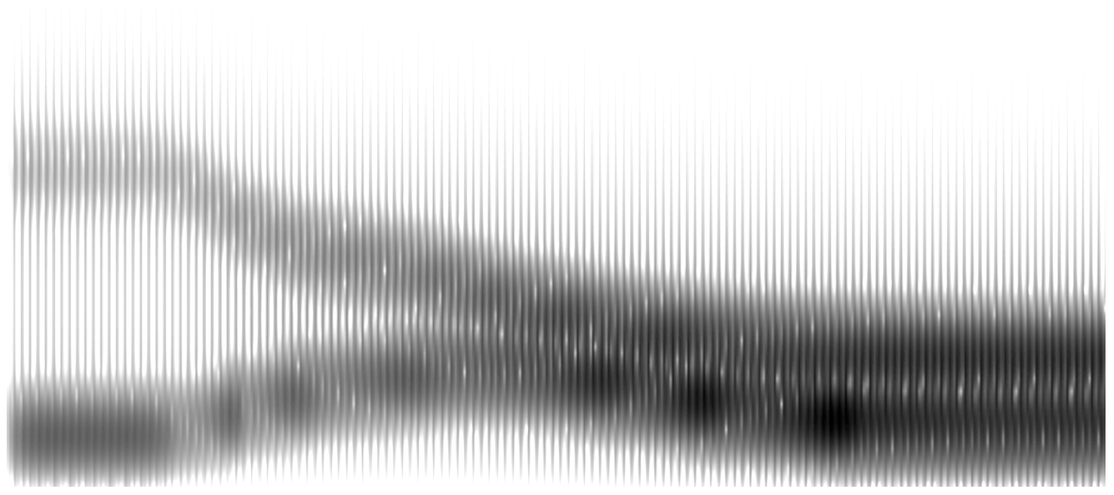
Speech Perception



Approaches to vowel perception

College of Arts and Sciences
Department of Linguistics

2019 LSA Institute (Beddor & McGowan)

What is a vowel?

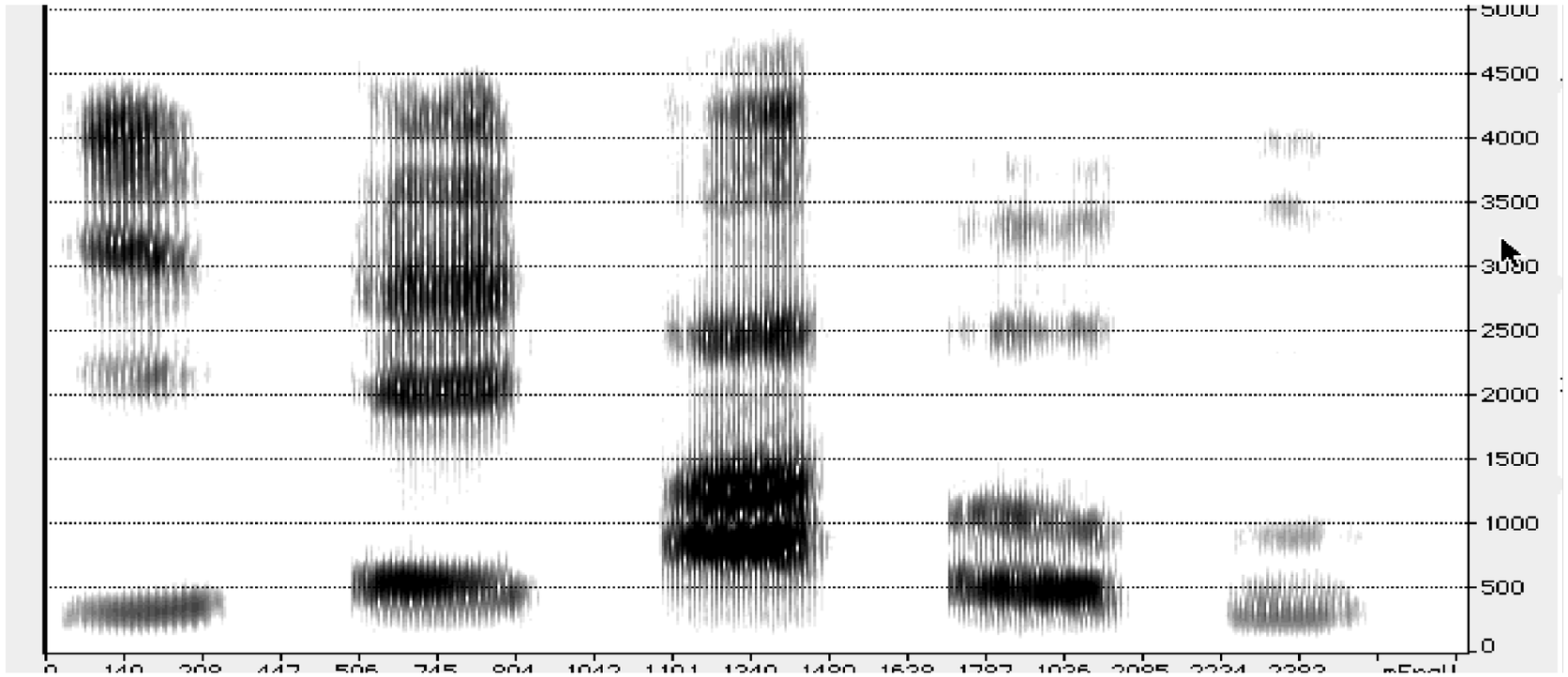
F2

F1

	Front	Central	Back
Close	i • y	ɨ • ʉ	ɯ • u
	ɪ ʏ		ʊ
Close-mid	e • ø	ɘ • ɵ	ɤ • o
		ə	
Open-mid	ɛ • œ	ɜ • ɞ	ʌ • ɔ
	æ	ɐ	
Open	a • ɶ		ɑ • ɒ

Where symbols appear in pairs, the one to the right represents a rounded vowel.

Spectrographic vowel formants



Target Theory

The approach to vowel perception that assumes a fairly straightforward mapping from articulatory target to acoustic target to percept is called **target theory**.

Target frequencies: center formant frequencies characteristic of a given steady-state vowel.

Target Theory: vowels can be represented as static points located by their target frequencies in an F_1 x F_2 (xF_3) vowel space.

Problems with Target Theory:

- Across speakers, target frequencies for different vowels **overlap**.

- Across contexts, in continuous speech vowels may not reach their targets due to coarticulation (aka **vowel undershoot**).

Target theory problems: overlap

- **Overlapping formant frequencies** across speakers were found in a classic study of vowel productions of 76 American English-speaking adults and children (Peterson & Barney, 1952, JASA 24).

