- www.experimetrix.com/ucsd
- Click on “new.user” at the left
- Fill in information
- Choose experiments to participate in under “sign.up”

Speech Perception

LIGN 170, Lecture 2

Problems facing the listener

- Speech is rapid
  - 125-180 words per minute
  - 5-6 syllables per second
  - 25-30 phonetic segments (distinguishable pieces of sound) per second
- Speech is continuous
  - No pauses for word boundaries (Czech BBC)

Lack of invariance problem

- Production of a given phoneme is highly dependent on context
  - Coarticulation
    - Aids in rapid speech, but blurs acoustic boundaries
- Speaker variation - Sounds differ depending on the size and shape of your vocal tract
  - Normalization means we still understand them

Normalization

Normalization

?

Normalization

/i/

Outline for today:

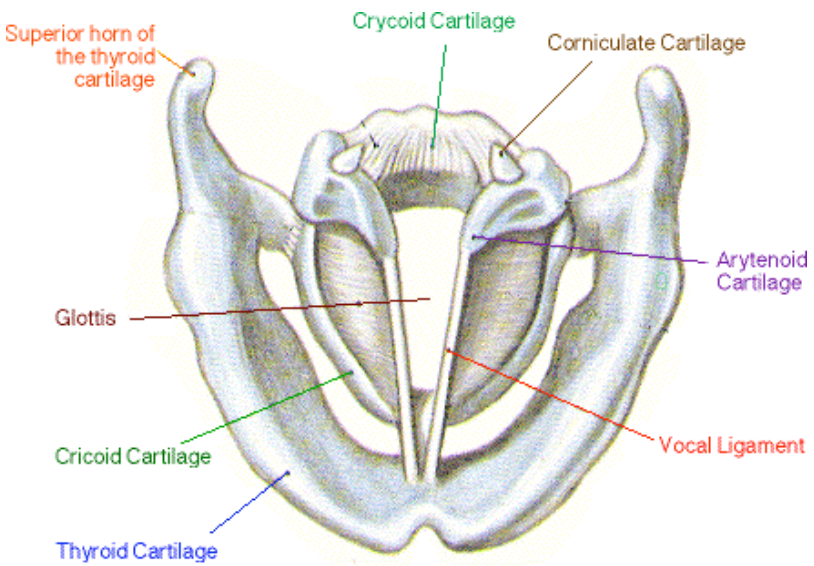
- Vocal apparatus for producing language sounds
- Acoustic properties of language sounds
- Perception of these sounds
  - What properties do we use to
    - Discriminate between sounds
    - Identify sounds

Outline for today:

- Vocal apparatus for producing language sounds
- Acoustic properties of language sounds
- Perception of these sounds
  - What properties do we use to
    - Discriminate between sounds
    - Identify of sounds

How is speech produced?

- Subglottal system
- Larynx
- Vocal Tract

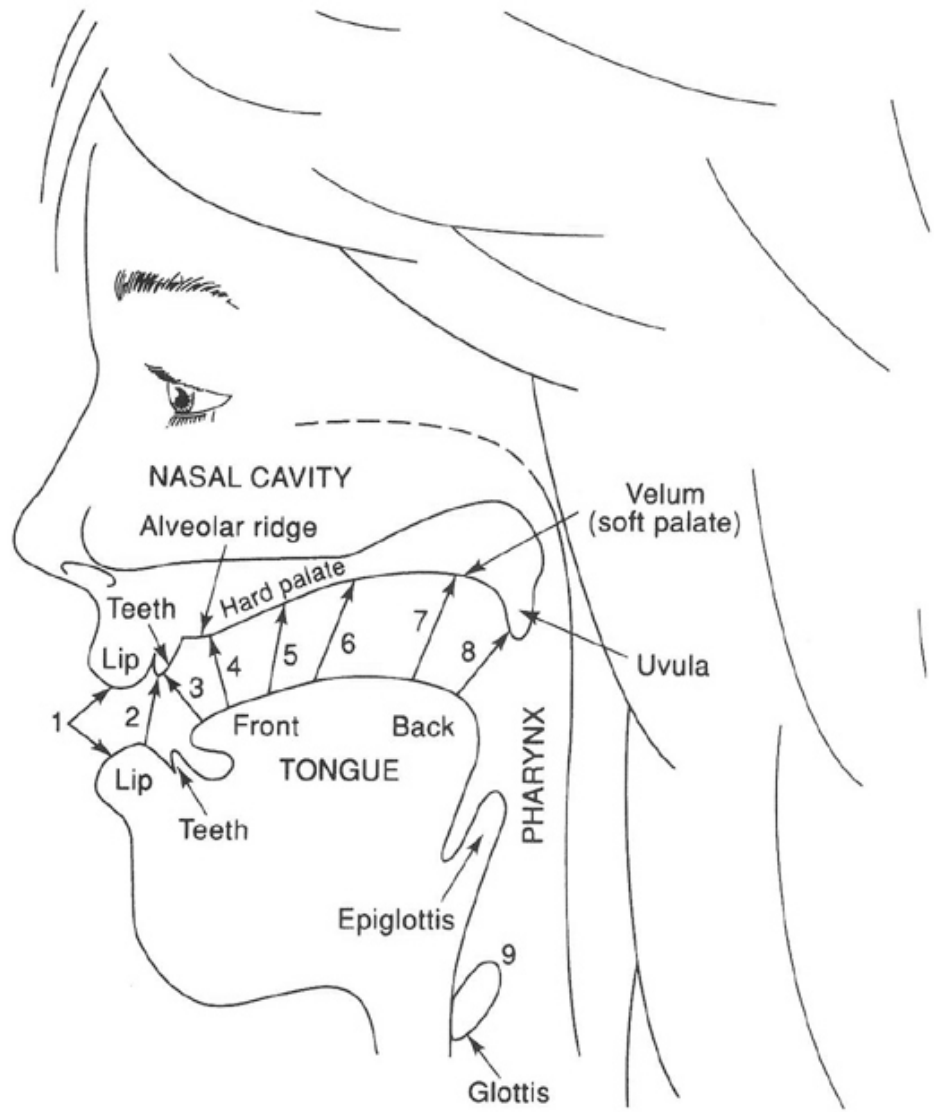


The Larynx: viewed from above

FIGURE 3.2

A schematic drawing of the vocal tract. Places of articulation: 1, bilabial; 2, labiodental; 3, dental or interdental; 4, alveolar; 5, palatoalveolar; 6, palatal; 7, velar; 8, uvular; 9, glottal.

Source: An Introduction to Language *(4th ed.) (Figure 2.1, p. 36), V. Fromkin and R. Rodman, 1988, New York: Holt, Rinehart and Winston.*



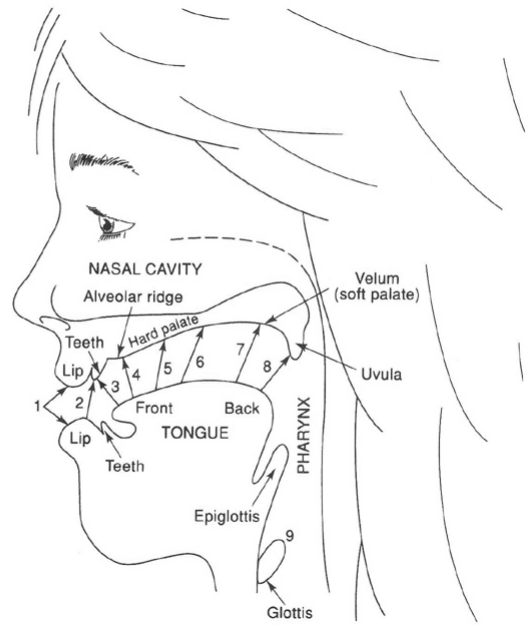
Consonants

Manner of production	Place of articulation: Bilabial	Labiodental	Interdental	Alveolar	Palatal	Velar
Oral stop voiceless	p (pin)			t (tin)		k (kin)
Oral stop voiced	b (bin)			d (din)		g (get)
Nasal stop voiced	m (map)			n (nap)		ŋ (sing)
Fricative voiceless		f (fin)	θ (thin)	s (sin)	ʃ (shin)	
Fricative voiced		v (van)	ð (than)	z (zone)	ʒ (leisure)	
Affricate voiceless					tʃ (chin)	
Affricate voiced					dʒ (gin)	
Liquid voiced				l (law) r (raw)		
Glides voiced					j (yes)	w (we)

FIGURE 3.2

A schematic drawing of the vocal tract. Places of articulation: 1, bilabial; 2, labiodental; 3, dental or interdental; 4, alveolar; 5, palatoalveolar; 6, palatal; 7, velar; 8, uvular; 9, glottal.

