Invariance and Vowels

- Vowels are “cued” by the frequencies of the first three formants.
- Duration, fundamental frequency and the shape of the short-term spectrum are acoustic correlates of vowel identity.
- Vowels in fluent speech are rarely steady-state.

F1, F2, and F3

Using synthetic speech, we can generate 2-formant (F1 and F2) or 3-formant (F1, F2, & F3) vowels. Except for /e/ (as in heard), all of the vowels in American English can be identified based on 2-formant, synthetic tokens.

Steady-state (formants do not change over time) vowels are not as intelligible as natural vowels.

For example, in a corpus of 1520 vowel tokens, overall intelligibility was 94.4 percent correct (Peterson & Barney). Steady-state versions of these tokens were identified correctly less than 75 percent of the time (Hillenbrand & Gayvert).

Vowels of American English

Formant measurements were made for vowels in hVd context by Peterson and Barney in the 1950's/. A more recent set of measurements can be found in:

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. Journal of the Acoustical Society of America, 97, 3099-3111.

The following figures and tables come from Hillenbrand et al.

Vowels of American English

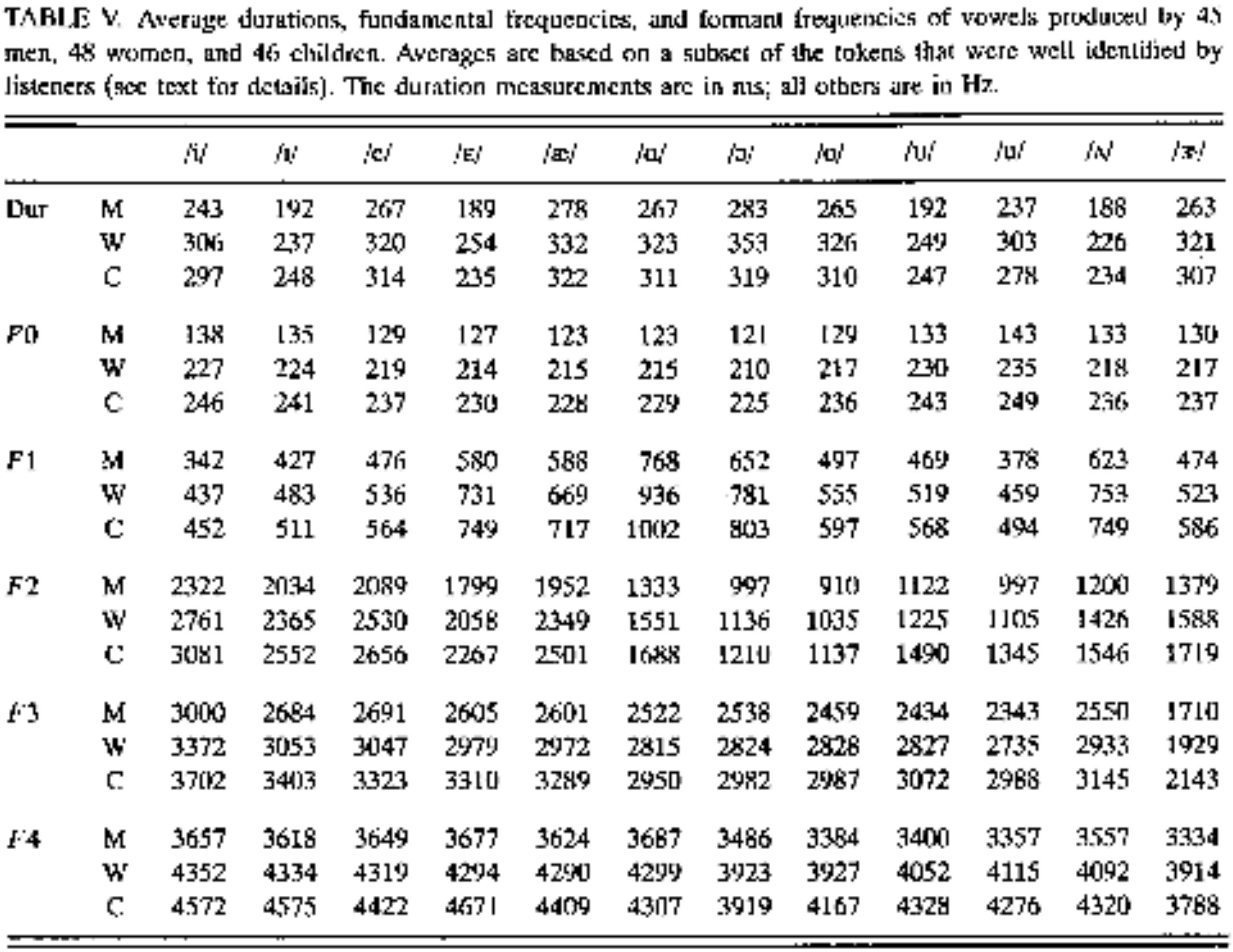