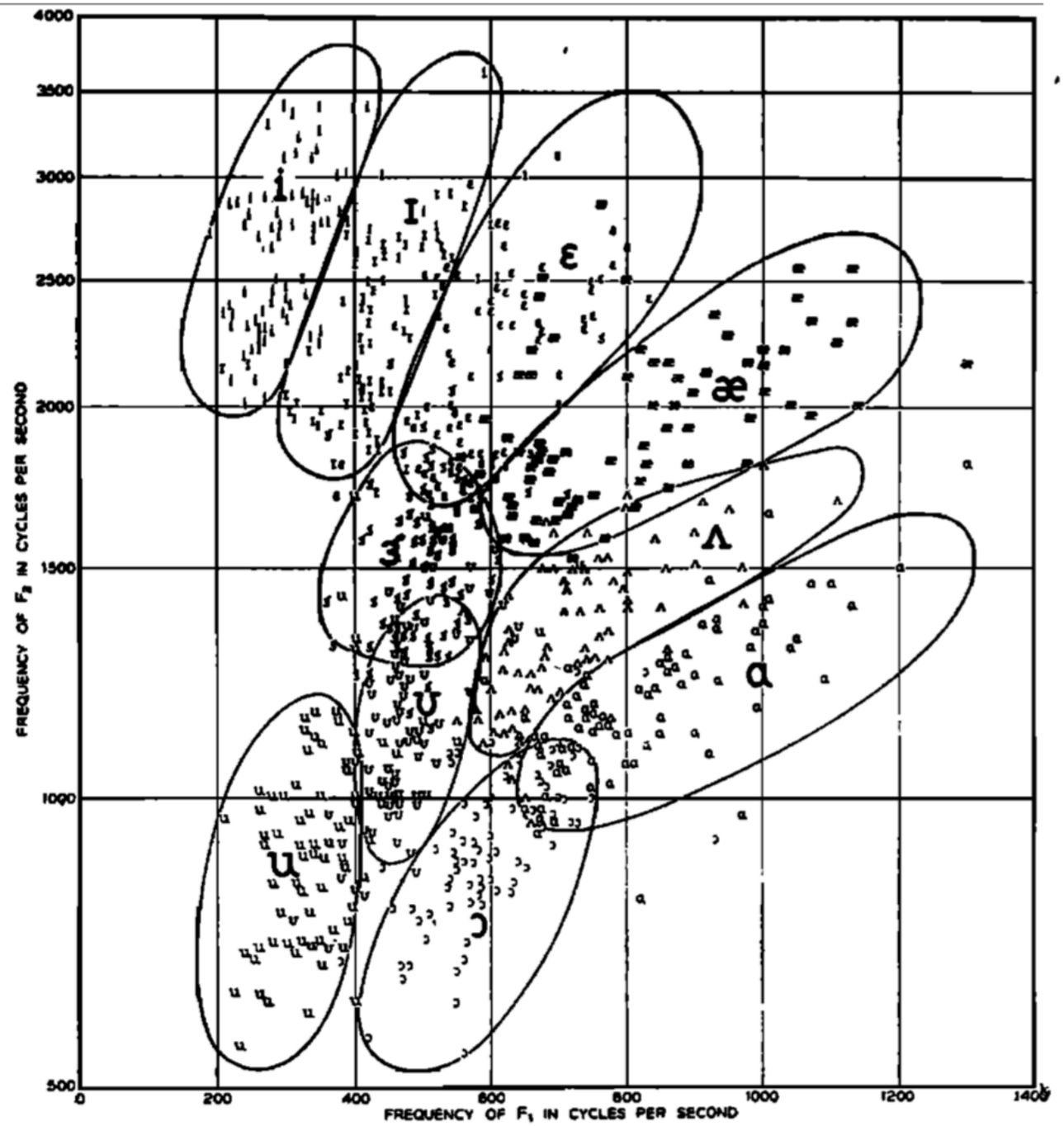
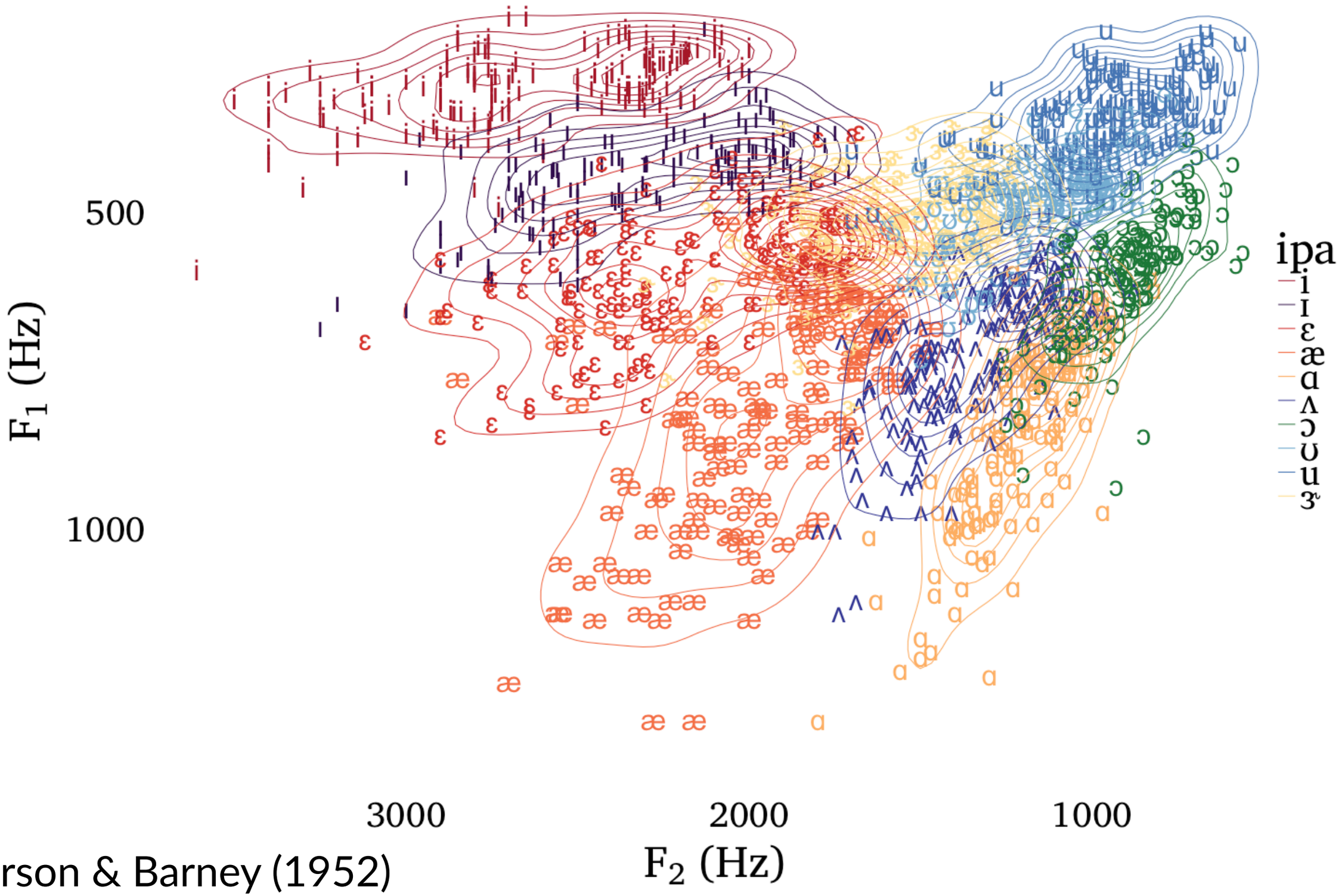
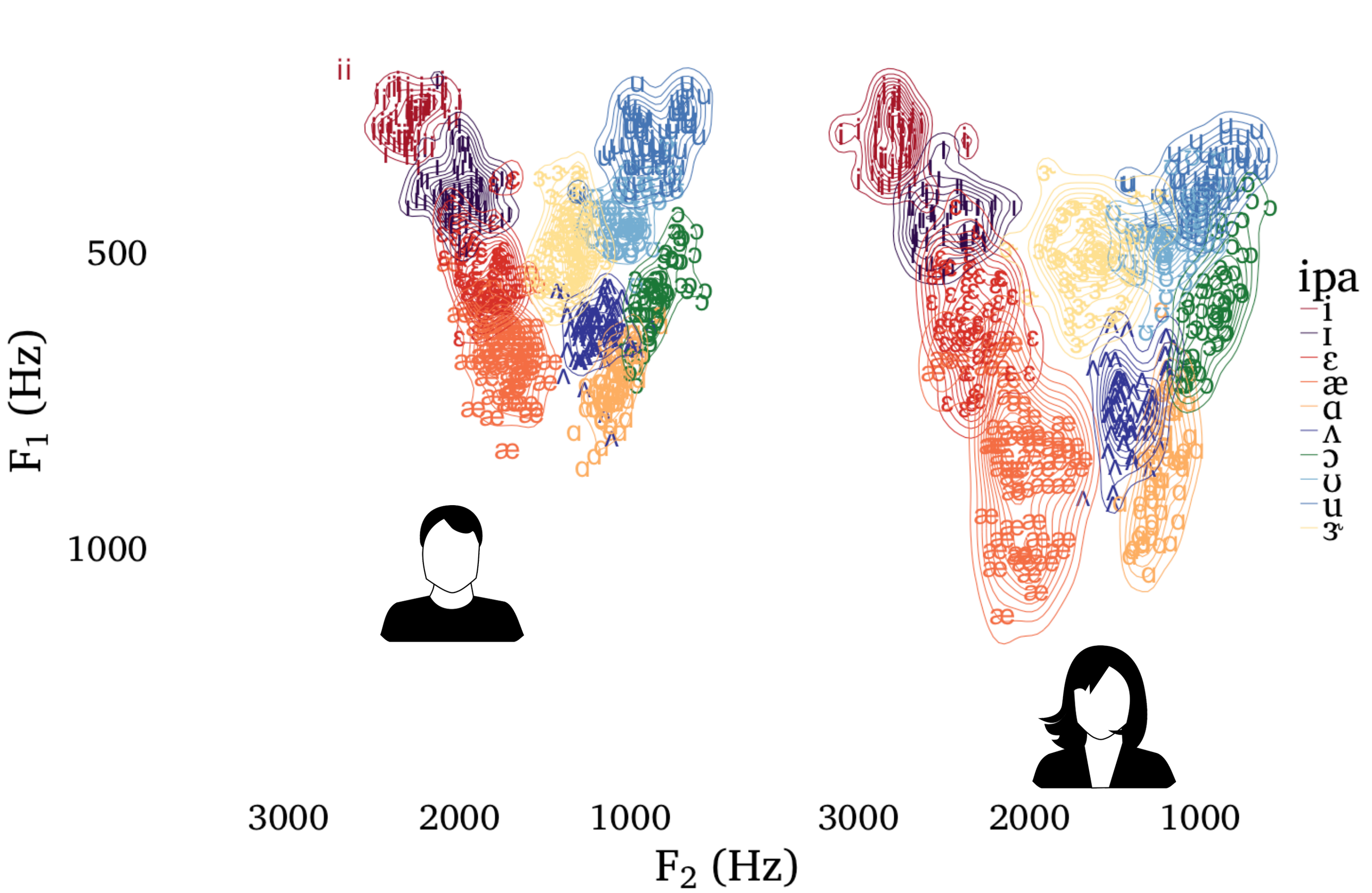


FIG. 8. Frequency of second formant *versus* frequency of first formant for ten vowels by 76 speakers.