Source: An Introduction to Language *(4th ed.) (Figure 2.1, p. 36), V. Fromkin and R. Rodman, 1988, New York: Holt, Rinehart and Winston.*



Vowels

Position of the tongue

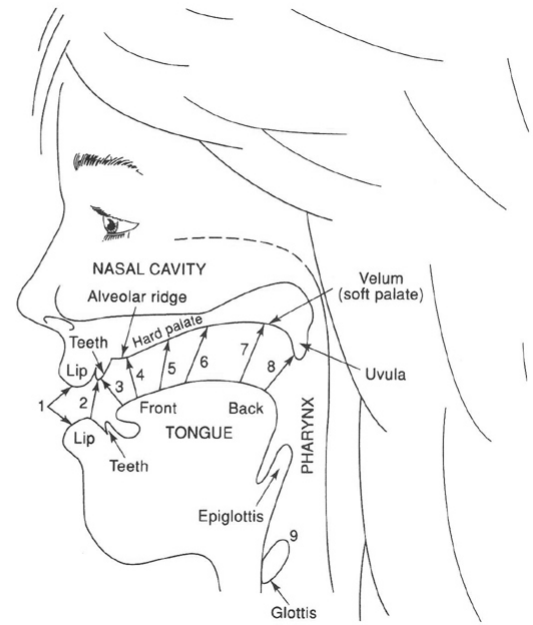


FIGURE 3.2

A schematic drawing of the vocal tract. Places of articulation: 1, bilabial; 2, labiodental; 3, dental or interdental; 4, alveolar; 5, palatoalveolar; 6, palatal; 7, velar; 8, uvular; 9, glottal.

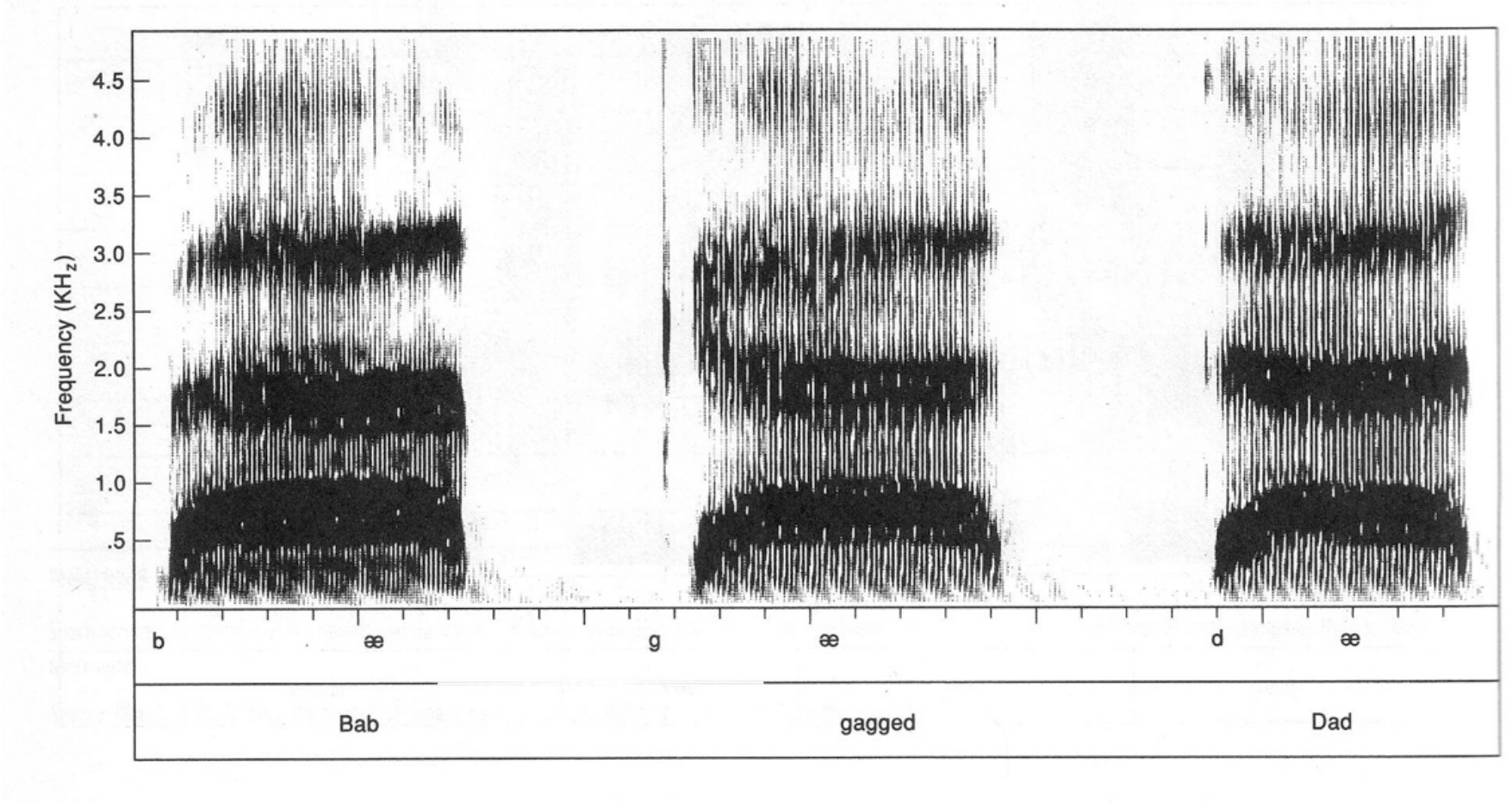
Source: An Introduction to Language *(4th ed.) (Figure 2.1, p. 36), V. Fromkin and R. Rodman, 1988, New York: Holt, Rinehart and Winston.*

Vowels	Front	Back
high	i	u
	I	U
mid	e ə	o
	ɛ ʌ	ɔ
low	æ	a

Outline for today:

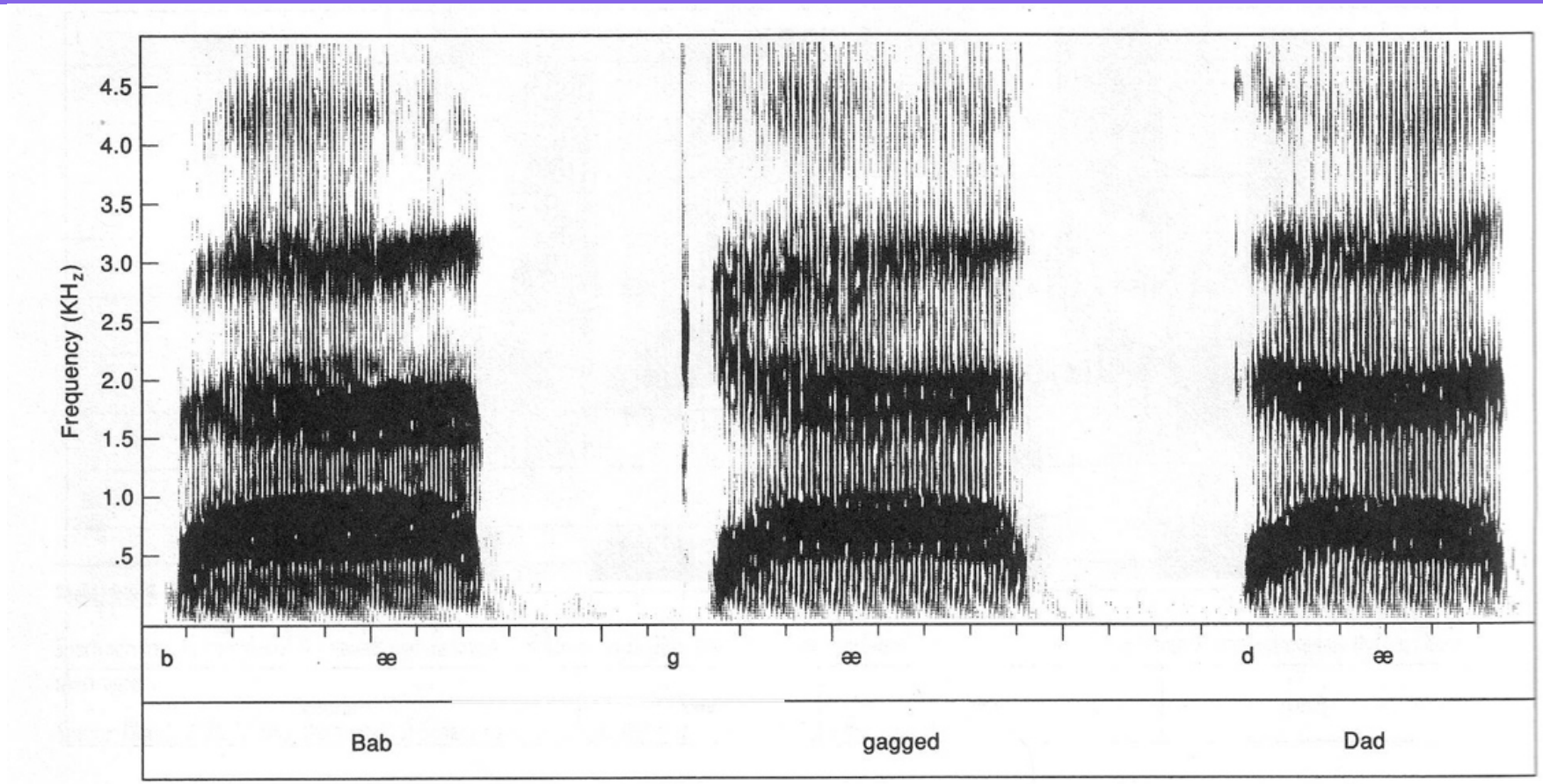
- Vocal apparatus for producing language sounds
- Acoustic properties of language sounds
- Perception of these sounds
  - What properties do we use to
    - Discriminate between sounds
    - Identify sounds

Acoustic properties of speech



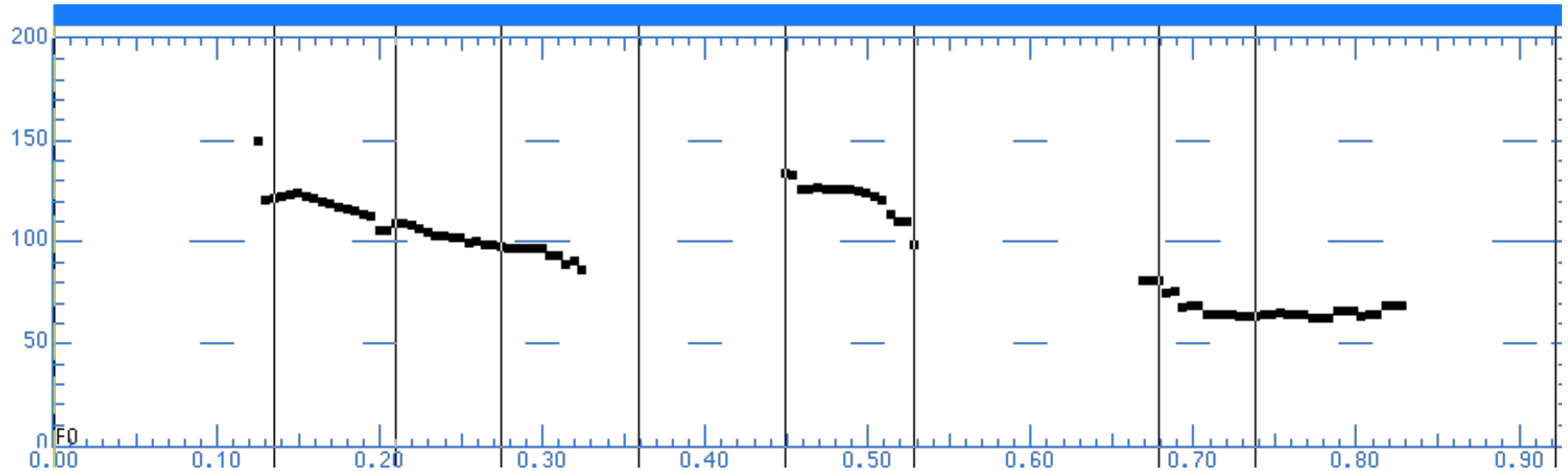
- Fundamental Frequency (F0)
- Formants (F1, F2, F3...)

Acoustic properties of speech

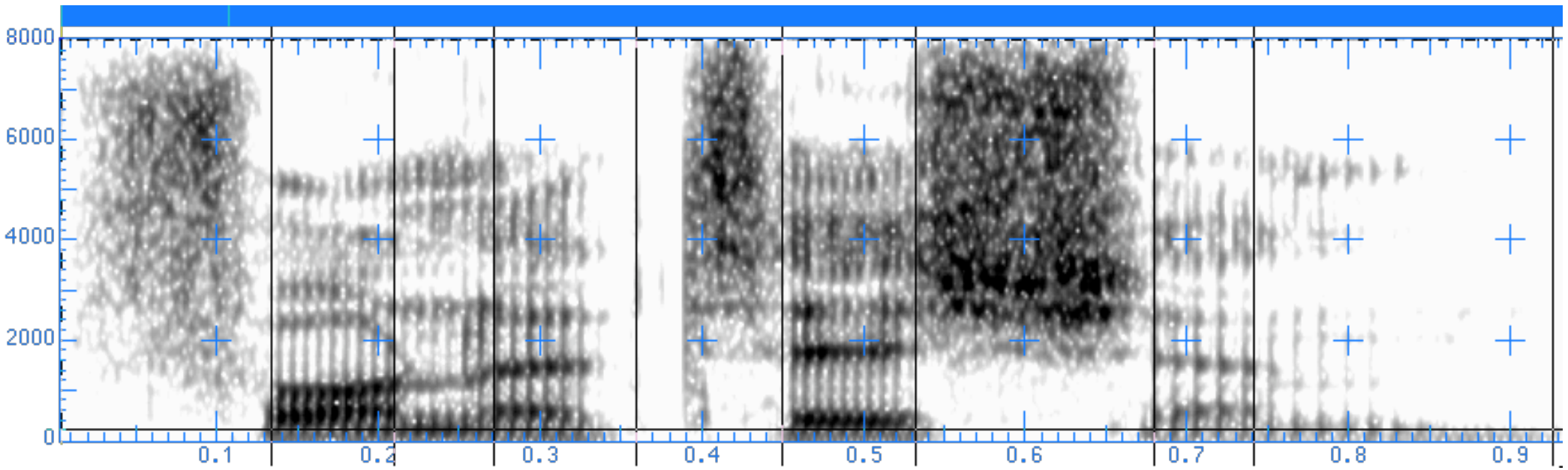


- Fundamental Frequency (F0)
- Formants

“Phonetician”

