

CECS401

Fundamentals of Spoken Language Processing

Note-2

Thursday 8/26/99

## **C. Summary of current system performances**

### **Automatic speech recognition**

<b>Task</b>	<b>Style</b>	<b>Vocabulary size</b>	<b>Word-error rate</b>
connected digit string	spontaneous	10	0.3%
airline travel information	spontaneous	2,500	2.0%
wall street journal	read	64,000	8.0%
radio broadcast	mixed	64,000	27%
switch board	conversation	10,000	38%
recorded telephone speech	conversation	-	54%

by Dr. John Makhoul, BBN, 1998.

## Speech & audio compression

Codec Type	Bit rate (per sec)
narrowband speech	6.4 KB
wideband speech	16 KB
wideband audio	64 KB
CD quality audio	96 - 128 KB

## Text-to-speech synthesis

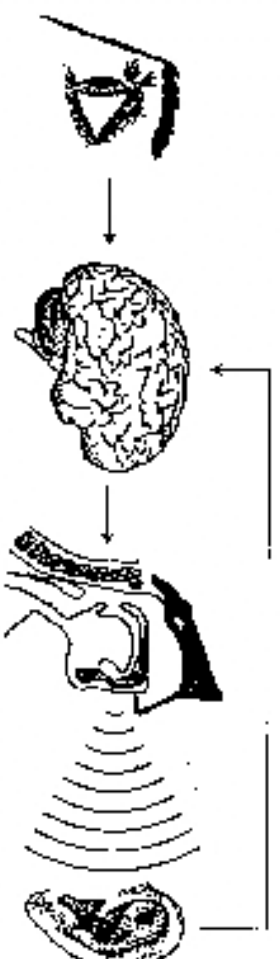
Language	Intelligibility	quality
English	high	low
Japanese, French, Scottish	high	high

by Dr. Rich Cox, 1998

## Commercial products

Look up the web, for example, <http://www.tiac.net/users/rwillcox/speech.h>

## Topic-2 Speech Production



Speaking and hearing is an interactive process.

Harvey Fletcher (1953): “We speak with our ears.”

## A. Sound

Sound: vibration waves in the frequency range of 20 to 20,000 Hz

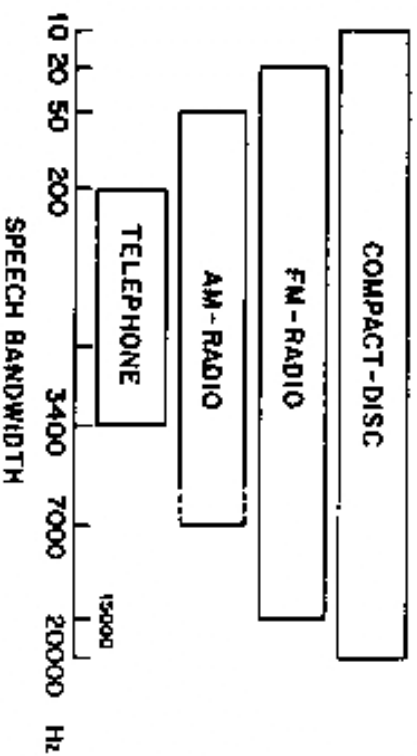
Speech: cover the frequency range of 100 — 8,000 Hz

Low frequency components (50 — 200 Hz) contribute to naturalness and presence

Mid frequency components (200 — 3,400 Hz) contribute to distinction of speech sounds

High frequency components (3,400 — 7,000 Hz) provide greater intelligibility and fricative differentiation (eg. s in sit vs. f in fit)

Four grades of audio signal bandwidth:



Sound perception by human ears:

Audible intensity (I) ranges from  $10^{-12}$  to  $10 \text{ watts}/m^2$

Reference intensity is defined as

$$I_{ref} = 10^{-12} \text{ watts}/m^2 \text{ in air}$$

Intensity level is defined as

$$IL = 10 \log (I/I_{ref}) \text{ (dB re } I_{ref})$$

Human ear is most sensitive to the frequency band of 500 Hz to 10,000 KHz.