Target theory problems: overlap



Peterson & Barney (1952)

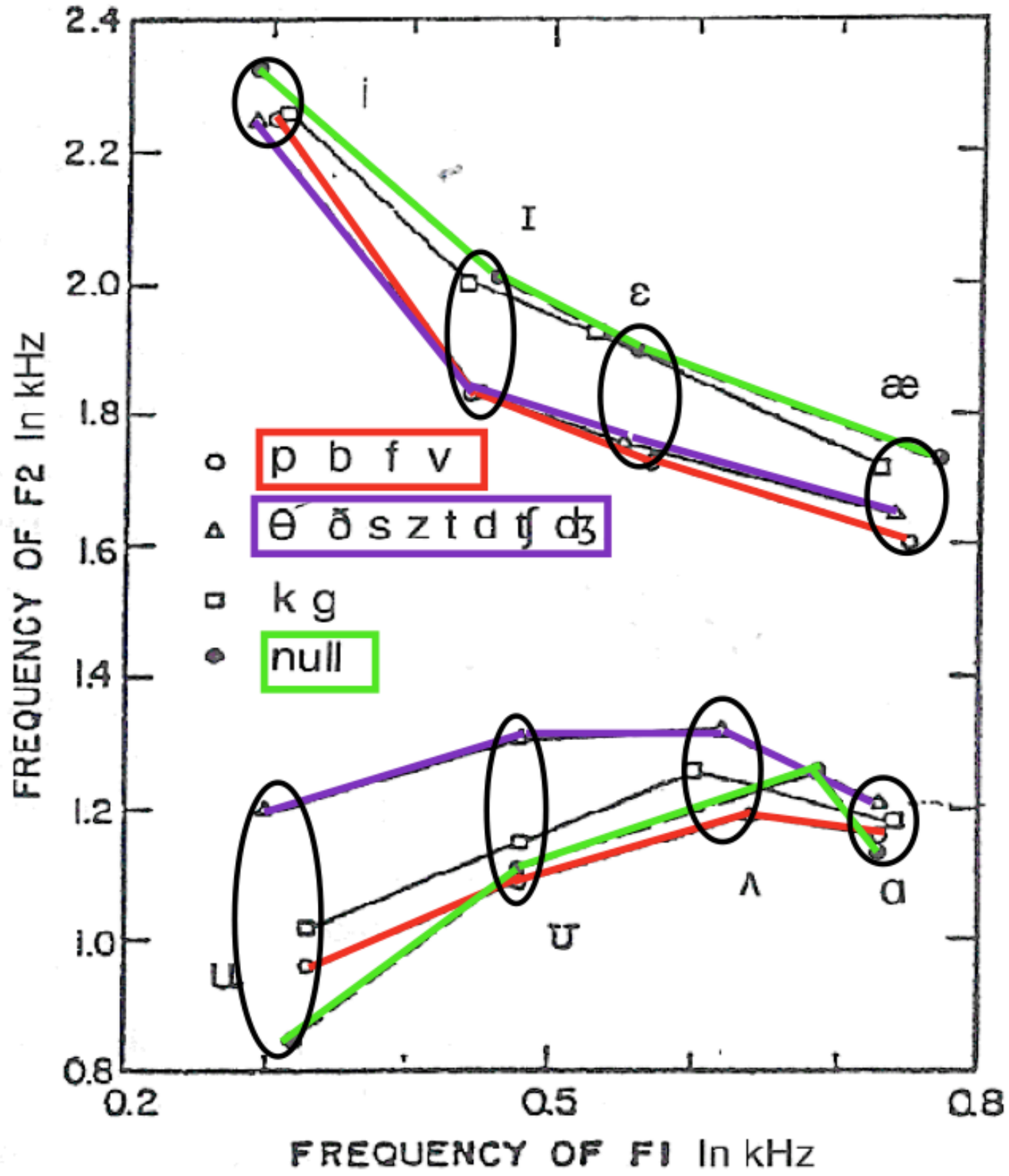


Peterson & Barney (1952)

Target Theory Problems: Undershoot

- **Target undershoot** due to coarticulation was shown in other classic work by Stevens & House (1963, JSHR 6) and Lindblom (1963, JASA 35).

- The figure (from Stevens & House, 1963) shows undershoot due to C context.



Addressing Target Theory Problems

- Let’s focus first on these across-speaker differences:

- Given acoustic overlap in formant frequencies for different vowels across different speakers, how do listeners achieve **perceptual constancy**?

- Some researchers have argued that perceptual constancy does not rest solely in the acoustic characteristics of the particular vowel currently being perceived. . .

Extrinsic Normalization Theory

Hypothesis: Vowels are perceived relationally. Listeners identify vowels by calibrating their perceptual mechanisms for each speaker's utterances on the basis of a sample of speech from that speaker.

Early supporting evidence: Ladefoged & Broadbent (1957, JASA 29)

Extrinsic Normalization Theory

- Ladefoged and Broadbent (1967) tested extrinsic normalization.

  Stimuli: 4 synthesized test words (A-D) preceded by a version of the precursor sentence: “Please say what this word is...”

  Experimental groups: Test words introduced with a synthetic precursor sentence. The 6 versions of this sentence differed in the vowels' F1 and F2 frequencies.

  Control group: no precursor.

  Without precursor, % test word ID was:
  A = bit: 87% B = bet: 77% C = bat 55% (bet 45%) D = but: 75%

Ladefoged & Broadbent (1957)

Please say what this word is:
bit bet bat but

F1 of CARRIER

1. 380-660 Hz [b?t]
2. 200-380 Hz [b?t]
3. (no carrier) [b?t]

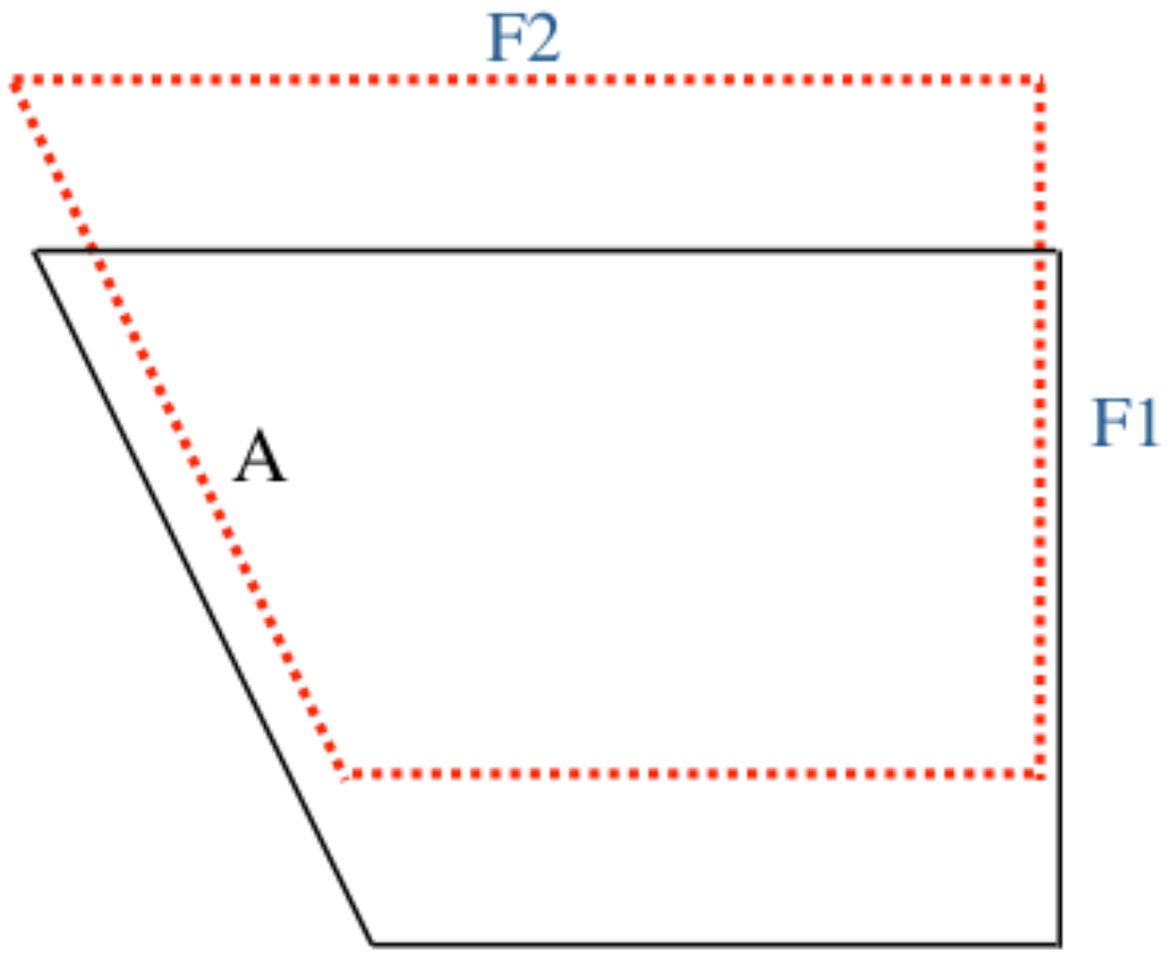
Ladefoged and Broadbent (1957) *JASA* 29, 98-104