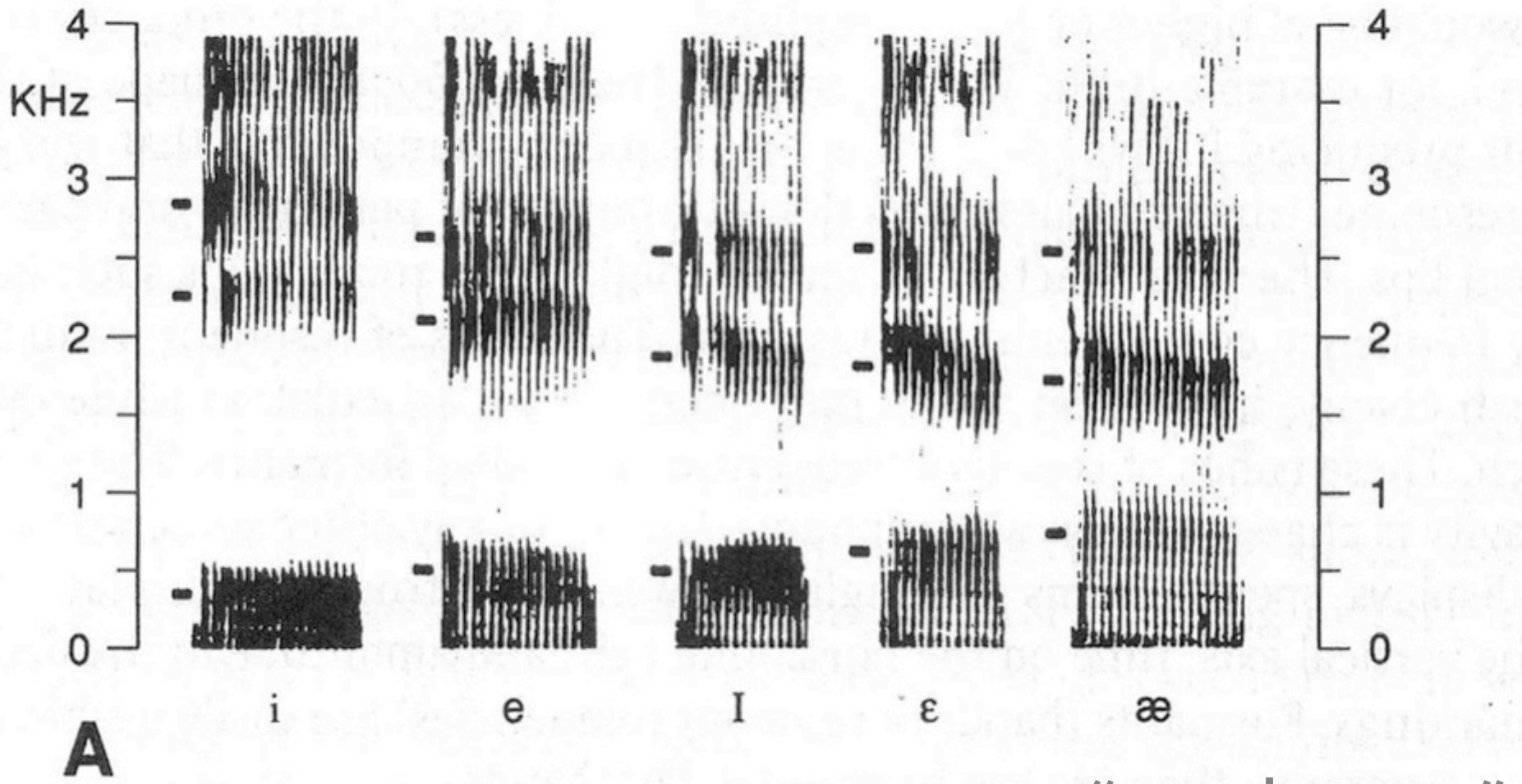


F0 - Pitch extraction



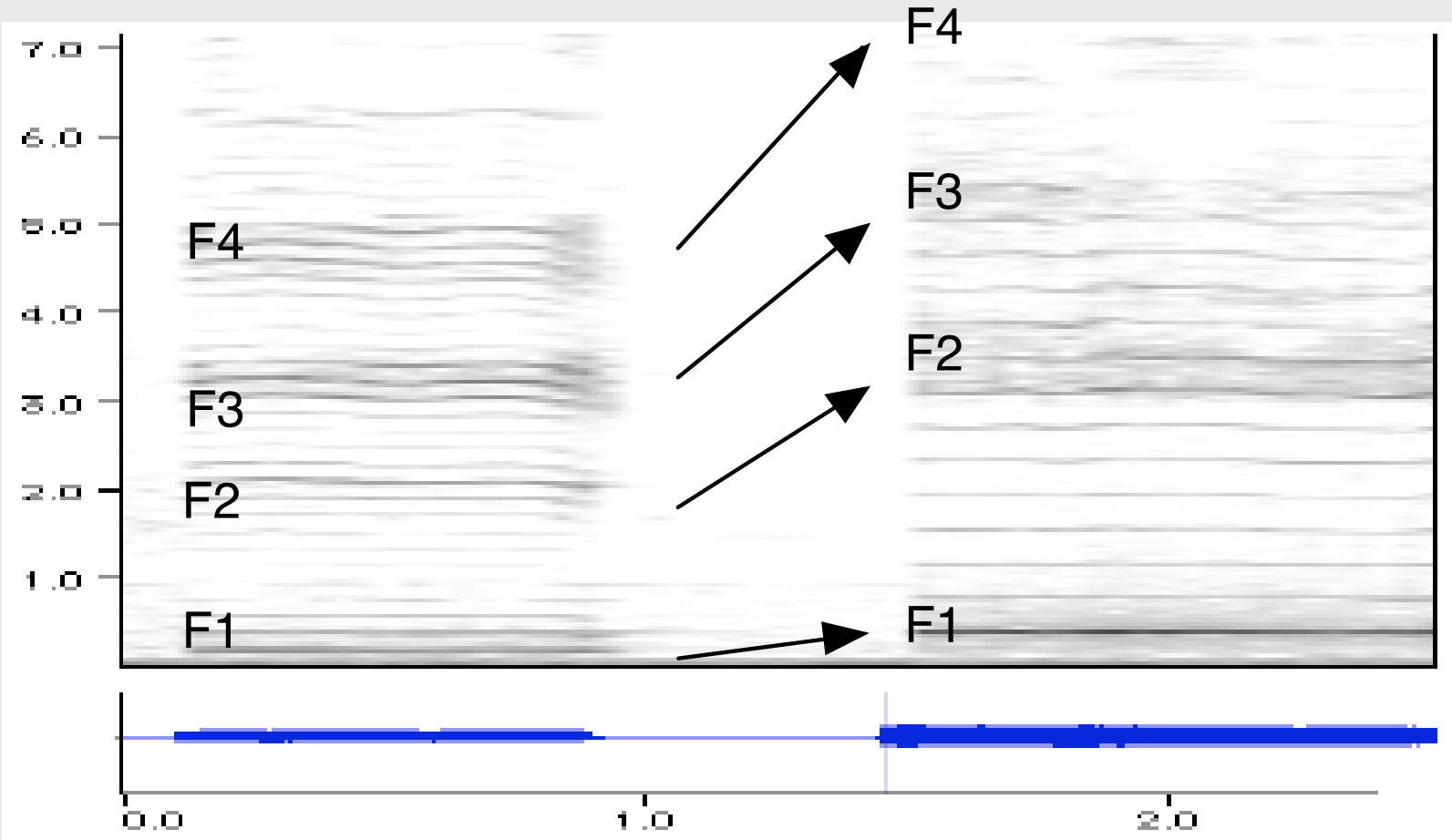
Spectrogram

Acoustic properties of vowels



"steady state"

