The next layer extracts features (i.e., partial recognition) from the speech in a local manner, as indicated by the network wiring. The output of this layer is measured in terms of the K=20 or so feature errors ek. Next, the features are mapped onto the M=20 or so phones. This process necessarily integrates across the entire tonotopic axis. Then syllables and words are formed.

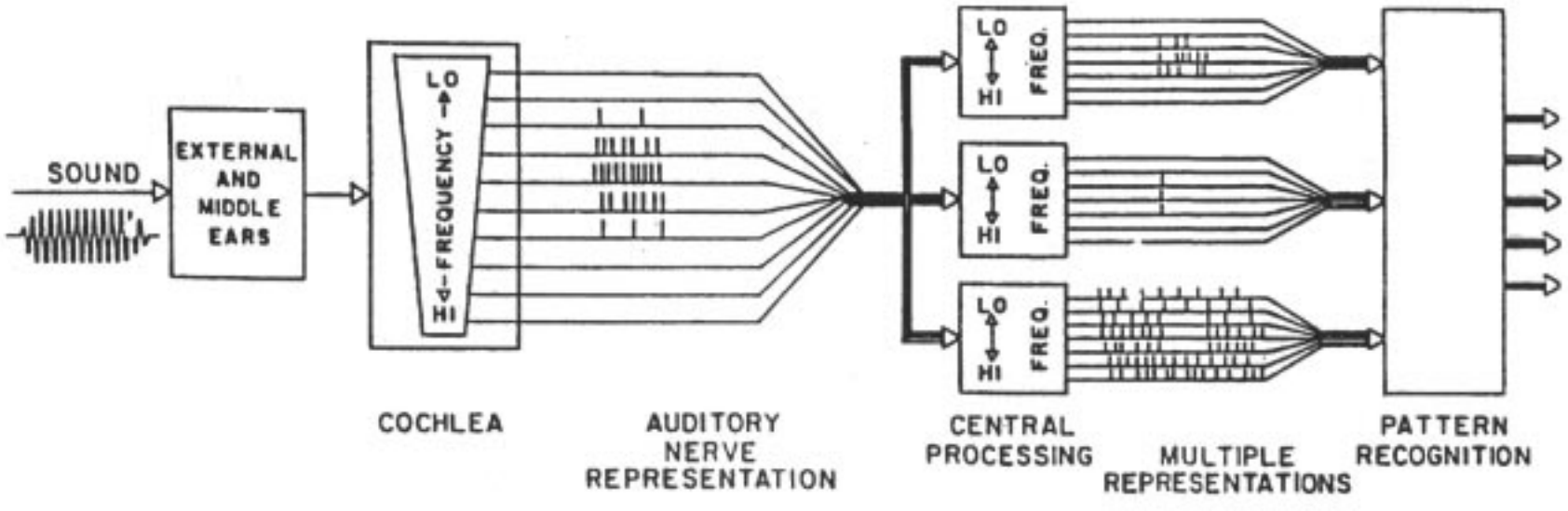
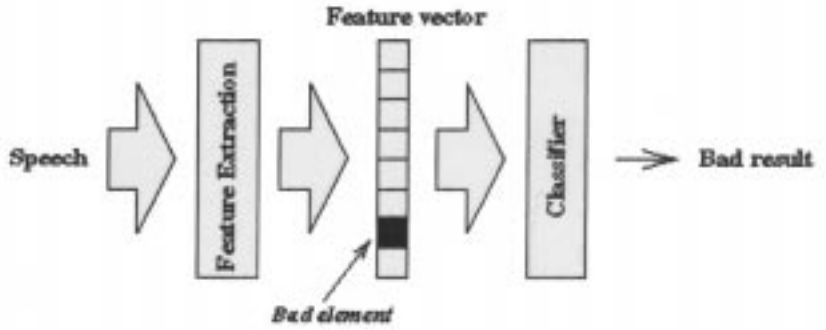
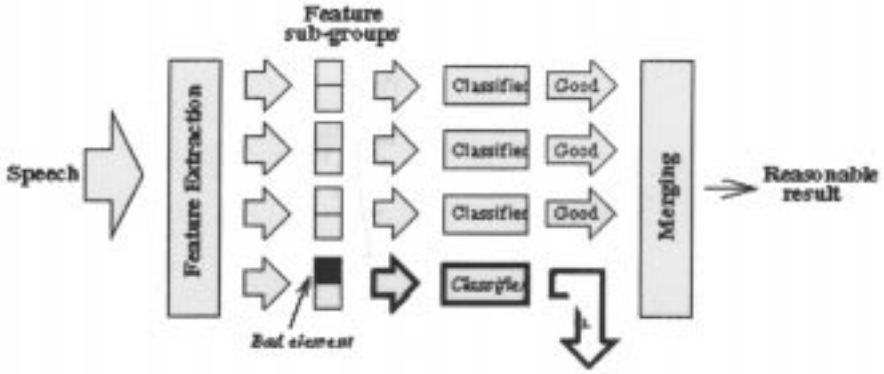


Figure 17.1 : Block diagram of sound representation in the auditory system.