Shifted F1 space

- The difference between the two precursor sentences was the vowel's F1 frequency space.

- Test word A was heard as:

	bit	**bet**
F1 up	88%	12%
F1 down	7%	90%



That is, lowering F1 of the vowels of the precursor sentence caused the vowel of Test Word A to be perceived as lower, because it occupied a relatively low position in the new vowel space (dashed read line).

Extrinsic Normalization Theory

- Ladefoged & Broadbent (1957)'s results tell us that identification of a particular vowel is at least somewhat influenced by knowledge of the rest of that speaker's vowel space.

- But is such prior knowledge *necessary* for accurate vowel identification?

Strong normalization theory predicts:

- More errors for listening tasks with multiple speakers

- More errors without prior V space information (i.e., without precursors)

Are these predictions upheld experimentally?

Extrinsic Normalization Theory

- For the most part these predictions are NOT upheld.

- The general picture emerging from research (since the 1970s and 80s) is that even in experimental conditions with multiple speakers and no prior knowledge of each speaker's vowel space (e.g., no precursors), identification of most vowels is fairly accurate.

- Thus, while speaker-specific information extrinsic to the particular vowel *can* bias listeners' vowel judgments, **extrinsic normalization is not necessary** for accurate vowel identification.

PVP id errors with and without precursors

Intended vowel	Condition: Mixed talker	Segregated talker	Point-vowel precursor	Central-vowel precursor
i	1.1	0.3	3.3	3.3
ɪ	1.6	3.6	2.7	1.7
ɛ	26.8	12.1	4.7	10.8
æ	18.9	1.8	20.7	18.3
ɑ	20.0 (10.0)	22.7 (3.9)	43.3 (26.7)	29.2 (12.5)
ɔ	27.4 (3.2)	18.5 (1.8)	18.7 (12.7)	13.3 (2.5)
ʌ	15.3	7.6	9.3	22.5
ʊ	38.9	17.6	26.7	29.2
u	2.6	0.9	7.3	5.8
Overall	**17.0 (13.2)**	**9.5 (5.5)**	**15.2 (12.7)**	**14.9 (11.9)**

Percent vowel identification errors (/p_p/ syllables).
Verbrugge et al. 1976, *JASA 60(1)*

[i ɑ u]

[ɛ ɑ ʌ]

Theoretical approaches to undershoot

- **Extrinsic normalization** focuses on relational information *between* the vowels in a word, phrase, sentence, discourse, community of practice, language, etc.

- **Intrinsic normalization** focuses on relational information *within* the vowel nucleus.

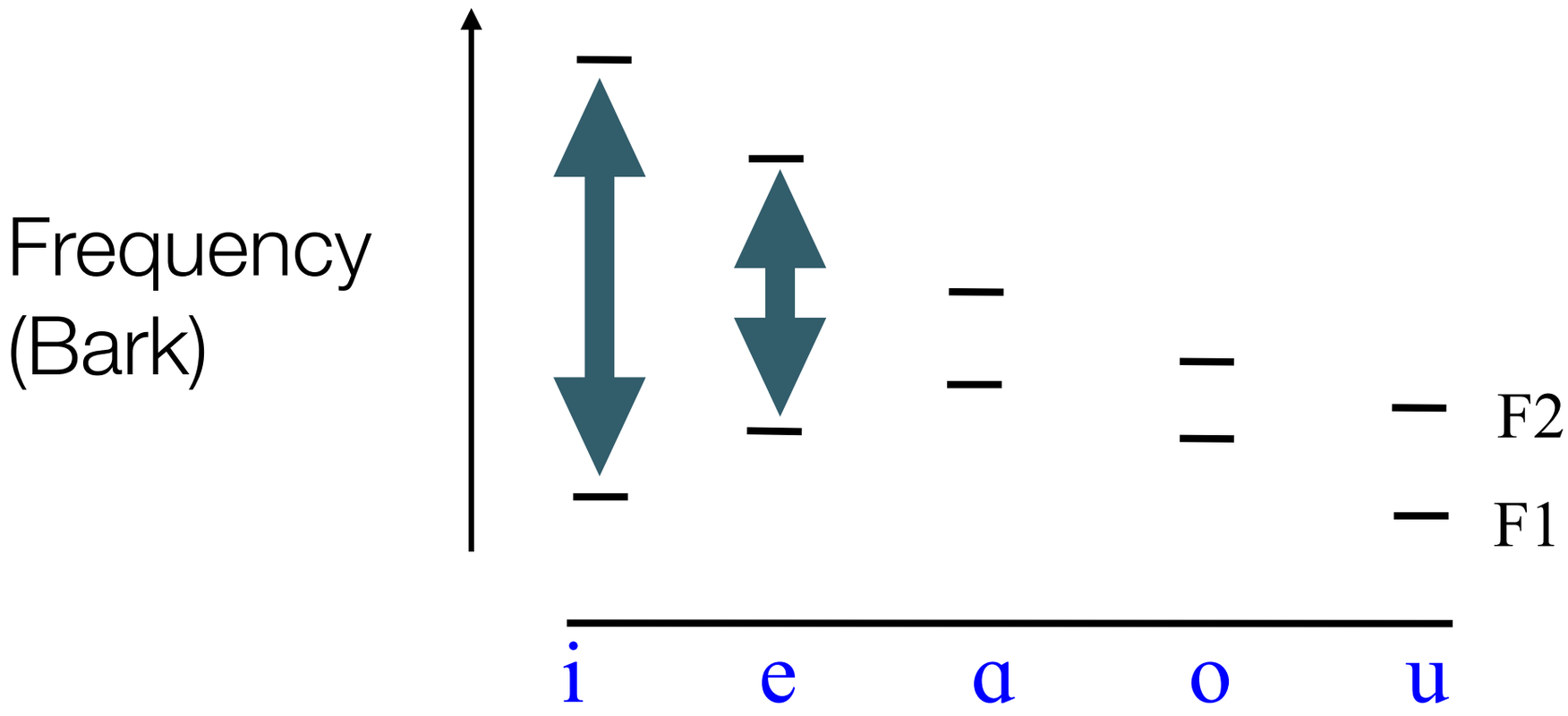
Intrinsic Vowel Normalization

Relational Information in the Vowel Nucleus:

Hypothesis: Variation across speakers can be minimized, and perceptual distinctions maximized, if the auditory relations between spectral prominences are taken into account.