Normalization



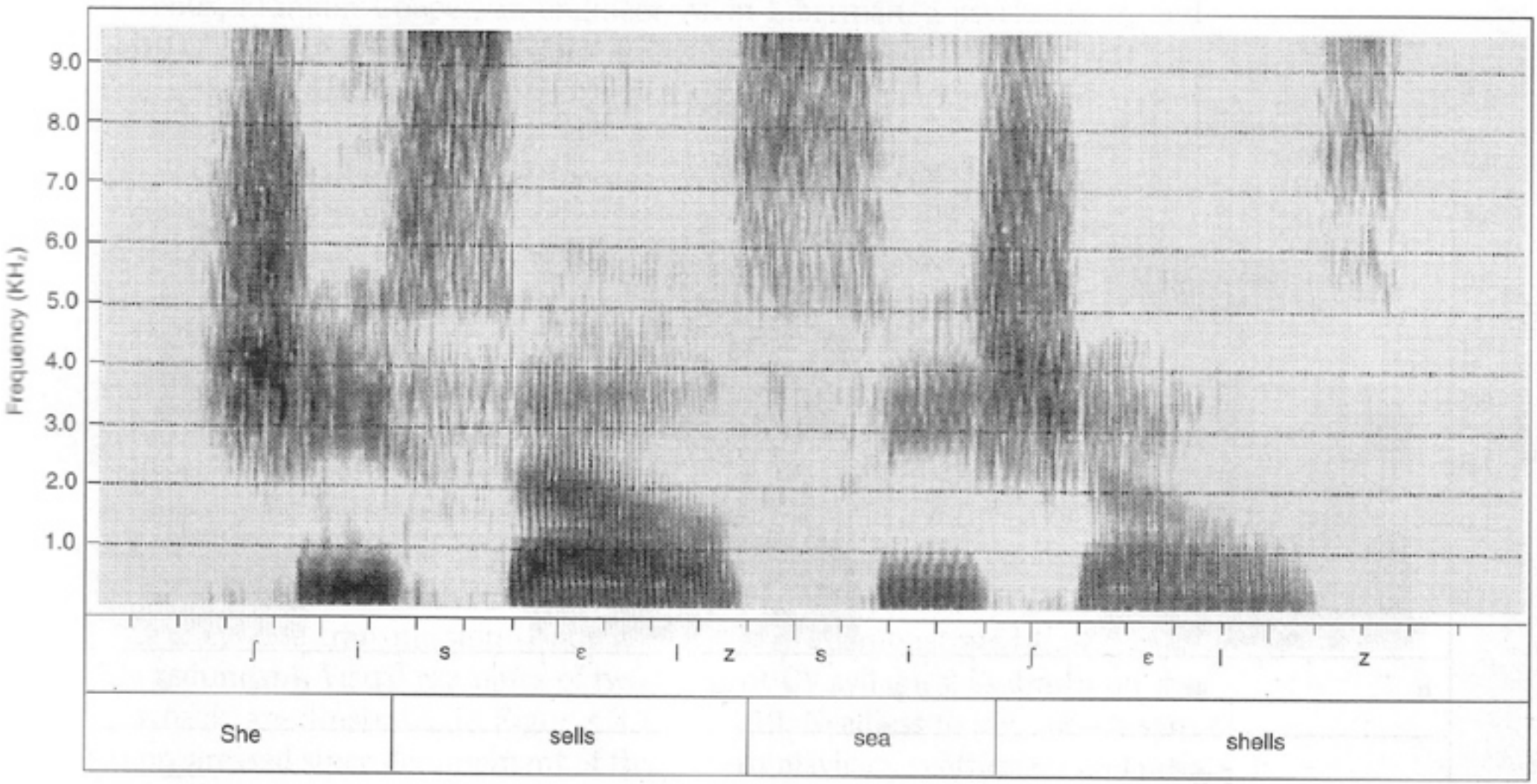
/i/

Speaker 1 adult male (Vocal tract length: 17cm)

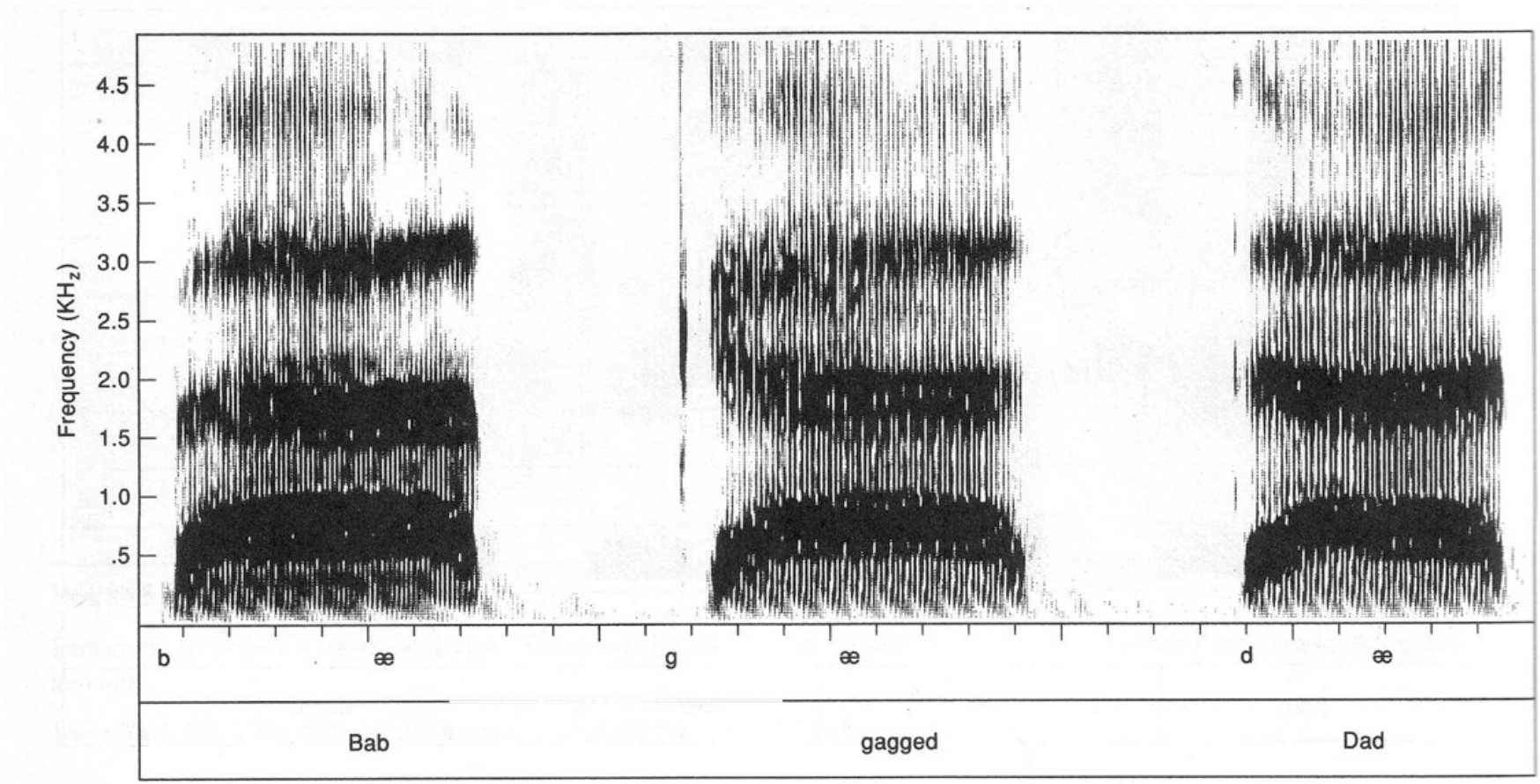
Speaker 2 child female (Vocal tract length: 9cm)