Conventional ASR



Multi-band Model



Multi-band approach potentialities:

- Robustness to frequency localized degradation
- Relaxing synchrony constraint between frequency bands
- Use of different recognition strategies in different bands
- Reverberation robustness

Issues to be considered:

- definition of sub-bands
- features in each sub-band
- merging level : word, syllable, phone, state

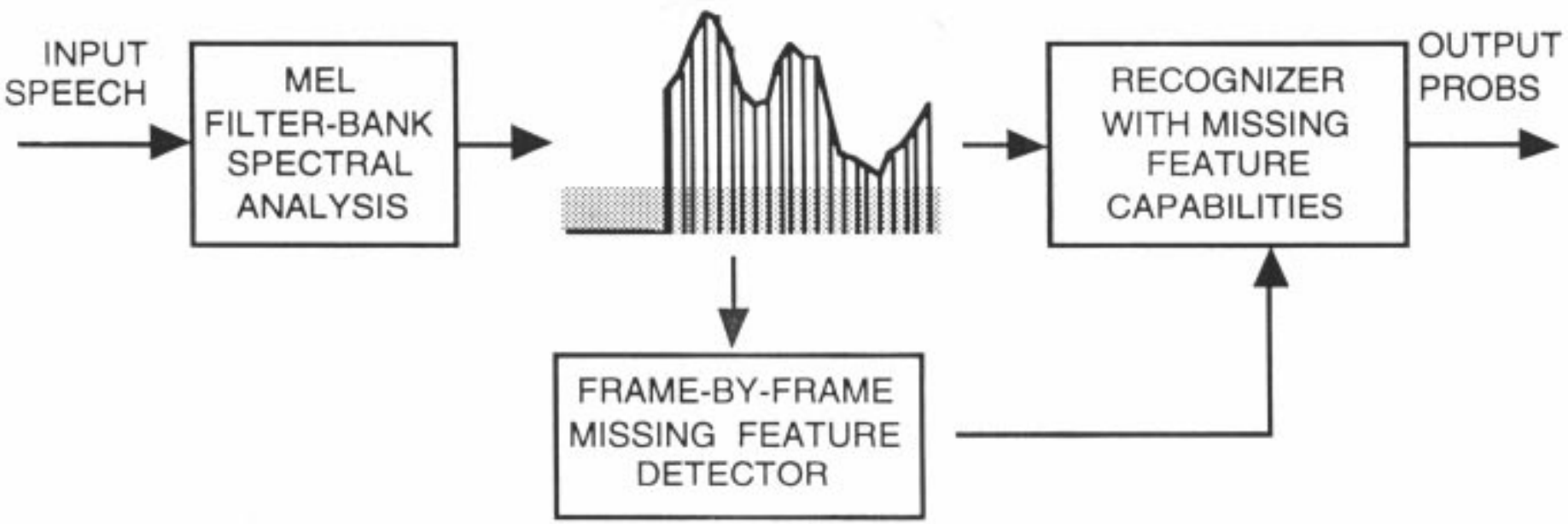
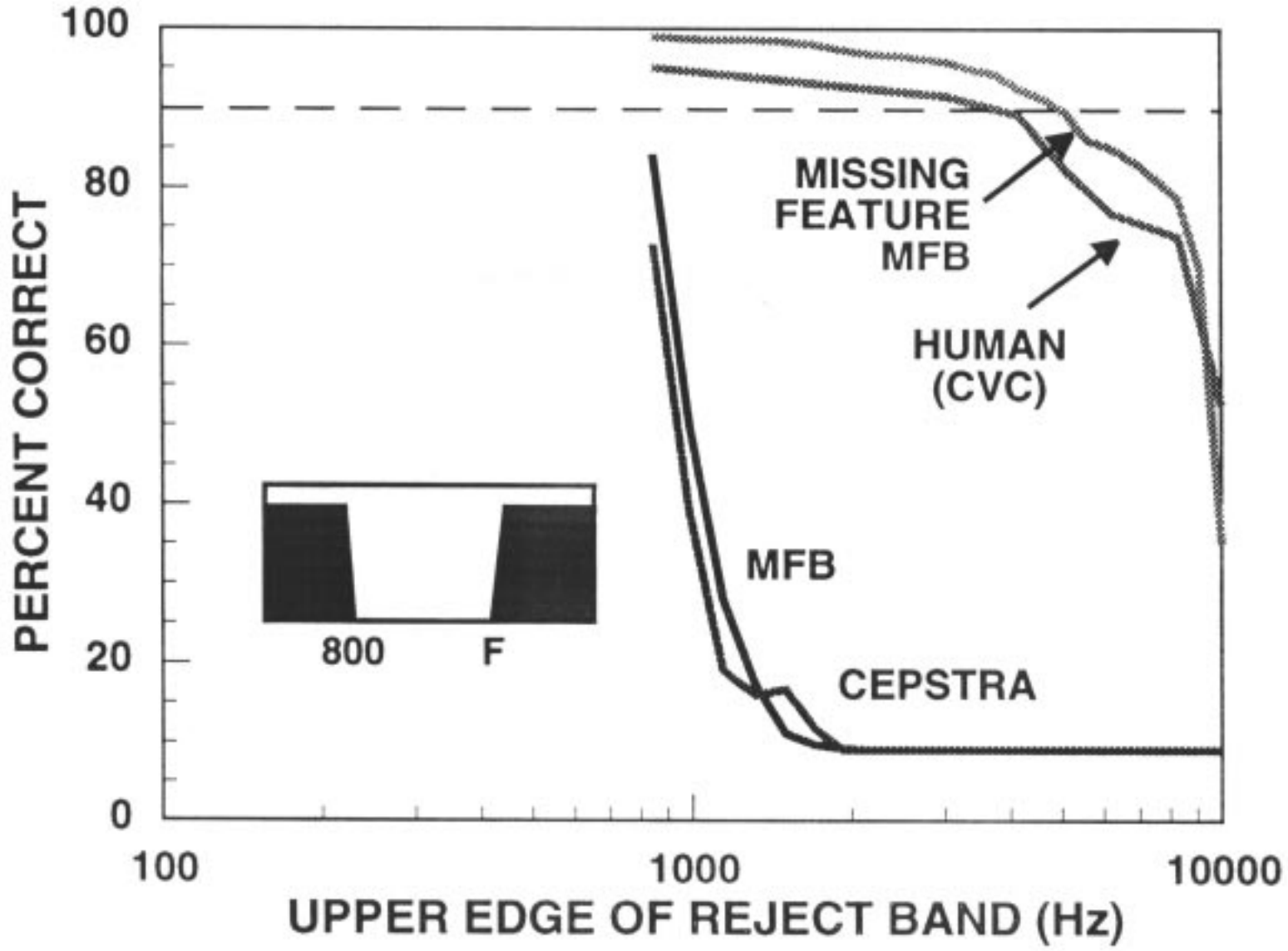