Evidence: Delattre et al. 1952, Chistovich et al. 1979, Syrdal & Gopal 1986 , Miller 1989, Beddor & Hawkins 1990

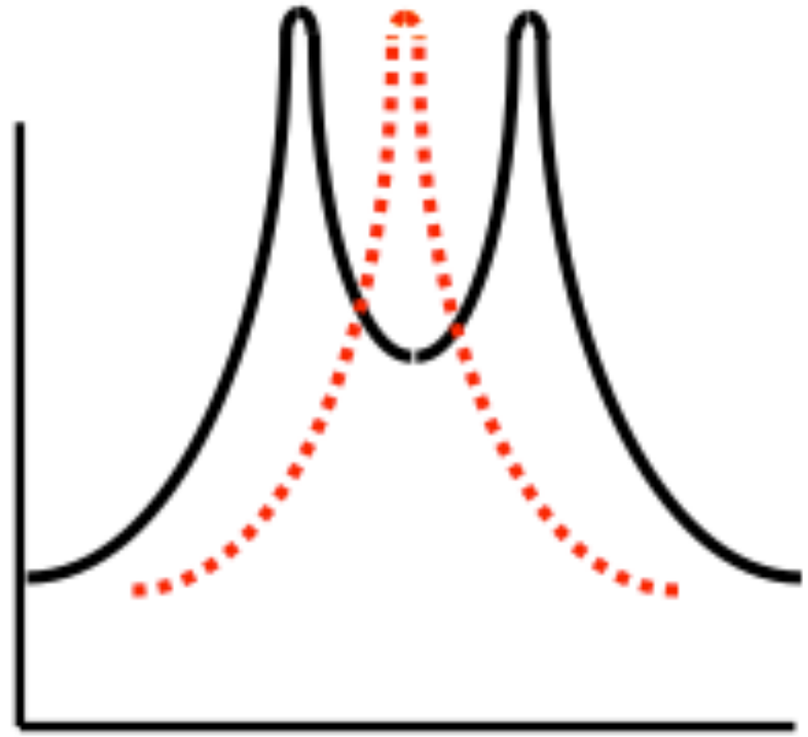
Intrinsic Vowel Normalization



- “Chiba and Kajiyama (1941) … hypothesized vowels are recognized on the basis of spatial patterns of excitation in the peripheral auditory system regardless of their specific location along the spatially coded frequency dimensions.” (Syrdal and Gopal, 1986)

Intrinsic Vowel Normalization: Example

- Spectral “center of gravity”: When two spectral prominences fall within a certain critical distance (3.0 to 3.5 Bark), listeners’ auditory systems effectively average the two prominences F1 and F2, resulting in a **percept** (F’) that is intermediate in frequency. (Chistovich et al., 1979)



Intrinsic Vowel Normalization

- Syrdal and Gopal (1986, JASA 79): proposed normalizing vowels by converting F0, F1, F2, and F3 from Hz to Bark and then using the *differences* between these measurements as the dimensions for vowel

Hertz

TABLE V. Classification matrix for ten vowels in hertz.

Group	Percent correct	Number of tokens classified into group /i/	/ɪ/	/ɛ/	/æ/	/ɜ/	/ʌ/	/a/	/ɔ/	/U/	/u/
/i/	89.3	134	16	0	0	0	0	0	0	0	0
/ɪ/	80.7	13	12	6	0	0	0	0	0	0	0
/ɛ/	81.3	0	27	122	1	0	0	0	0	0	0
/æ/	85.3	0	1	19	128	0	1	1	0	0	0
/ɜ/	96.7	0	1	3	1	145	0	0	0	0	0
/ʌ/	82.7	0	0	0	0	0	124	19	7	0	0
/a/	77.3	0	0	0	0	0	26	116	7	1	0
/ɔ/	75.0	0	0	0	0	0	10	8	108	4	14
/U/	75.3	0	0	0	0	0	3	0	5	113	29
/u/	78.7	0	0	0	0	1	0	0	2	29	118
Total	82.3	147	166	160	130	146	164	144	129	147	161
Jackknifed total	81.8										

Intrinsic Vowel Normalization

- Syrdal & Gopal (1986, JASA 79): proposed normalizing vowels by converting F0, F1, F2, and F3 from Hz to Bark and then using the *differences* between these measurements as the dimensions for vowel

Bark

TABLE VI. Classification matrix for ten vowels in bark differences.

Group	Percent correct	Number of tokens classified into group /i/	/ɪ/	/ɛ/	/æ/	/ɝ/	/ʌ/	/a/	/ɔ/	/ʊ/	/u/
/i/	95.3	143	7	0	0	0	0	0	0	0	0
/ɪ/	84.0	10	126	14	0	0	0	0	0	0	0
/ɛ/	86.7	0	20	130	0	0	0	0	0	0	0
/æ/	86.7	0	0	20	130	0	0	0	0	0	0
/ɝ/	94.0	0	2	6	1	141	0	0	0	0	0
/ʌ/	88.7	0	0	0	0	0	133	14	3	0	0
/a/	88.7	0	0	0	0	0	10	133	6	1	0
/ɔ/	79.9	0	0	0	0	0	3	11	115	4	11
/u/	77.3	0	0	0	0	0	4	0	3	116	27
/u/	77.3	0	0	0	0	0	0	0	3	31	116
Total	85.9	153	155	170	131	141	150	158	130	152	154
Jackknifed total	85.7										

Classification accuracy: Syrdal & Gopal

Hertz

Group	Percent correct
/i/	89.3
/ɪ/	80.7
/ɛ/	81.3
/æ/	85.3
/ɝ/	96.7
/ʌ/	82.7
/a/	77.3
/ɔ/	75.0
/U/	75.3
/u/	78.7
Total	82.3
Jackknifed total	81.8

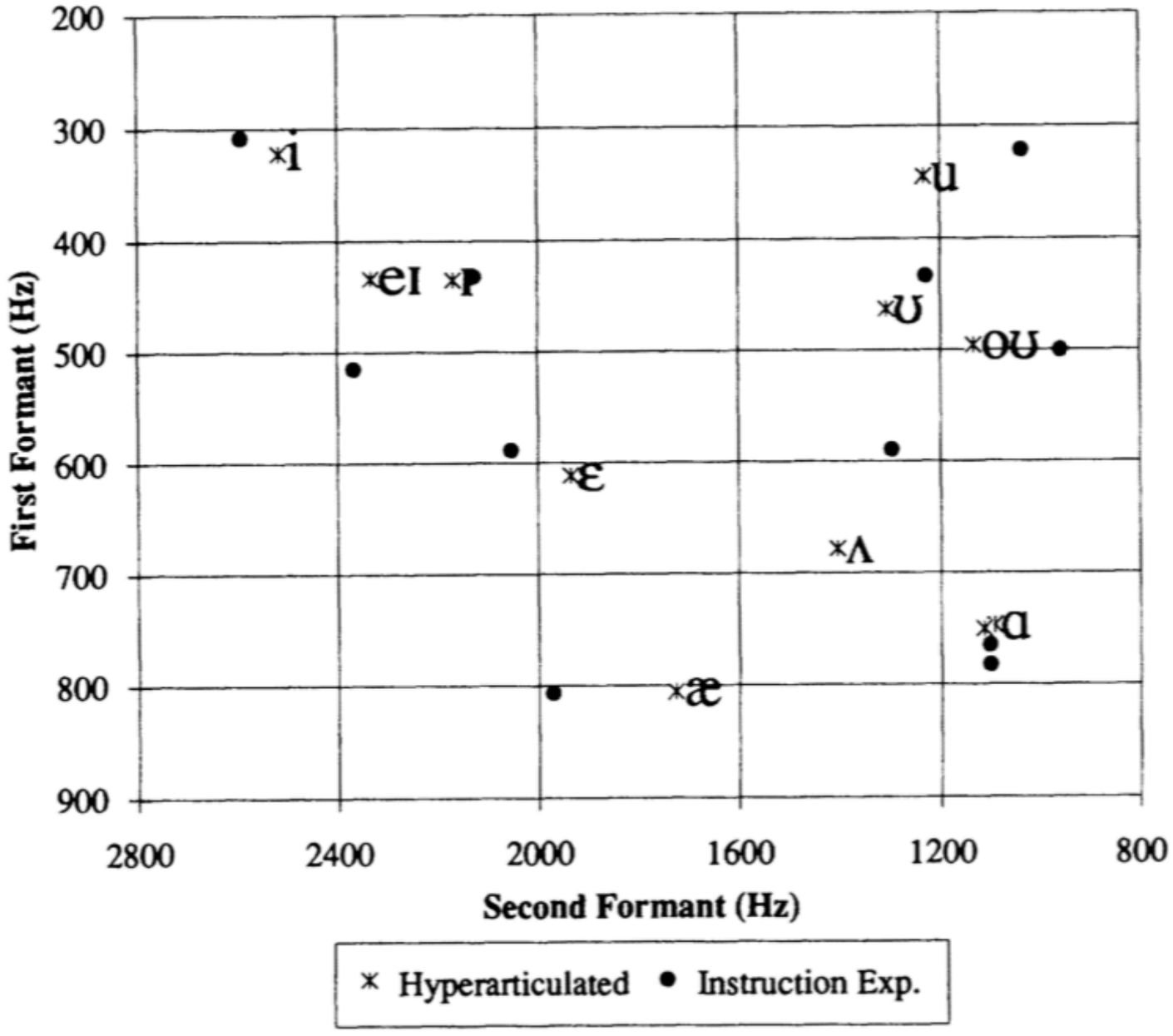
Bark

Group	Percent correct
/i/	95.3
/ɪ/	84.0
/ɛ/	86.7
/æ/	86.7
/ɝ/	94.0
/ʌ/	88.7
/a/	88.7
/ɔ/	79.9
/u/	77.3
/u/	77.3
Total	85.9
Jackknifed total	85.7

Hyperspace effect

- Johnson, Fleming, & Wright (1993) (auditory theorists) found evidence that listeners expect F1/F2 targets that are more extreme than those found in (even their own) casual speech.