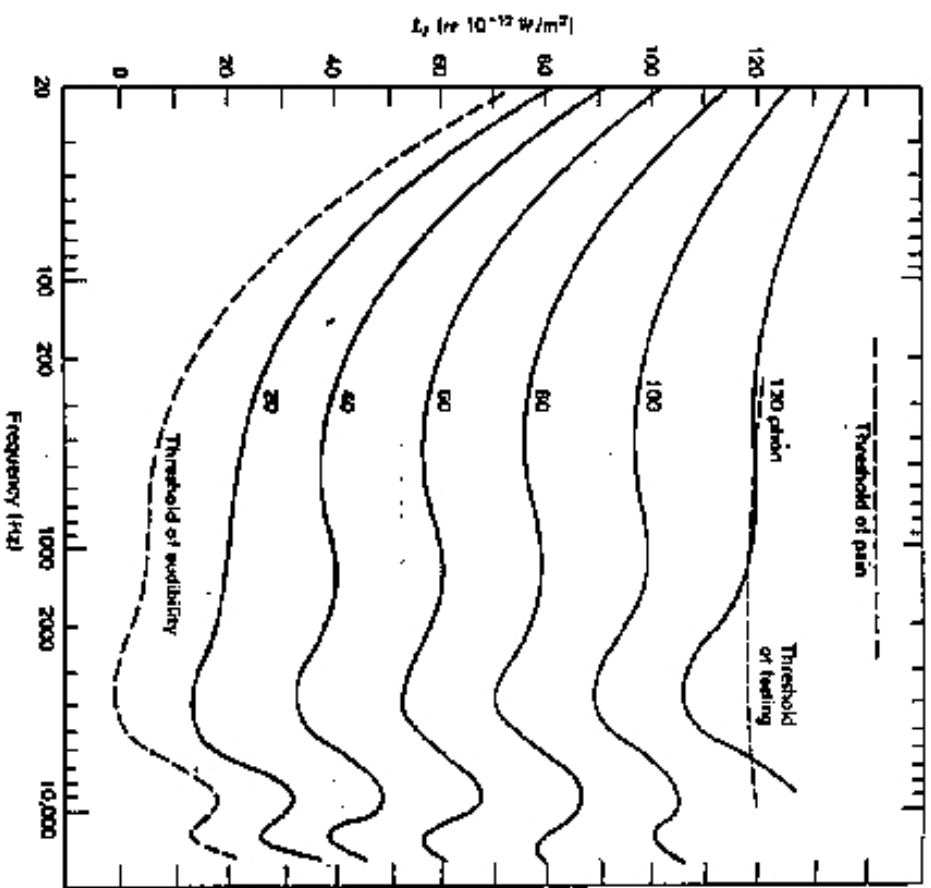
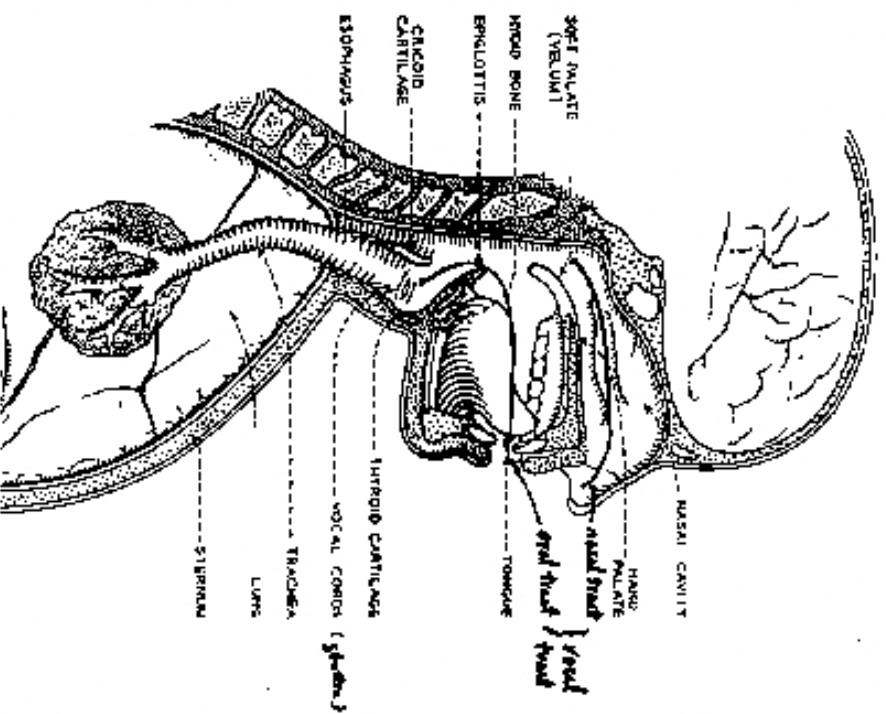