Figure 1 : Block diagram of recognizer using missing-feature adaptation.

Questions about Articulation Index

- Based on phones - the right unit for fluent speech?
- Lost correlation between distant bands?
- Lippmann experiments, disjoint bands


100
80
60
40
20
0
PERCENT CORRECT
MISSING
FEATURE
MFB
HUMAN
(CVC)
MFB
CEPSTRA
800
F
100
1000
10000
UPPER EDGE OF REJECT BAND (Hz)

HSR vs ASR
Quantitative Comparisons

- Lippmann compilation
- Range of tasks

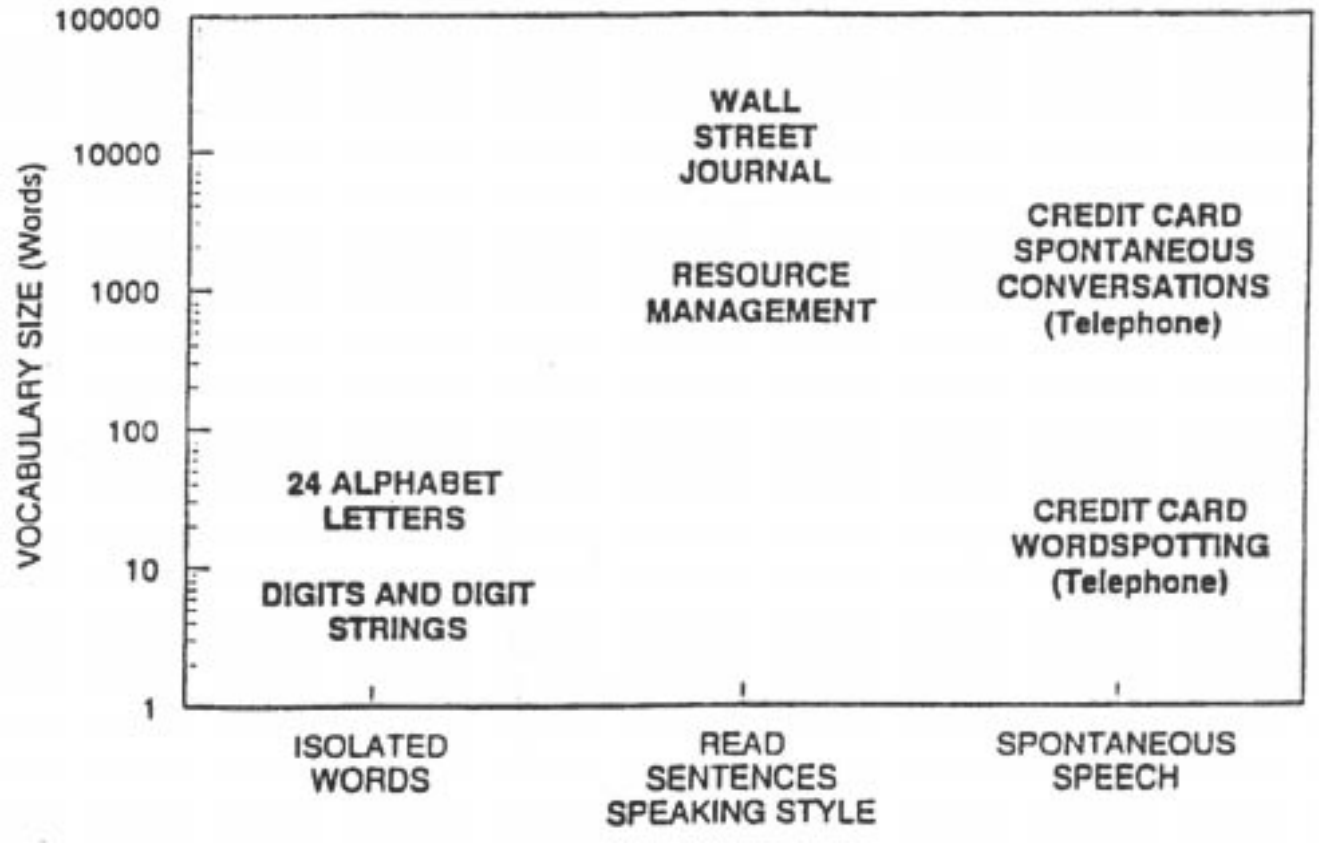
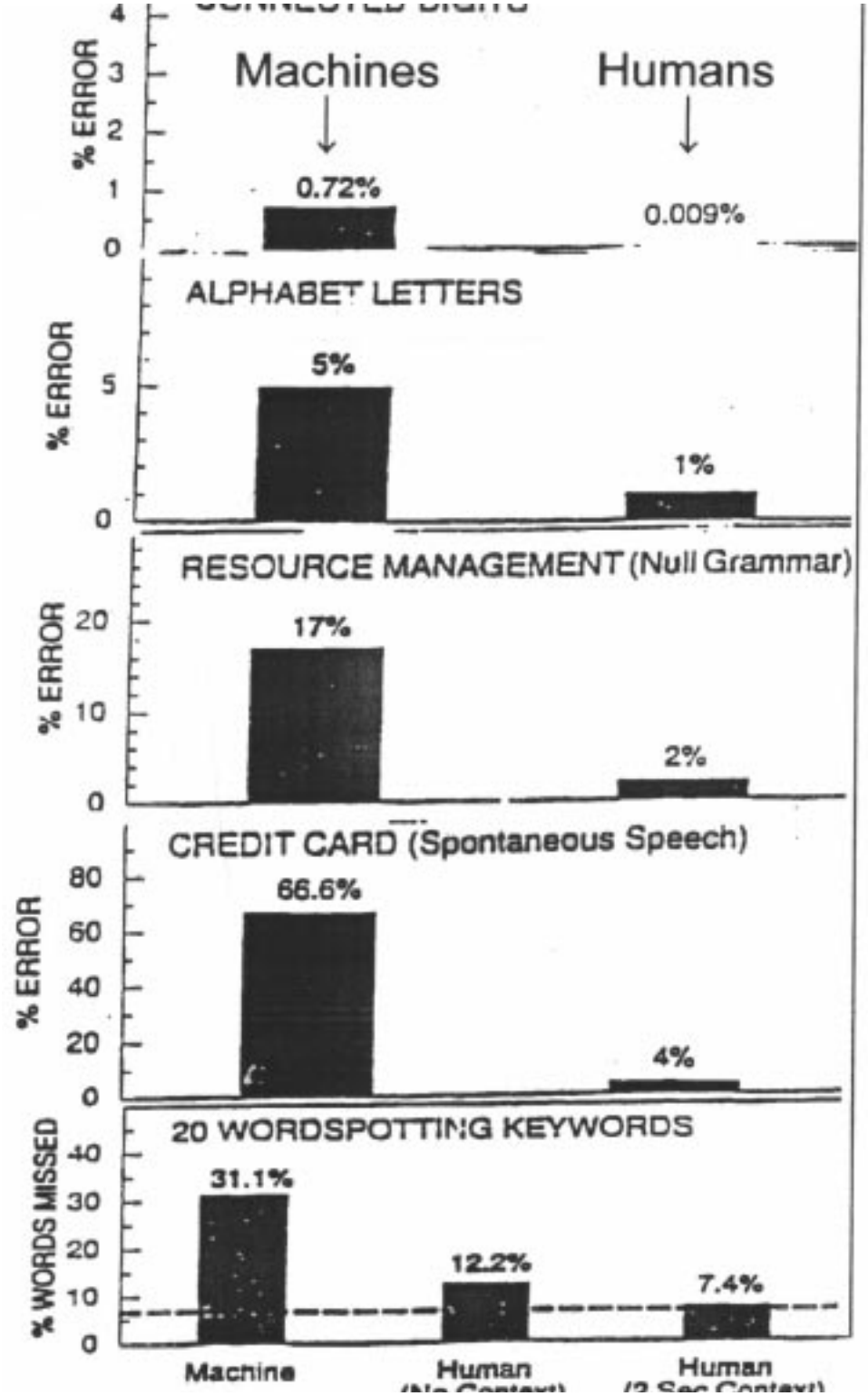