TABLE V. Average durations, fundamental frequencies, and formant frequencies of vowels produced by 45 men, 48 women, and 46 children. Averages are based on a subset of the tokens that were well identified by listeners (see text for details). The duration measurements are in ms; all others are in Hz.

		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɑ/	/ɔ/	/o/	/ʊ/	/u/	/ʌ/	/ɝ/
Dur	M	243	192	267	189	278	267	283	265	192	237	188	263
	W	306	237	320	254	332	323	353	326	249	303	226	321
	C	297	248	314	235	322	311	319	310	247	278	234	307
$F0$	M	138	135	129	127	123	123	121	129	133	143	133	130
	W	227	224	219	214	215	215	210	217	230	235	218	217
	C	246	241	237	230	228	229	225	236	243	249	236	237
$F1$	M	342	427	476	580	588	768	652	497	469	378	623	474
	W	437	483	536	731	669	936	781	555	519	459	753	523
	C	452	511	564	749	717	1002	803	597	568	494	749	586
$F2$	M	2322	2034	2089	1799	1952	1333	997	910	1122	997	1200	1379
	W	2761	2365	2530	2058	2349	1551	1136	1035	1225	1105	1426	1588
	C	3081	2552	2656	2267	2501	1688	1210	1137	1490	1345	1546	1719
$F3$	M	3000	2684	2691	2605	2601	2522	2538	2459	2434	2343	2550	1710
	W	3372	3053	3047	2979	2972	2815	2824	2828	2827	2735	2933	1929
	C	3702	3403	3323	3310	3289	2950	2982	2987	3072	2988	3145	2143
$F4$	M	3657	3618	3649	3677	3624	3687	3486	3384	3400	3357	3557	3334
	W	4352	4334	4319	4294	4290	4299	3923	3927	4052	4115	4092	3914
	C	4572	4575	4422	4671	4409	4307	3919	4167	4328	4276	4320	3788