Fig. 11.9. Thresholds and free-field, equal-loudness-level contours<sup>3</sup> for pure tones with subject facing the source.

## B. Mechanism of Speech production

Cross-section of vocal apparatus of an adult



## Vocal tract:

	start	end	length	cross-section area
oral tract	glottis	lips	17 cm	0-20 cm sq.
nasal tract	velum	nostrils	12 cm	0 - 5 cm sq.

the measures are based on average adult males

### Functions of vocal apparatus:

speaking, breathing, eating

Breathing:

air  $\leftrightarrow$  nostrils  $\leftrightarrow$  nasal cavity  $\leftrightarrow$  velum port  $\leftrightarrow$  trachea

Eating:

food  $\rightarrow$  mouth  $\rightarrow$  esophagus (the gate under the epiglottis to the trachea is closed)  $\rightarrow$  stomach

## Speaking:

Producing voiced, unvoiced, and plosive sounds

*Voiced* sounds (e.g. a)

contract muscles → push air out of the lungs → vibrate vocal cords → generate quasi-periodic pulses of air → excite vocal tract → produce voiced sounds

Vibration of vocal cords:

air pressure forces the vocal cords apart → pressure between the cords is reduced to draw the two cords together → pressure is built up again to force the cords apart

The vibration frequency of the vocal cords is called fundamental frequency (F0)

	men	women	children
F0 (Hz)	70-200	150-400	200-600

The positions of the articulators (jaw, tongue, velum, lips, teeth) define the shape of vocal tract, and determine the type of voiced sounds.

Production of *unvoiced* sounds (e.g. s)

contract muscles → push air out of the lungs → form constriction in the vocal tract → generate turbulent air flow → excite vocal tract → produce unvoiced sounds

Production of *plosive* sounds (e.g. p)

form a closure in the vocal tract → build up pressure behind the closure → release the pressure abruptly

## **C. Acoustic-phonetics**

**Phoneme:**

A set of abstract symbolic units that can be used for writing a language down in a systematic and unambiguous way

**Phone:**

Acoustics realizations of phoneme

**Allophone:**

phone produced in the context of neighboring phones, for example, the phone unit *i* in *bishop* and *king*

**Syllable:**

The smallest possible unit of words, every word must contain at least one syllable (each syllable must contain a vowel with optional surrounding consonants)