- Normalization: Listeners can also extract information about speaker's vocal tract length from previous exposure and use it to identify the vowel in a subsequent word
  - Tricking this normalization process:
    - Sentence from speaker with long vocal tract: "Now I will say 'but'."
    - insert "but" from medium vocal tract:
      - heard as "bat"
    - insert "bat", heard as "bit"

Consonants


Frequency (KHz)
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
ʃ i s ɛ l z s i ʃ ɛ l z
She
sells
sea
shells

Coarticulation



Outline for today:

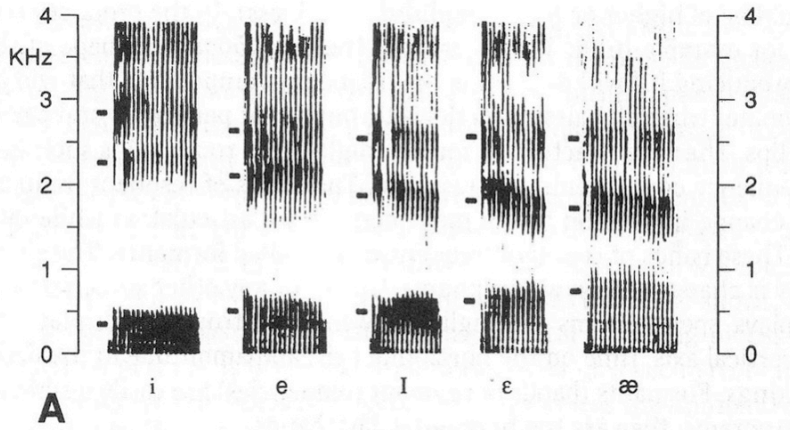
- Vocal apparatus for producing language sounds
- Acoustic properties of language sounds
- Perception of these sounds
  - What properties do we use to
    - Discriminate between sounds
    - Identify sounds

Studies of speech perception

- Methods
  - Artificial speech
  - Take existing speech and alter it
- What acoustic cues are necessary for perception?
  - Discriminate between two sounds
  - Correctly identify sounds

Vowel discrimination

- In isolation (steady state), formant spacing gives reliable cues for vowels



- During connected speech other acoustic cues exist: duration and transitions between different sounds
  - How do these cues interact?
  - Which ones are most important?

- Jenkins, Strange & Edman (1983)
- CVC syllables with /b/ and different Vs

  /bab/ /bib/ /baeb/ /bIb/ /baib/ ...
- Altered syllables to create 5 conditions:

			Steady-state	Duration	Transition
Unmodified	[A][B][C]	/bab/	Y	Y	Y
Silent V	[A][ ][C]	/b_b/	N	Y	Y
Vowel only	[][B][]	/a/	Y	Y	N
Fixed Vowel	[][B][]	/a/	Y	N	N
Cs abutted	[A][C]	/bb/	N	N	Y

			Steady-state	Duration	Transition
Unmodified	[A][B][C]	/bab/	Y	Y	Y
Silent V	[A][ ][C]	/b_b/	N	Y	Y
Vowel only	[][B][]	/a/	Y	Y	N
Fixed Vowel	[][B][]	/a/	Y	N	N
Cs abutted	[A][C]	/bb/	N	N	Y

- Identification task accuracy:

			Steady-state	Duration	Transition
Unmodified	[A][B][C]	/bab/	Y	Y	Y
Silent V	[A][ ][C]	/b_b/	N	Y	Y
Vowel only	[][B][]	/a/	Y	Y	N
Fixed Vowel	[][B][]	/a/	Y	N	N
Cs abutted	[A][C]	/bb/	N	N	Y

- Identification task accuracy:
  - Silent V = Unmodified

			Steady-state	Duration	Transition
Unmodified	[A][B][C]	/bab/	Y	Y	Y
Silent V	[A][ ][C]	/b_b/	N	Y	Y
Vowel only	[][B][]	/a/	Y	Y	N
Fixed Vowel	[][B][]	/a/	Y	N	N
Cs abutted	[A][C]	/bb/	N	N	Y

- Identification task accuracy:
  - Silent V = Unmodified

			Steady-state	Duration	Transition
Unmodified	[A][B][C]	/bab/	Y	Y	Y
Silent V	[A][ ][C]	/b_b/	N	Y	Y
Vowel only	[][B][]	/a/	Y	Y	N
Fixed Vowel	[][B][]	/a/	Y	N	N
Cs abutted	[A][C]	/bb/	N	N	Y

- Identification task accuracy:
  - Silent V = Unmodified
  - More errors for Vowel only & Cs abutted

			Steady-state	Duration	Transition
Unmodified	[A][B][C]	/bab/	Y	Y	Y
Silent V	[A][ ][C]	/b_b/	N	Y	Y
Vowel only	[][B][]	/a/	Y	Y	N
Fixed Vowel	[][B][]	/a/	Y	N	N
Cs abutted	[A][C]	/bb/	N	N	Y

- Identification task accuracy:
  - Silent V = Unmodified
  - More errors for Vowel only & Cs abutted
  - Most errors for Fixed Vowel

			Steady-state	Duration	Transition
Unmodified	[A][B][C]	/bab/	Y	Y	Y
Silent V	[A][ ][C]	/b_b/	N	Y	Y
Vowel only	[][B][]	/a/	Y	Y	N
Fixed Vowel	[][B][]	/a/	Y	N	N
Cs abutted	[A][C]	/bb/	N	N	Y

- Identification task accuracy:
  - Silent V = Unmodified
  - More errors for Vowel only & Cs abutted
  - Most errors for Fixed Vowel
- Steady-state information alone not enough

			Steady-state	Duration	Transition
Unmodified	[A][B][C]	/bab/	Y	Y	Y
Silent V	[A][ ][C]	/b_b/	N	Y	Y
Vowel only	[][B][]	/a/	Y	Y	N
Fixed Vowel	[][B][]	/a/	Y	N	N
Cs abutted	[A][C]	/bb/	N	N	Y

- Identification task accuracy:
  - Silent V = Unmodified
  - More errors for Vowel only & Cs abutted
  - Most errors for Fixed Vowel
- Steady-state information alone not enough

- Jenkins, Strange & Edman (1983)
- Steady-state information alone not enough
- Transition and duration information also relevant

			Steady-state	Duration	Transition
Unmodified	[A][B][C]	/bab/	Y	Y	Y
Silent V	[A][ ][C]	/b_b/	N	Y	Y
Vowel only	[][B][]	/a/	Y	Y	N
Fixed Vowel	[][B][]	/a/	Y	N	N
Cs abutted	[A][C]	/bb/	N	N	Y

Consonants

- Not as accurately perceived in general
  - Quieter, shorter than vowels
- Stops (/b/, /p/, /k/ ...)
  - Lose identity in isolation
  - Must have formant transitions into vowel
  - Encoded: CV merged or fused for stops
  - Acoustic properties change based on context