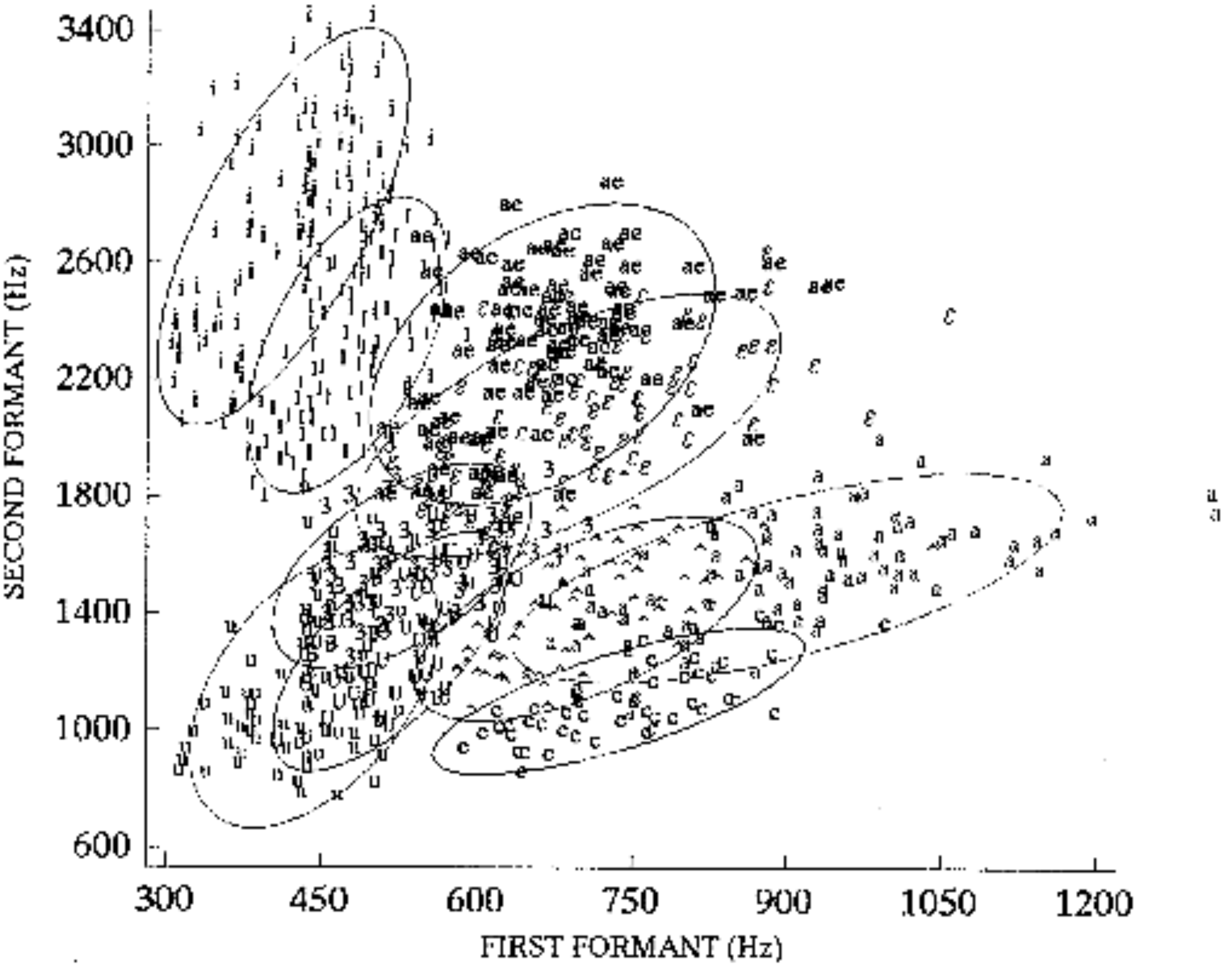


FIG. 4. Values of $F1$ and $F2$ for 46 men, 48 women, and 46 children for 10 vowels with ellipses fit to the data ("ae"=/æ/, "a"=/ɑ/, "c"=/ɔ/, "A"=/ʌ/, "u"=/ʊ/). Measurements for /e/ and /o/ have been omitted, and the data have been thinned of redundant data points.

Men and Women Averages

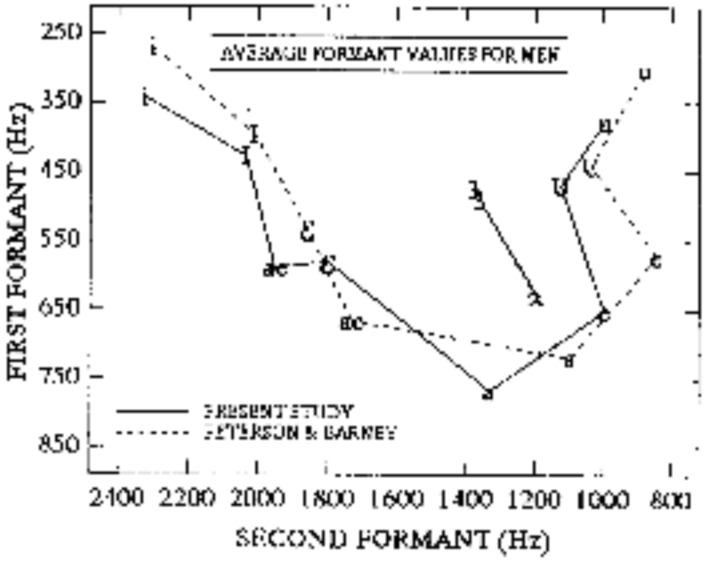


FIG. 5. Acoustic vowel diagrams showing average formant frequencies for men from the present study and from Peterson and Barney ("ae"=/æ/, "a"=/ɑ/, "c"=/ɔ/, "ʌ"=/ʌ/, "3"=/ɝ/).