- Whalen, Magen, Pouplier, Min Kang, and Iskarous (2004) claim that the observed effect is due to Californian participants fronting back vowels! And argue that there is no need to posit a hyperspace effect.



Target Theory

The approach to vowel perception that assumes a fairly straightforward mapping from articulatory target to acoustic target to percept is called **target theory**.

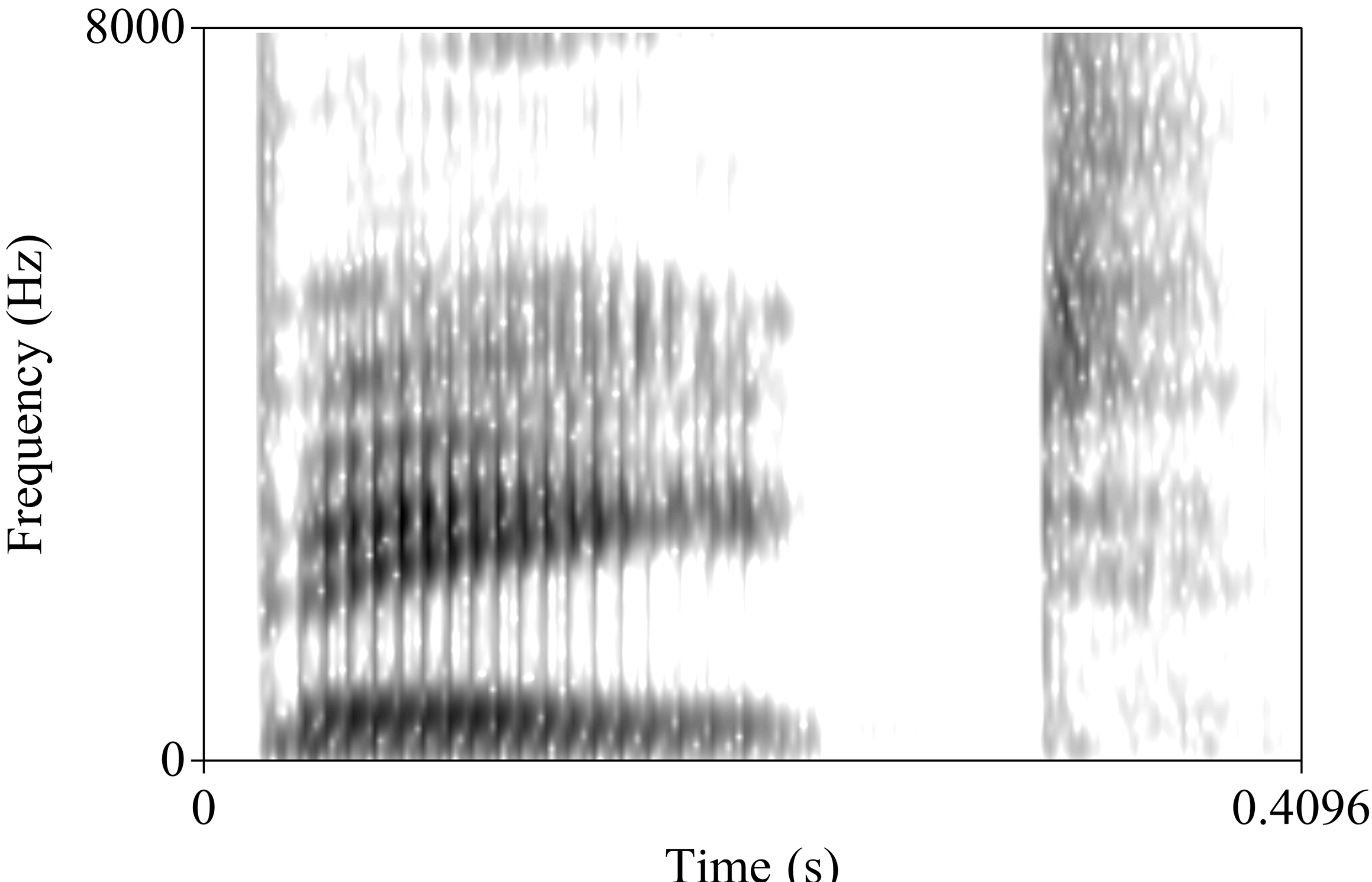
Target frequencies: center formant frequencies characteristic of a given steady-state vowel.

Target Theory: vowels can be represented as static points located by their target frequencies in an F_1 x F_2 (xF_3) vowel space.

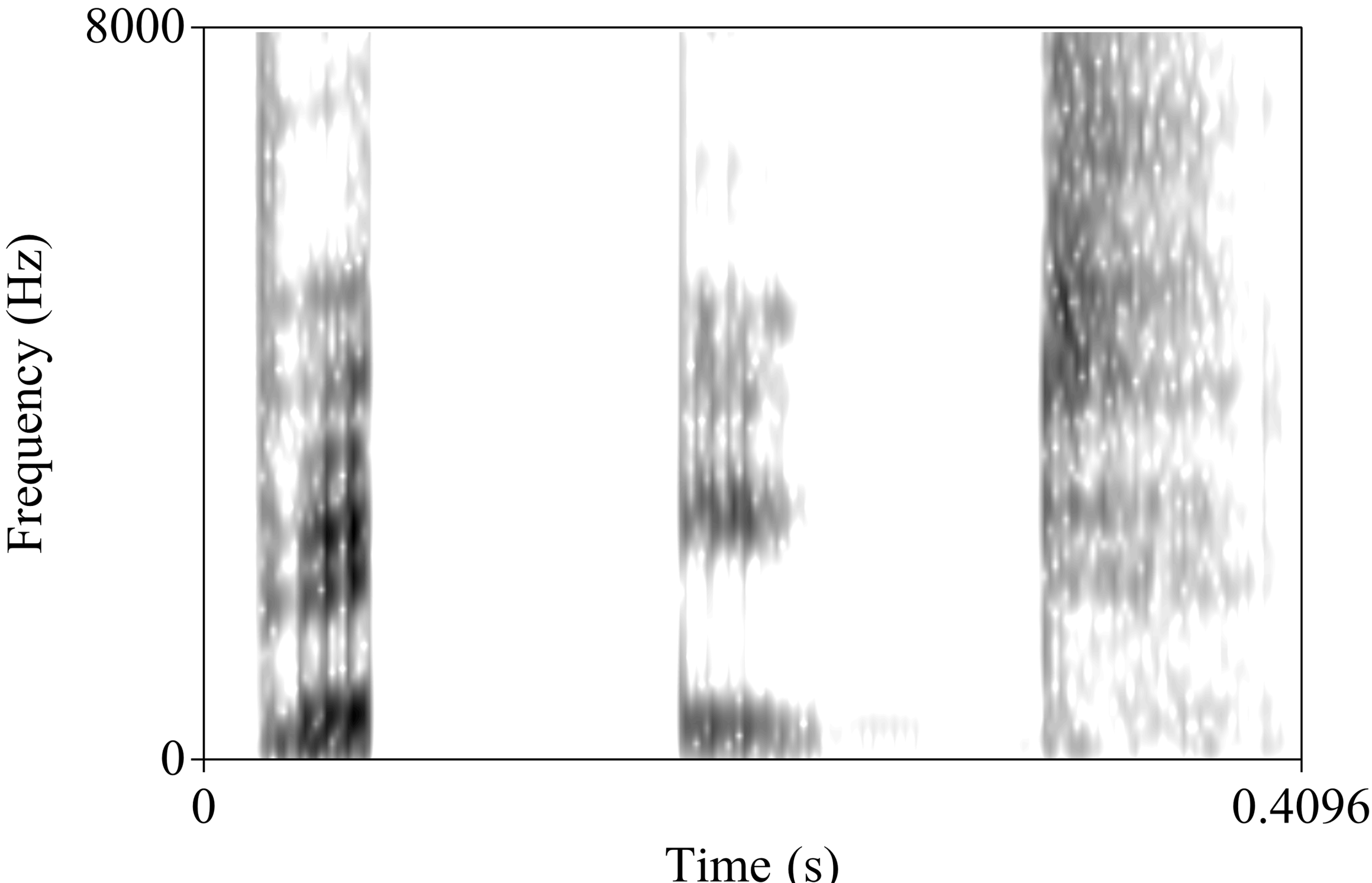
Problems with Target Theory:

- Overlap and undershoot (aka the lack of invariance problem!) are problems, sure, but the real problem with Target Theory is *much* bigger…

Where is a vowel most like itself?



Silent Centers



Silent Centers

- Strange, Jenkins, & Johnson (1983), among other manipulations, excised 65% of the center of the vowel without a significant reduction in listeners' ability to perceive the vowel quality.