**TABLE 2.1.** A condensed list of phonetic symbols for American English.

Phoneme	ARPABET	Example	Phoneme	ARPABET	Example
/ɪ/	IY	beat	/ɪ/	NX	sing
/I/	IH	bit	/ɪ/	P	<u>pe</u> t
/e/ (e')	EY	bat	/ɪ/	T	ten
/e/	EH	bet	/k/	K	kit
/ɛ/	AE	bat	/b/	B	bat
/ɛ/	AA	bat	/d/	D	bat
/ɛ/	AH	bat	/d/	H	bat
/ɔ/	AO	boat	/g/	G	get
/o/ (o')	OW	boat	/h/	H	hat
/U/	UH	boot	/f/	F	fat
/u/	UW	boot	/θ/	TH	thing
/a/	AX	about	/s/	S	sat
/ɪ/	IX	roses	/s/	SH	shut
/ɪ/	ER	bird	/v/	V	vat
/ɔ/	AXR	burd	/θ/	DH	that
/a*	AW	down	/z/	Z	zoo
/a*/	AY	buy	/z/ (zh)	ZH	azure
/ɔ/	OY	buy	/z/ (sh)	CH	church
/ɪ/	Y	you	/ʃ/ (sh, j)	JH	judge
/w/	W	wit	/w/	WH	which
/r/	R	rat	/w/	EL	warble
/l/	L	let	/r/	EM	barrel
/m/	M	met	/r/	EN	barren
/n/	N	net	/r/	DX	barren
			/r/	Q	barren (glottal stop)