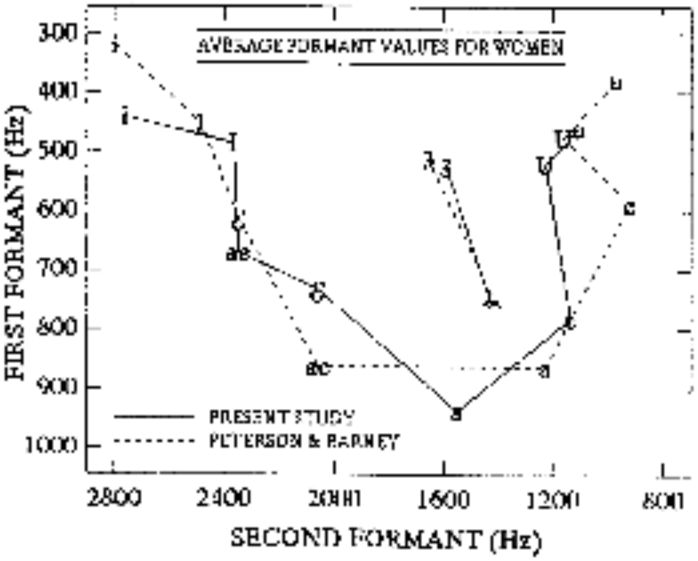


FIG. 6. Acoustic vowel diagrams showing average formant frequencies for women from the present study and from Peterson and Barney ("ae"=/æ/, "a"=/ɑ/, "c"=/ɔ/, "ʌ"=/ʌ/, "3"=/ɝ/).

3104 J. Acoust. Soc. Am., Vol. 97, No. 5, Pt. 1, May 1995

Hillenbrand *et al.*: Acoustic characteristics of vowels 3104

Duration

Ainsworth synthesized a steady-state vowel intermediate between /i/ and /I/. When duration was varied, long tokens were classified as /i/ and short tokens were classified as /I/.

Duration can be used by listeners.

Why are F1, F2 and F3 not “invariant”?

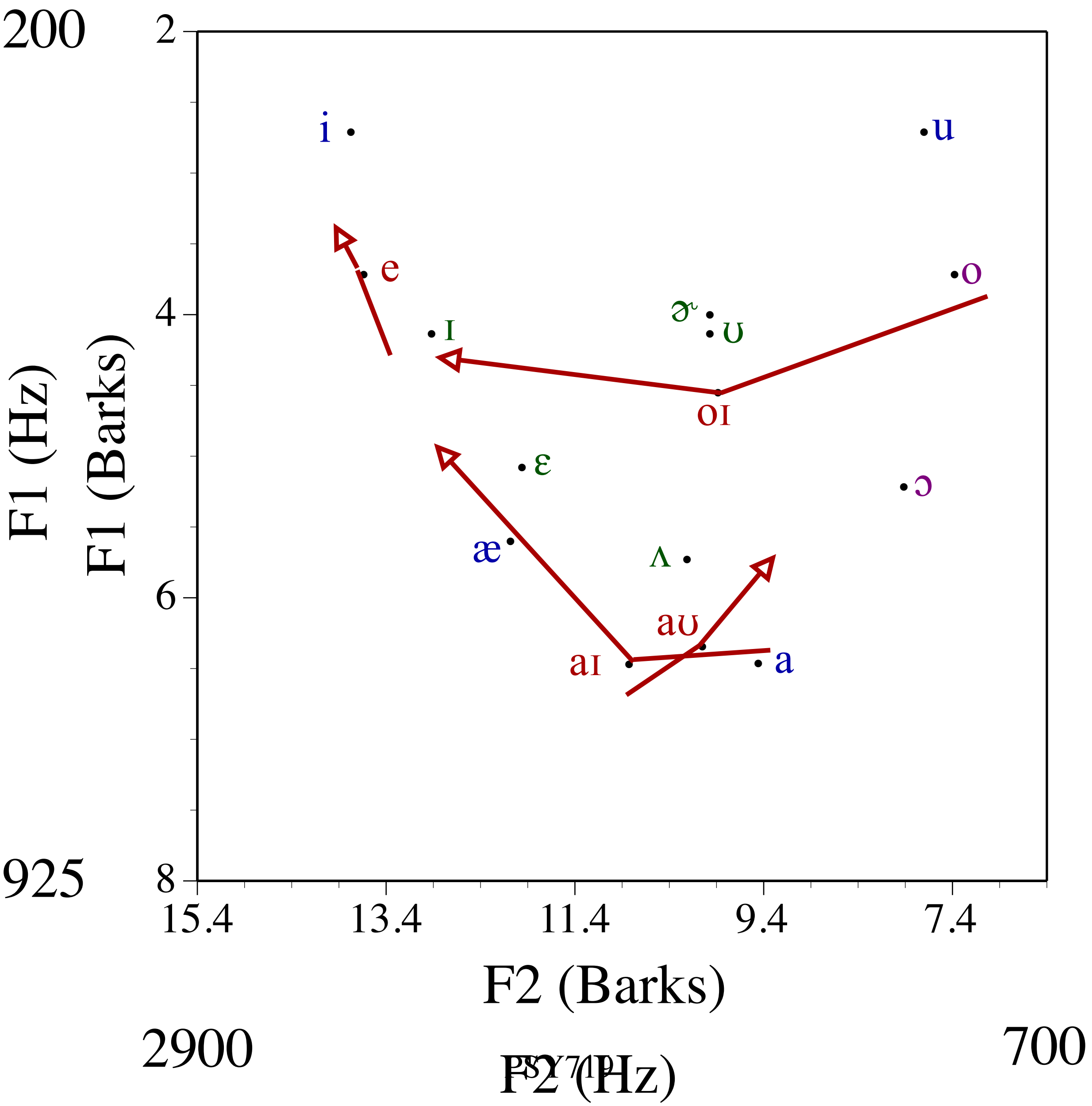
1) Coarticulation - Vowels formants do not hit their idealized values because of coarticulation with adjacent phonemes. /i/ in /bib/ and /did/ are different.