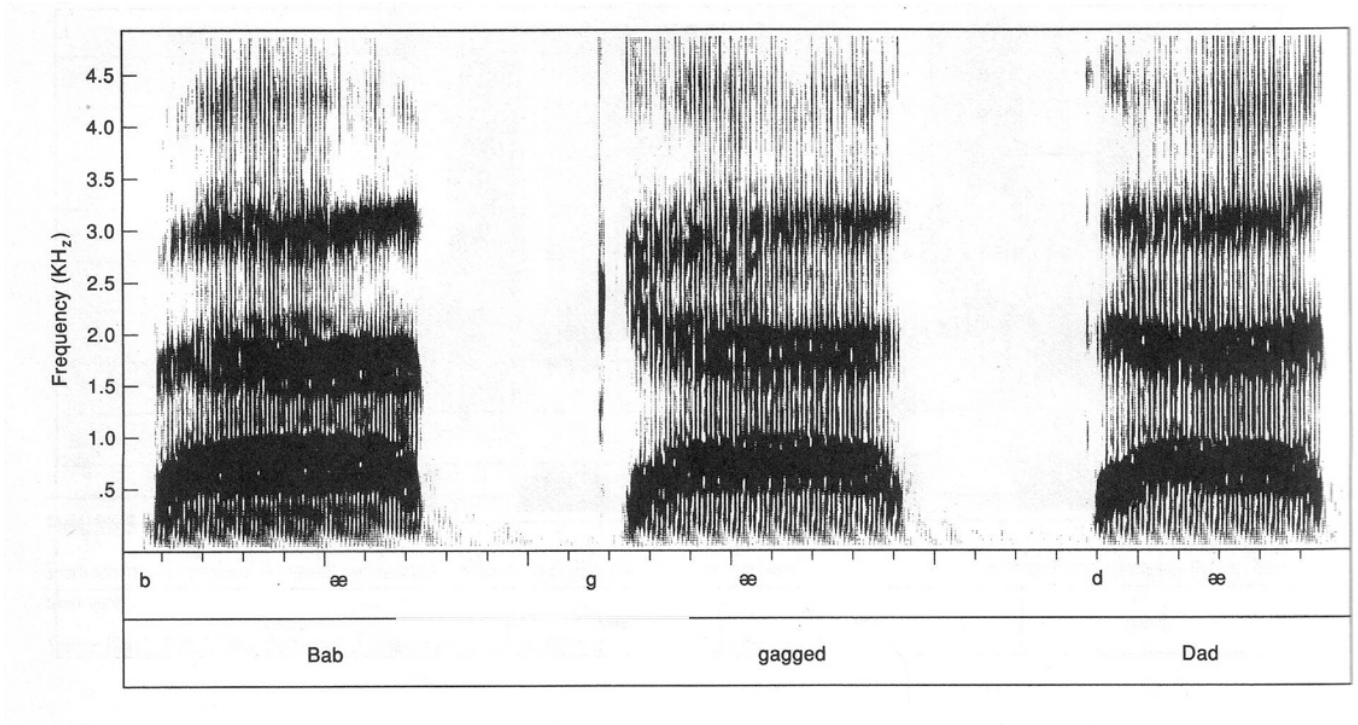
Voice-onset-time (VOT)

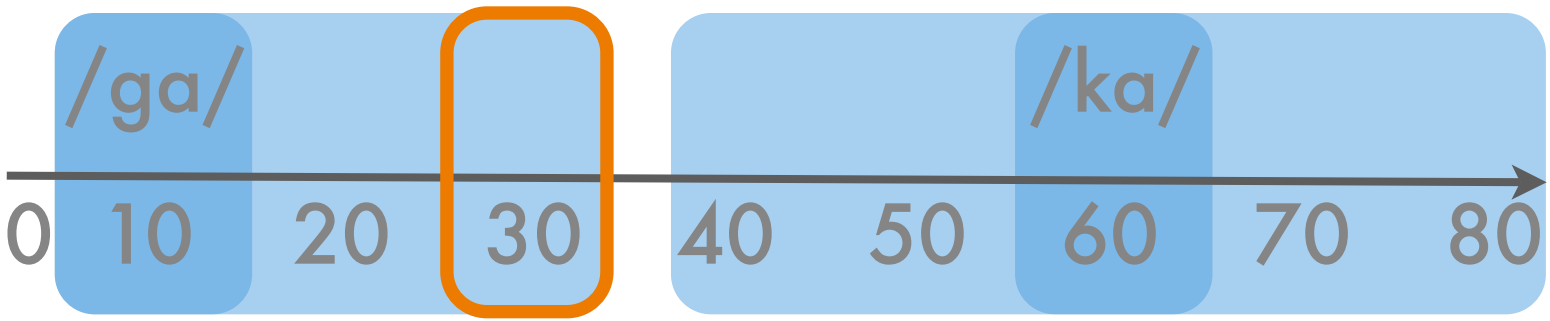
- Best single measure for distinguishing voiced vs. unvoiced consonants

/b/ -- /p/ /d/ -- /t/ /k/ -- /g/

- VOT: time between the release of air and the onset of the voicing for the adjacent vowel

Voice-onset-time (VOT)