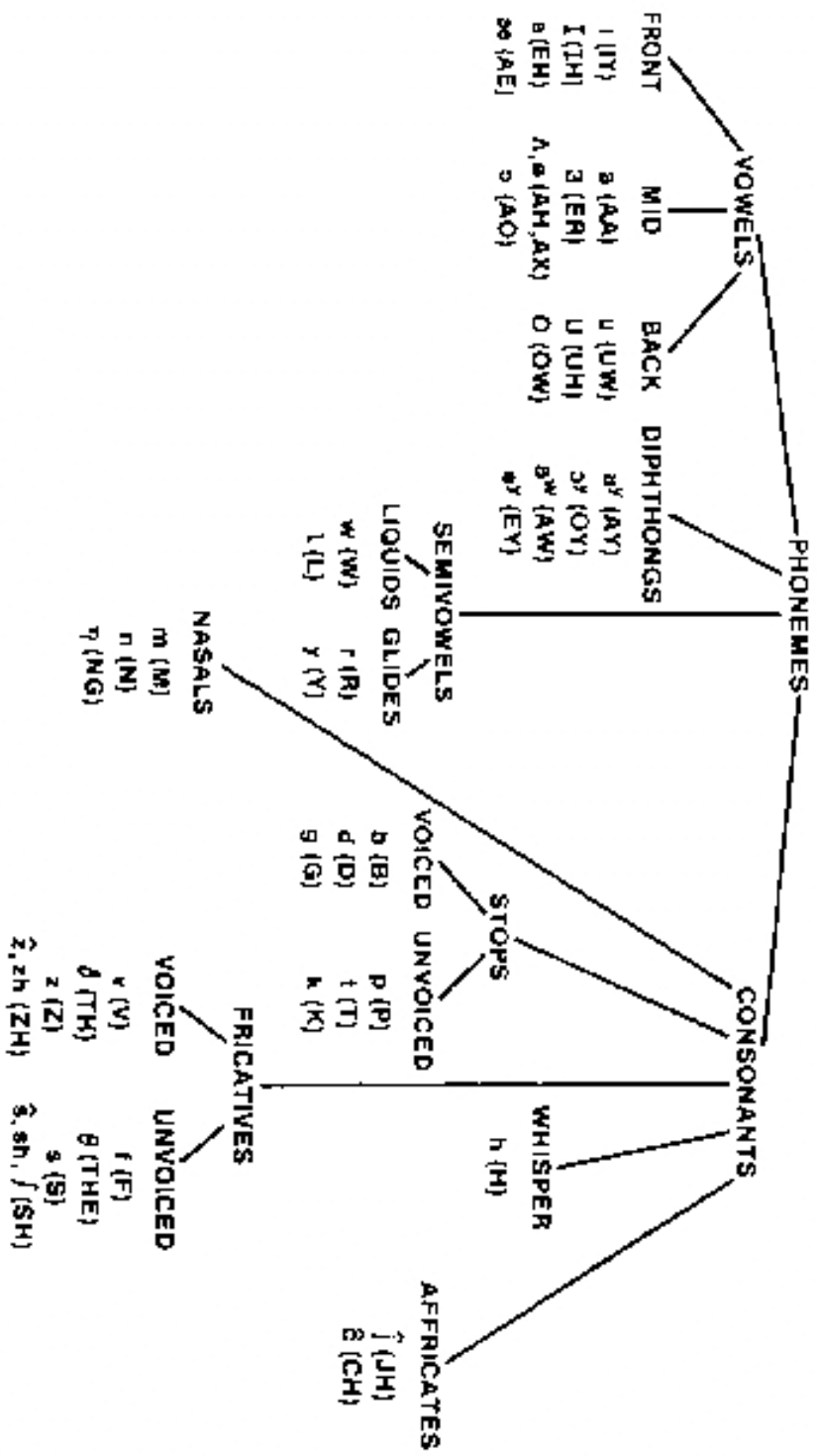


Figure 2.12 Chart of the classification of the standard phonemes of American English into broad sound classes.