2) Formant frequencies vary from talker to talker and additional information is needed to “normalize”.

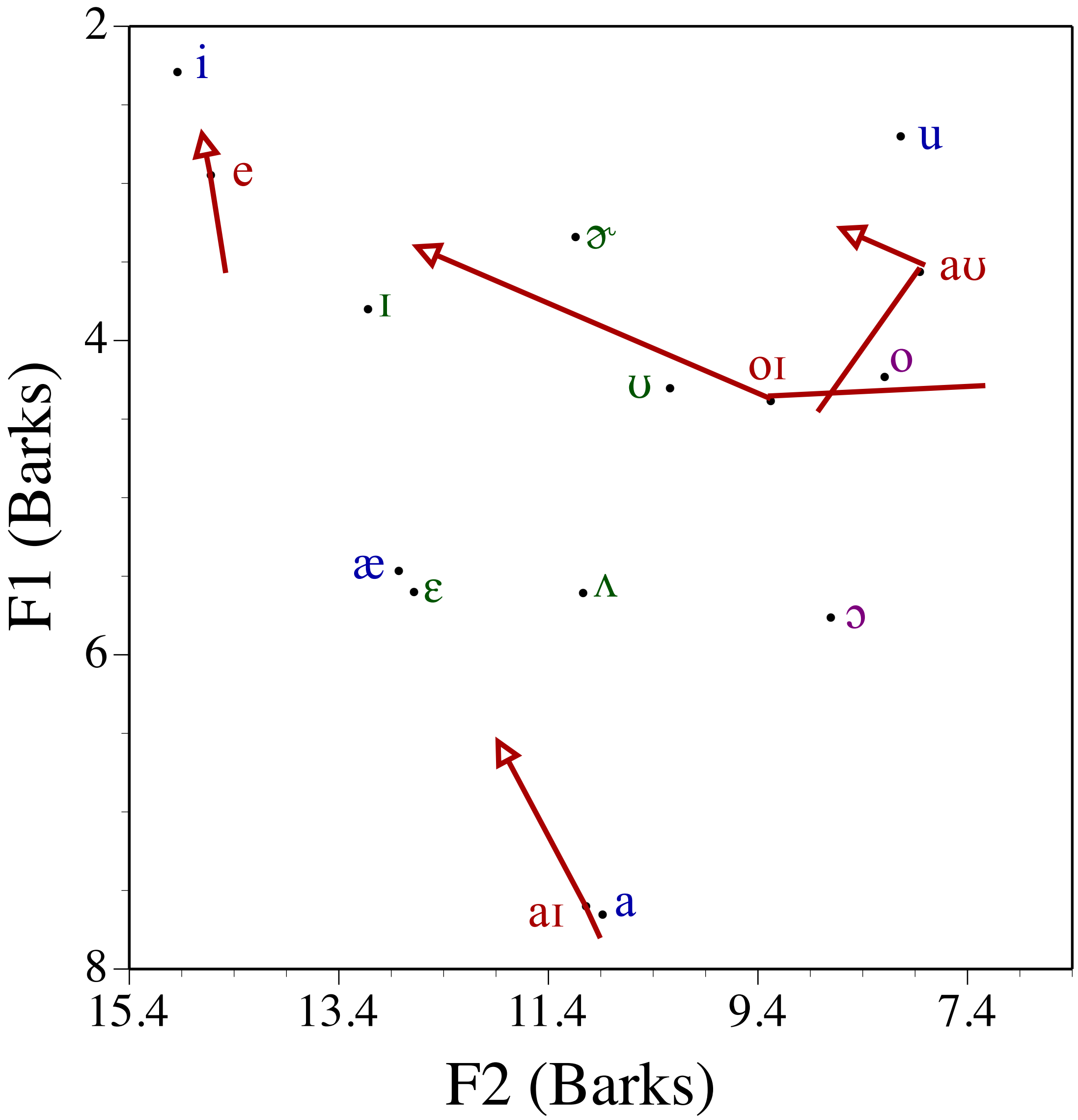
3) There are other acoustic correlates besides F1, F2, and F3.