- How do listeners interpret the acoustic –or, rather, auditory– signal as vowels so reliably?

- Could the object of perception still be a target (either a pair of frequencies or a gestural target)?

So… what's a vowel?

- Maybe a "vowel" is the whole thing.

- Remember that the human ear is good at slow changes at these frequencies. Vowels are slowly changing patterns of resonance. Think about how much more informative (and robust to communication errors) that is than 2 points!

- Maybe we've construed the goal wrong. Maybe we shouldn't be trying to throw away variation to extract phonemes.

It matters how we frame the question

- What do we mean by “linguistic message?” What is the domain or level of the message?

Feature?

Gesture?

Syllable?

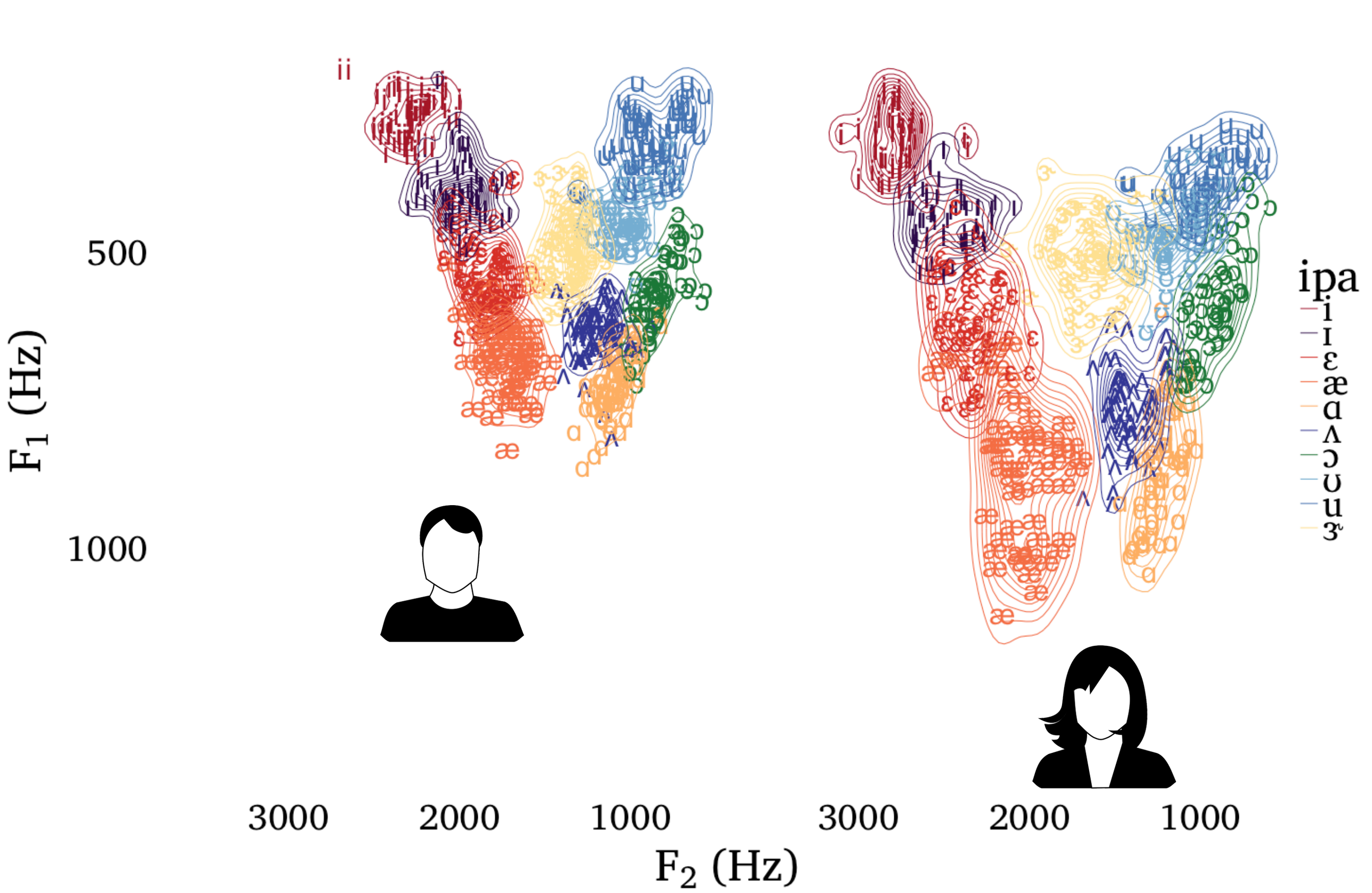
Phoneme?

Word?

Meaning?

Subcategorical variation isn't always noise!

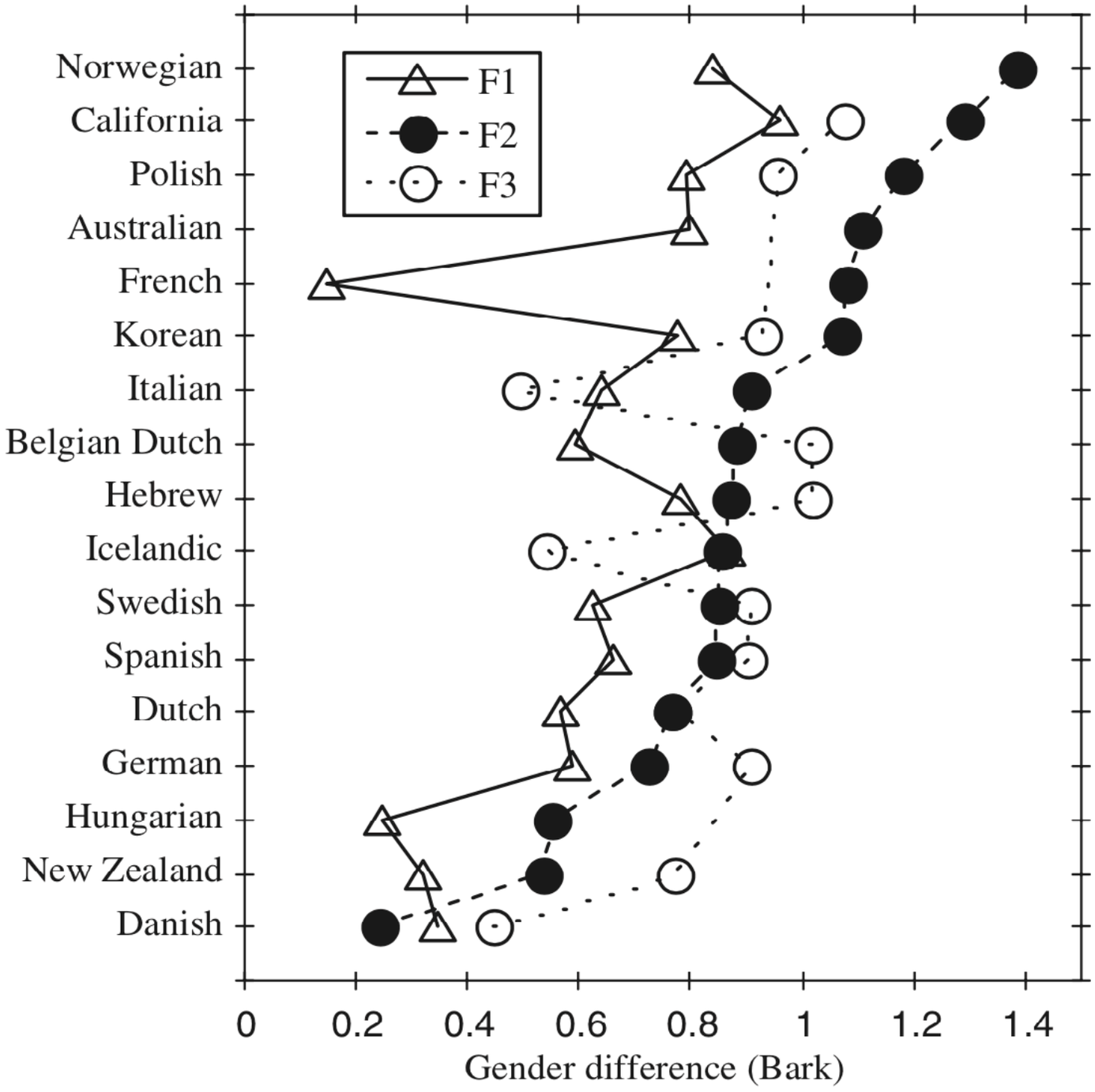
- Listeners are highly sensitive to subcategorical phonetic detail
  - Whalen, 1984; McMurray et al., 2002; Hawkins, 2003; Clayards et al., 2008; Sumner, 2011
- We do not filter out phonetic variation, we store it
  - Church & Schacter, 1994; Goldinger, 1996; Johnson, 1997; Nygaard & Pisoni, 1998; Nosofsky, 1986; etc.
- Some within-category variation, e.g., coarticulation, is used as soon as it becomes available
  - Beddor, McGowan, Boland, Coetzee, and Brasher, 2012; Beddor, Styler, Coetzee, Boland, and McGowan, 2019



Peterson & Barney (1952)

Gender works differently in different languages

- Johnson (2006) collects evidence from a number of studies to reveal systematic differences in the way gender is communicated through vowel formants cross-linguistically



Exemplar Theory

- Exemplar theories (e.g. Goldinger 1998; Johnson 1997; Pierrehumbert 2001, etc.) avoid many of these problems by offering “Speech Perception without Speaker Normalization”

    - Rather than processing the speech signal to identify the cues that link particular formant frequencies to particular vowels,

    - Listeners compare what they are hearing to, in some models, everything they have heard before and

    - Classify this new word according to the **exemplar cloud** it best matches.