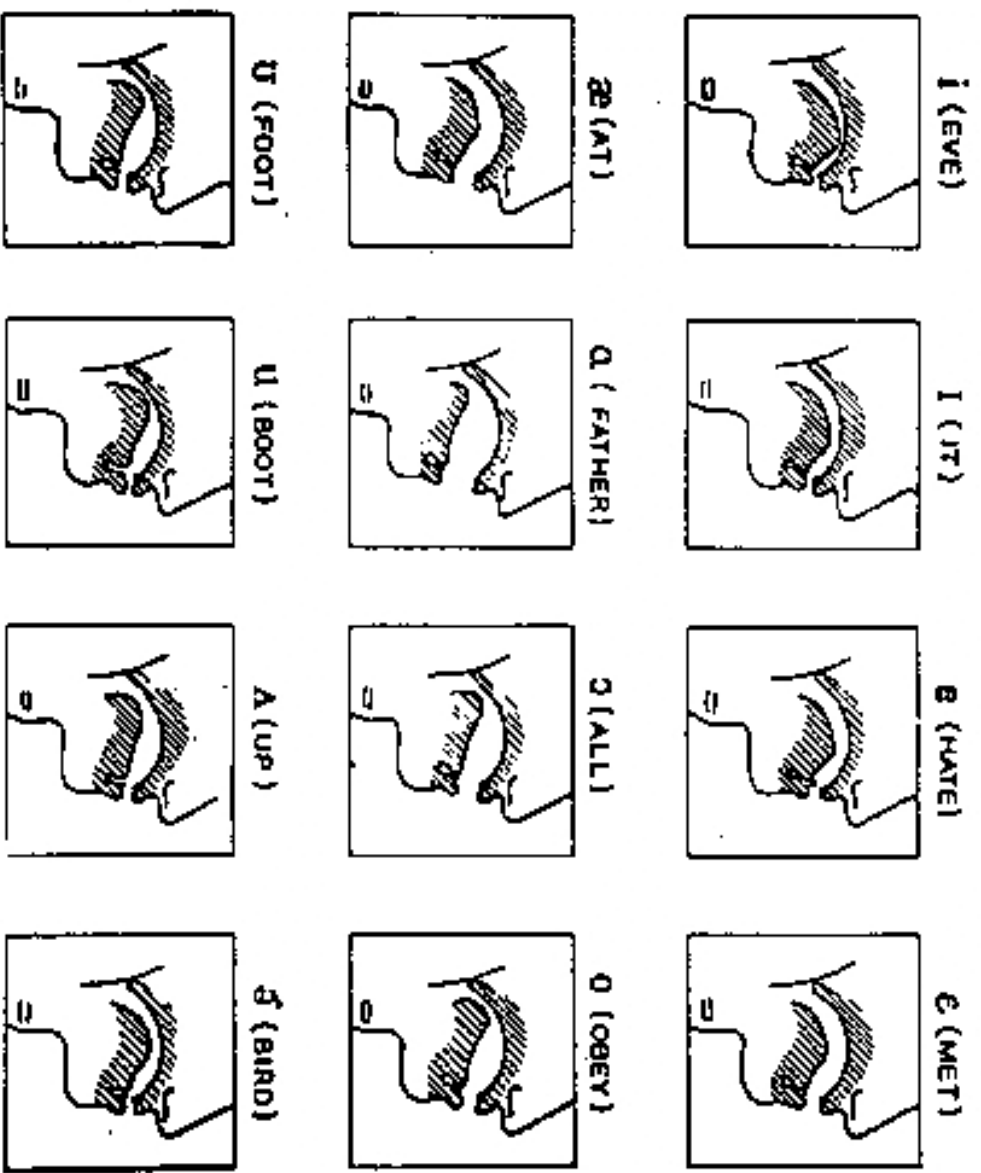


Figure 2.13 Articulatory configurations for typical vowel sounds (after Flanagan [3]).

## Diphthongs

movement from one vowel to another within a single syllable

## Semivowels

articulated in a similar way as vowel, but vocal tract is less steady and open.

## Nasal consonants

- excited by vocal cord vibration
- oral tract is closed by lips or tongue, nasal tract is open
- radiation is through nostrils

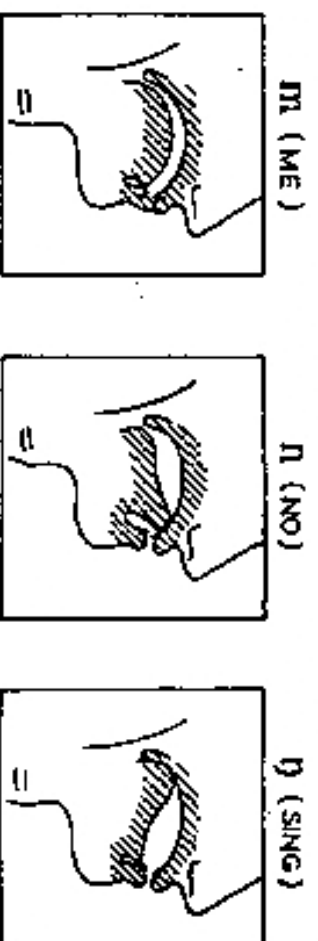


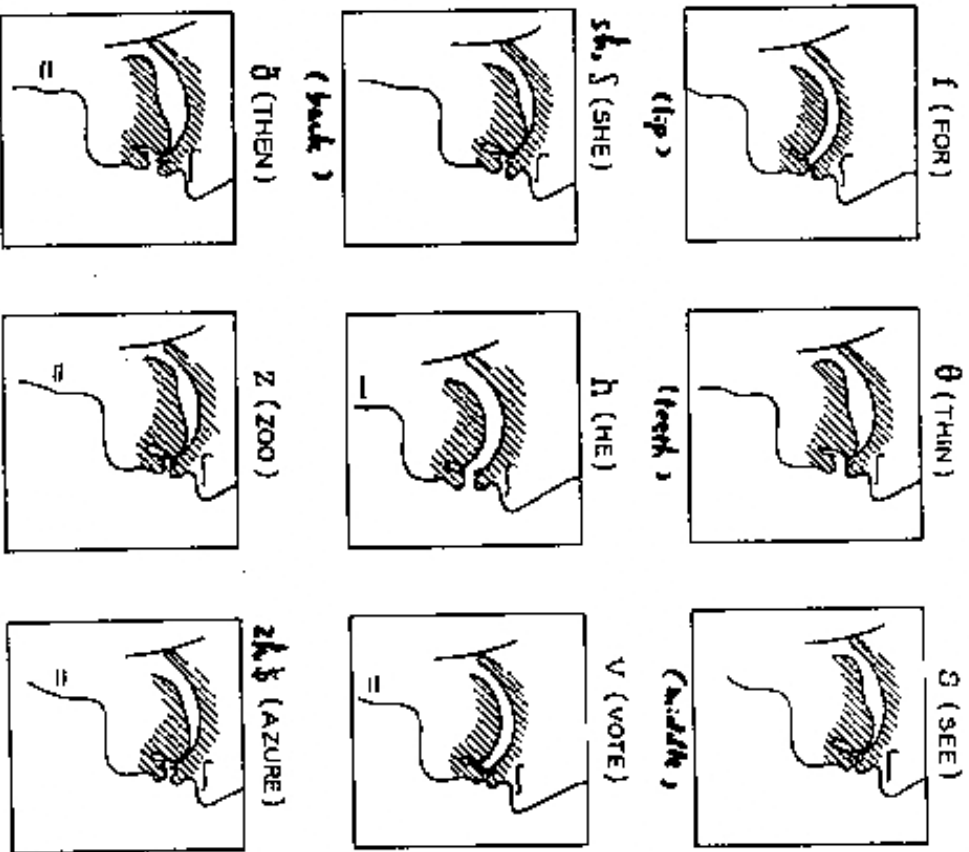
Fig. 2.8. Vocal profiles for the nasal consonants (after POTTUR, KOPP and GERSEN)