Figure 18.1 : Six speech recognition corpora.

Corpus	Description	Number of Talkers	Vocabulary Size	Number of Utterances	Total Duration	Recognition Perplexity
TI Digits	Read Digits	326	10	25,102	4 hrs	11
Alphabet Letters	Read Alphabet Letters	150	26	7,800	1 hr	26
Resource Management	Read Sentences	109	1,000	4,000	4 hrs.	60-1,000
Wall Street Journal	Read Sentences	84 - 284	5,000 - 20,000	7,200 - 37,200	12 hrs - 62 hrs	45-160
Credit-Card Continuous Speech Recognition	Spontaneous Telephone Conversations	70	2,000	35 Conversations. 1,600 Segments	2 hrs	100
Credit-Card Wordspotting	Spontaneous Telephone Conversations	70	20 Keywords	2,000 Keyword Occurrences	2 hrs	—

Figure18.2 : Characteristics of six takler-independent recogniton corpora

Figure 18.2 : Five comparisons between human and ASR Devices.



System	10dB SNR	16dB SNR	Quiet
Baseline HMM ASR	77.4%	42.2%	7.2%
ASR with noise comp	12.8%	10.0%	-
Human Listner	1.1%	1.0%	0.9%

Table 18.1: Word error rate for 5000 word Wall Street journal task using additive automotive noise

HSR vs ASR
Qualitative Comparisons

- Signal processing
- Subword recognition
- Temporal integration
- Higher levels

HSR vs ASR: Signal Processing

- Many maps versus one
- Sampled in frequency and time

  vs sampled in time (10ms)
- Some aspects of hearing already in ASR

HSR vs ASR: Subword Recognition

- Knowing what is important
- Combining it optimally

HSR vs ASR: Temporal Integration

- Using or ignoring duration
- Compensating for rapid speech
- Incorporating multiple time scales

HSR vs ASR: Higher Levels

- Syntax
- Semantics
- Pragmatics
- Getting the gist
- Dialog to learn more

Conclusions

- Under good conditions, human recognition

  much better

- Humans need to pay attention

- Some human approaches going into ASR

- Much more to do