Declarative Memory

- There are two types of declarative memory:
  - Semantic memory
    - General facts like “The first moon landing was 50 years ago this month”
  - Episodic memory
    - Personal facts like the experience of watching or listening to the first moon landing 50 years ago

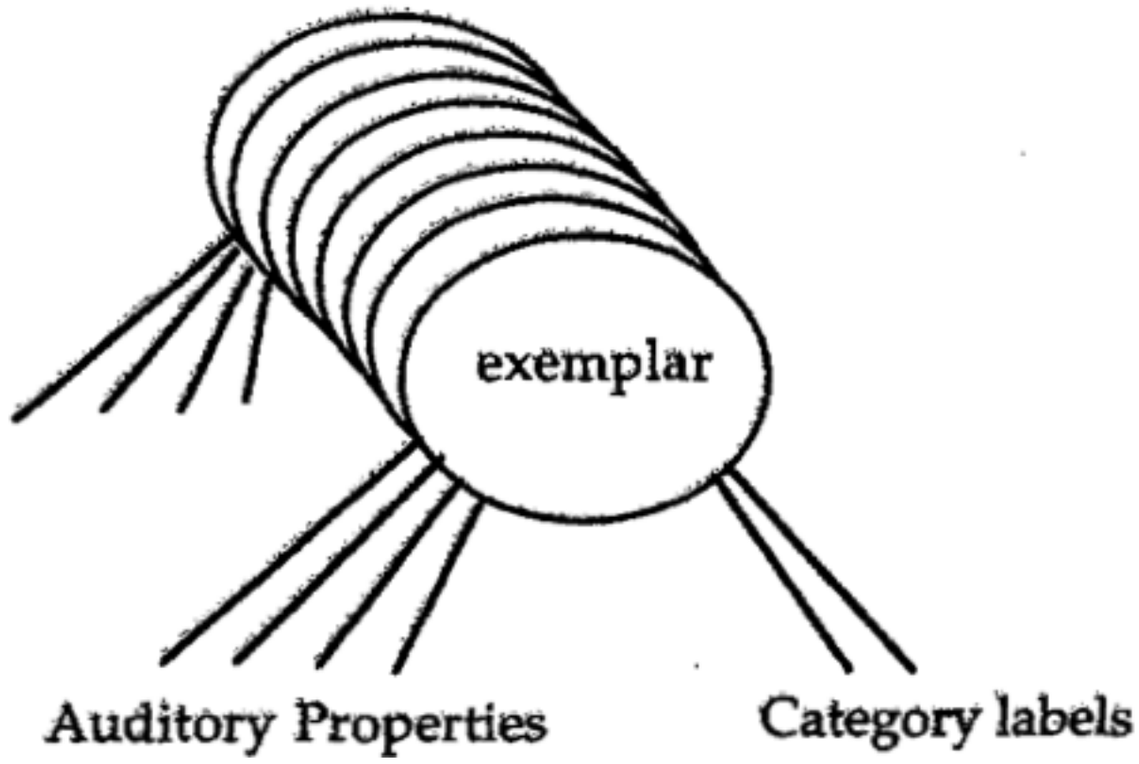
Episodic traces

- Exemplar models propose that we store our personal memories of experiences with speech.

- These memories are stored along with/labeled with the linguistic message we have correctly perceived

- For example: [‘kæt]

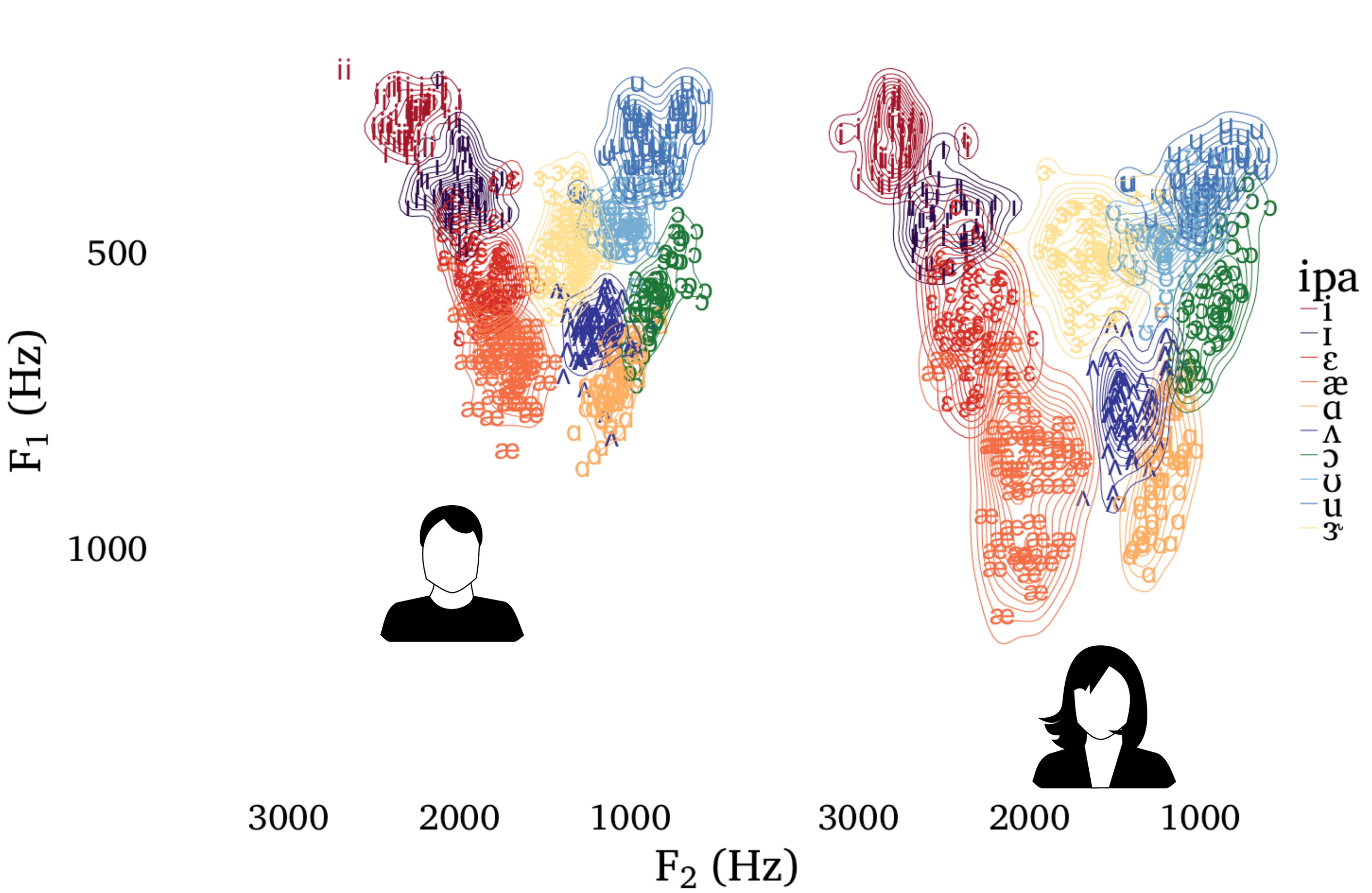
Exemplar Theory



Exemplar Theory: Strengths

- Exemplar models give us our first formal way to link speech perception theories to sociolinguistic theory and linguistic anthropology

- Experience with general *types* of people can be leveraged to help us understand individuals



Peterson & Barney (1952)

Exemplar models: Spoken word perception

- Does this solve the Segmentation problem?
- Does this solve the Lack of Invariance problem?
- Does it settle the auditory/gestural debate?

Exemplar Theory: Final observation

- The kind of speech we hear least often should be the most problematic, but:

- Casual speech is far more common than careful speech and yet careful speech is what we do when we want to be clearer! (see: Sumner, Kim, King, McGowan, 2014)