## **Unvoiced fricatives**

- excited by steady air flow
- vocal tract is constricted by tongue or teeth
- radiation is through mouth

## **Voiced fricatives**

similar to unvoiced fricatives but vocal cords are vibrating



16. Vocal tract profiles for the fricative consonants of English. The short pairs of lines drawn on the throat represent vocal cord operation (adapted from PORTER, KOPF and GRANN)

## Voiced and unvoiced stops (plosives)

- form a constriction in the oral tract and build up pressure (during this period vocal cords vibrate for voiced stops)
- constriction is formed by lips or tongue
- release pressure suddenly

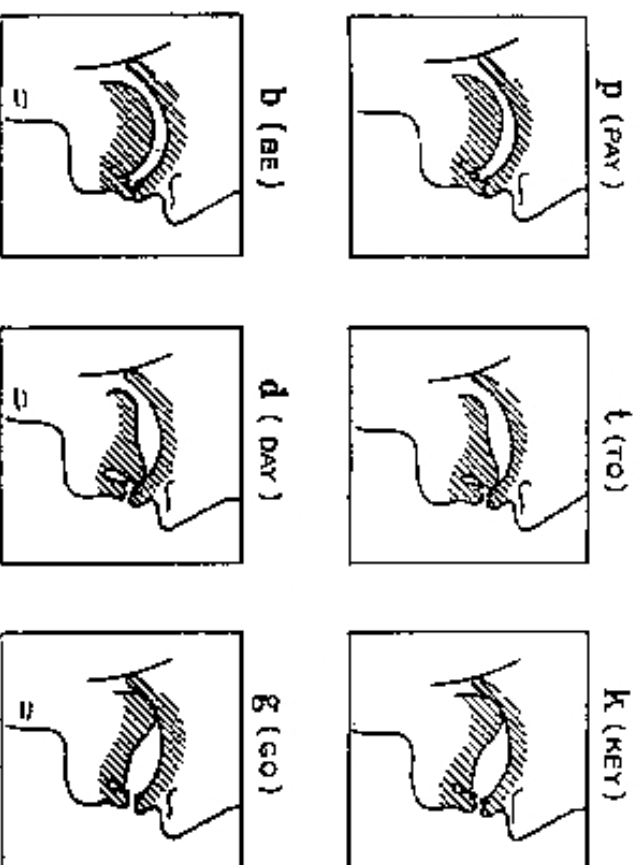


Fig. 2.7. Articulatory profiles for the English stop consonants (after Torter, Kopp and Green)

## **Whisper**

- excited by steady air flow
- vocal tract is constricted at glottis

### **Affricates**

- form a constriction in the oral tract and build up pressure
- slacken the constriction to produce turbulent air