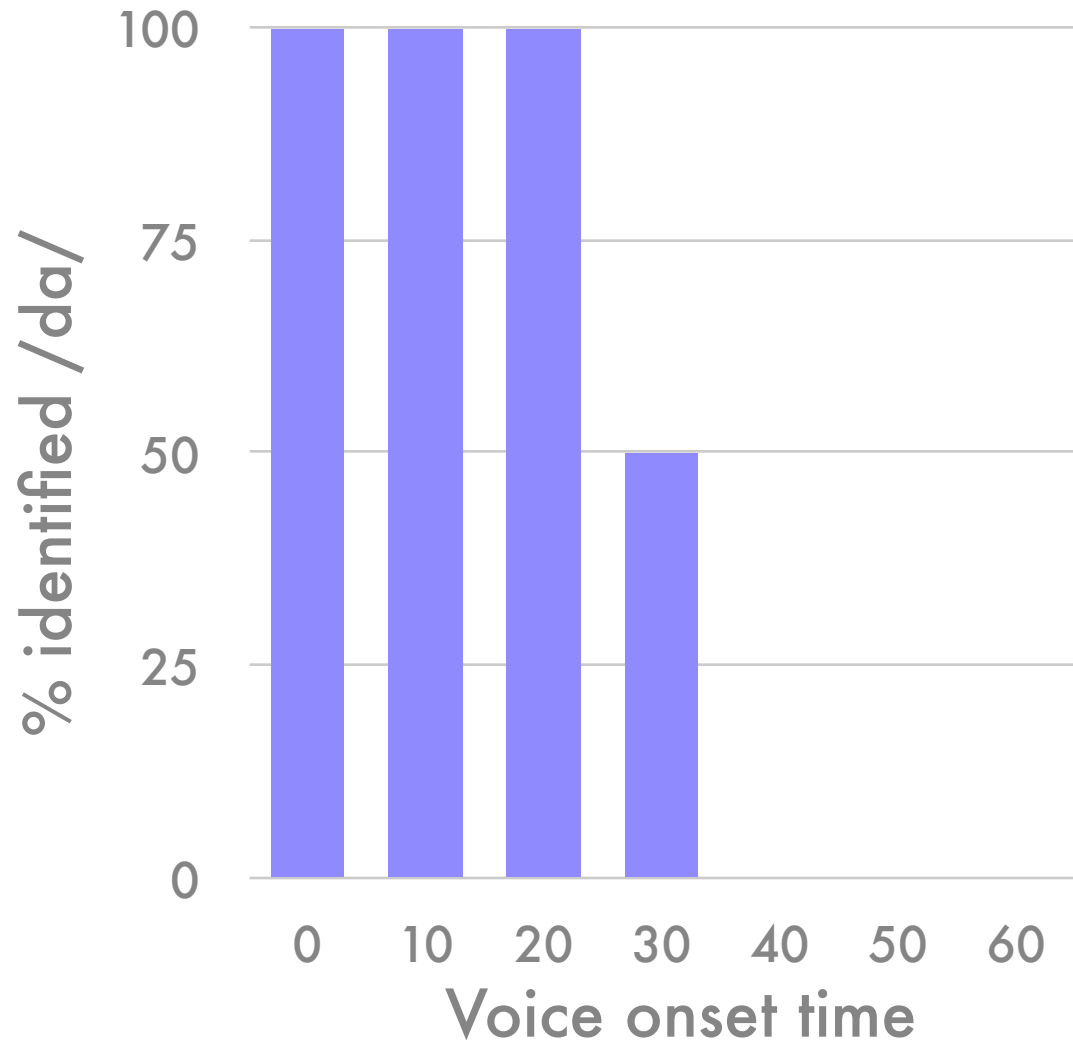




VOT in milliseconds

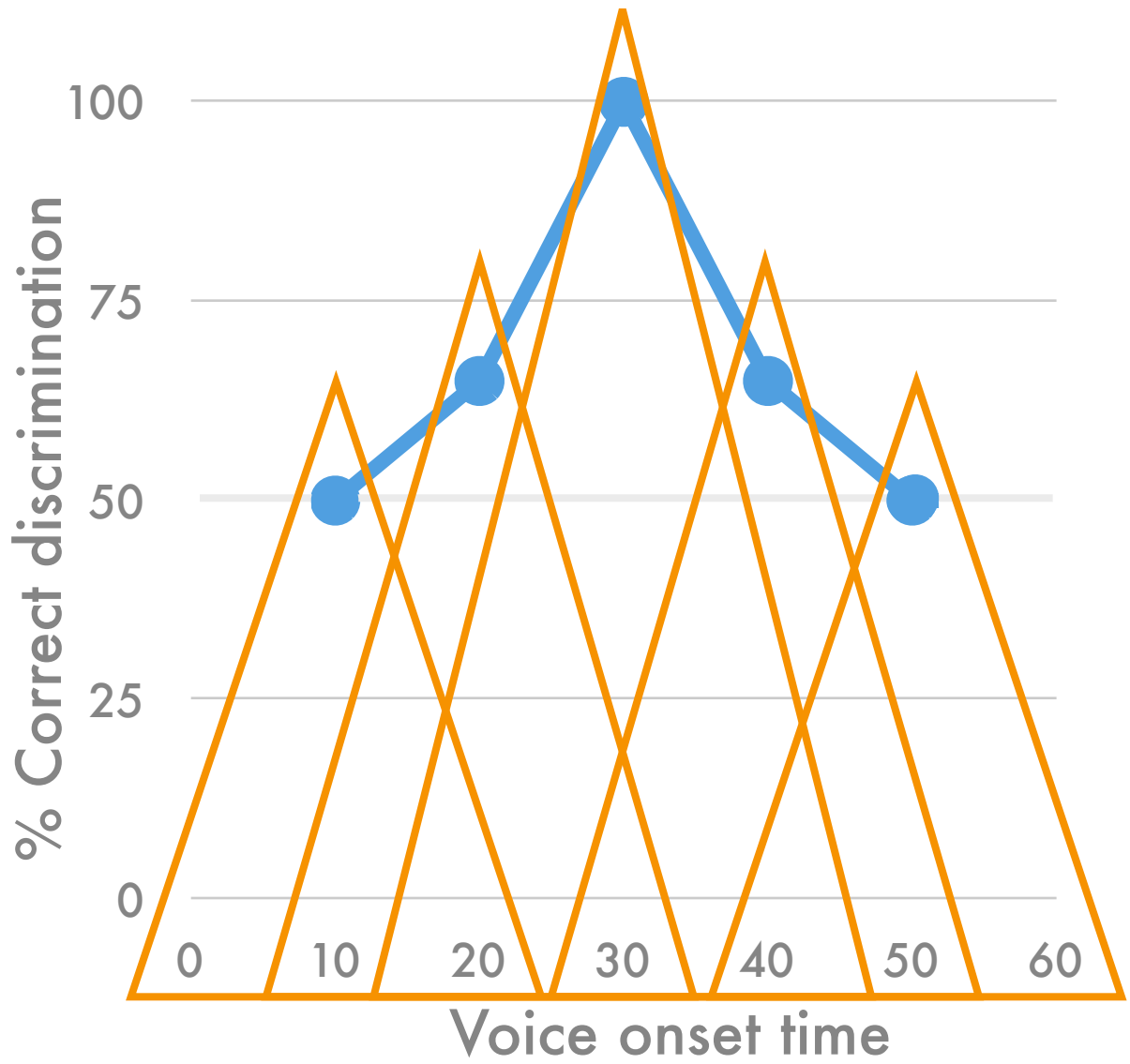
Categorical perception

- Vary VOT
- Ask subjects to identify CV pairs as /da/ or /ta/
- At 30ms VOT, subjects are markedly less accurate



Categorical perception

- Vary VOT between CV pairs
  - Each pair has 20msec VOT difference
- Discrimination task
- Pairs judged same unless 30 msec boundary is crossed



Categorical perception

Contrast	Major Acoustic Cue
Voicing in initial stops [ba-pa]	Voice Onset Time
Voicing in final stops [ab-ap]	Duration of preceding vowel
Place in stops [ba-da-ga]	Start & Duration of F2
Voicing in final fricatives [as-az]	Duration of preceding vowel
Place in fricatives [sa-sha]	Frequency of noise
Liquids [la-ra]	Frequency of F3

Categorical perception of sound

- Categorical perception not limited to language
  - Music - can categorize instruments, tones
- Categorical perception not limited to humans
  - Vervet monkeys
- VOT contrast not limited to humans
  - Chinchillas and Japanese Quail

VOT = 20

Food if /ba/

Speaker

Food if /pa/

VOT = 20

Food if /ba/

Speaker

Food if /pa/

VOT = 40

Food if /ba/

Speaker

Food if /pa/

VOT = 40

Food if /ba/

Speaker

Food if /pa/

- Last note on VOT and categorical perception:
- Bilinguals could have two different category perception systems, but they don't
  - Instead: single system of perceptual category boundaries at a midpoint between the two categories of the monolingual versions of the languages.

Top-down effects

- Phonemic restoration effect (Warren, 1970)
- Replace one phonetic segment with coughing sound

The governor met with the state legi*lature earlier in the year.

  - Subjects report hearing the phonetic segment
  - Subjects are not accurate in locating the cough in sentence

"Shadowing task" also revealed restoration effect (Warren & Warren, 1970)

- It was found that the *eel was on the axle.
- It was found that the *eel was on the shoe.
- It was found that the *eel was on the orange.
- It was found that the *eel was on the table.

- We live in a noisy world!

- When DO we notice errors or irregularities?
- Cole & Jakimik (1980)
  - Subjects instructed to press a button when they noticed an error
    - Most accurate for:
      - beginnings of words
      - place of articulation

Context Effects

- Mann & Repp (1981)

- Fooli[sh] _t_apes
- Christma[s] _c_apes
  - Ambiguous: /k/ or /t/

Context Effects

- Garnes & Bond (1976)
- Here's the fishing gear and the _____.
- Check the time and _____.
- Paint the fence and _____.

/___ate/

- If ambiguous, then bait, date, gate.
- If not ambiguous, then no context effect

- Wrapping up acoustic processing
  - Acoustic cues can be used to identify phonemes
    - Vowels vs. consonants
    - Normalization
    - Coarticulation
  - Context aids identification process
    - Top-down processing