4) Vowels are not processed with respect to steady-state values. Like stop-consonant place of articulation, vowels are inherently dynamic.

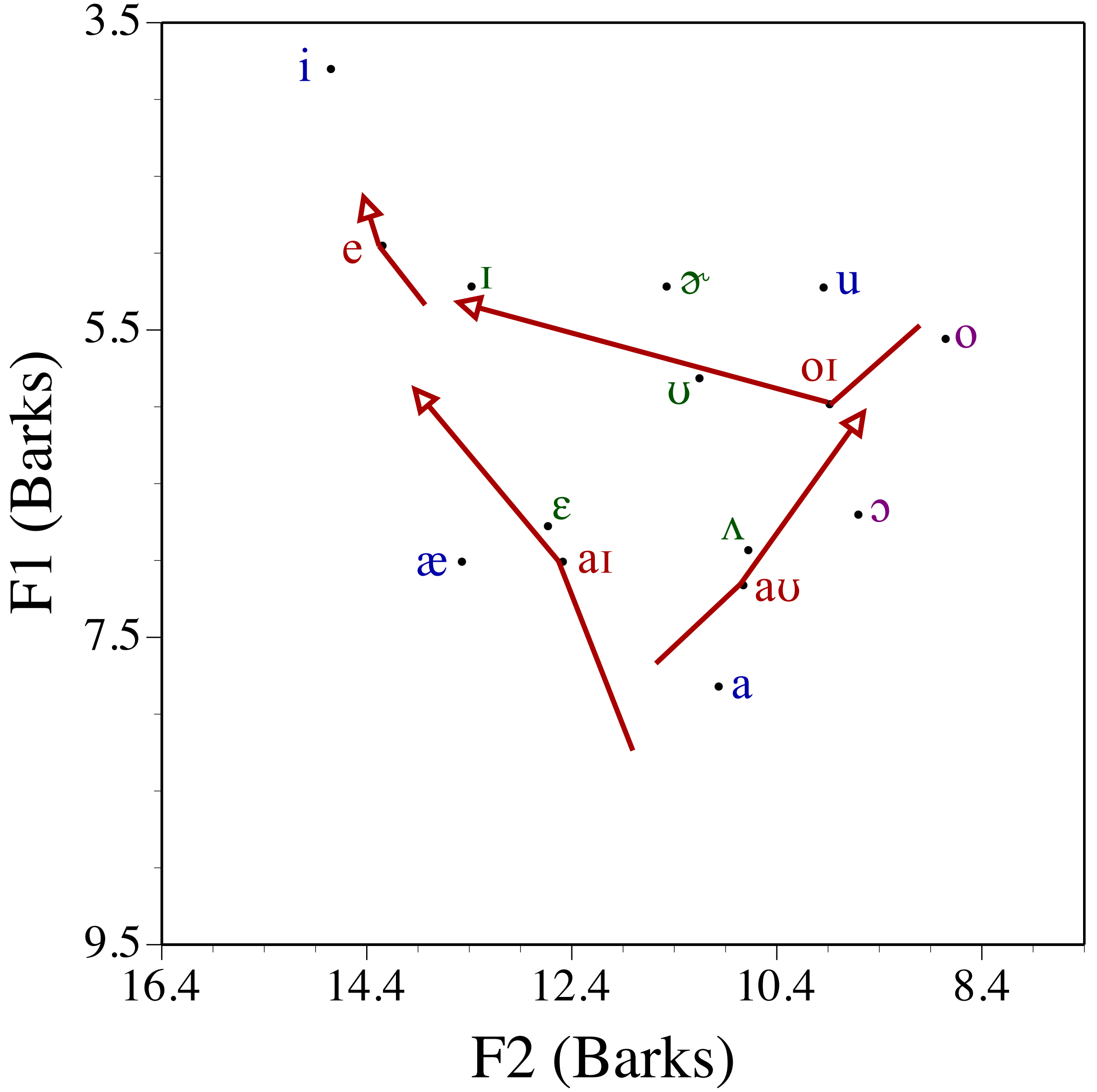
Vowel Space for JRS



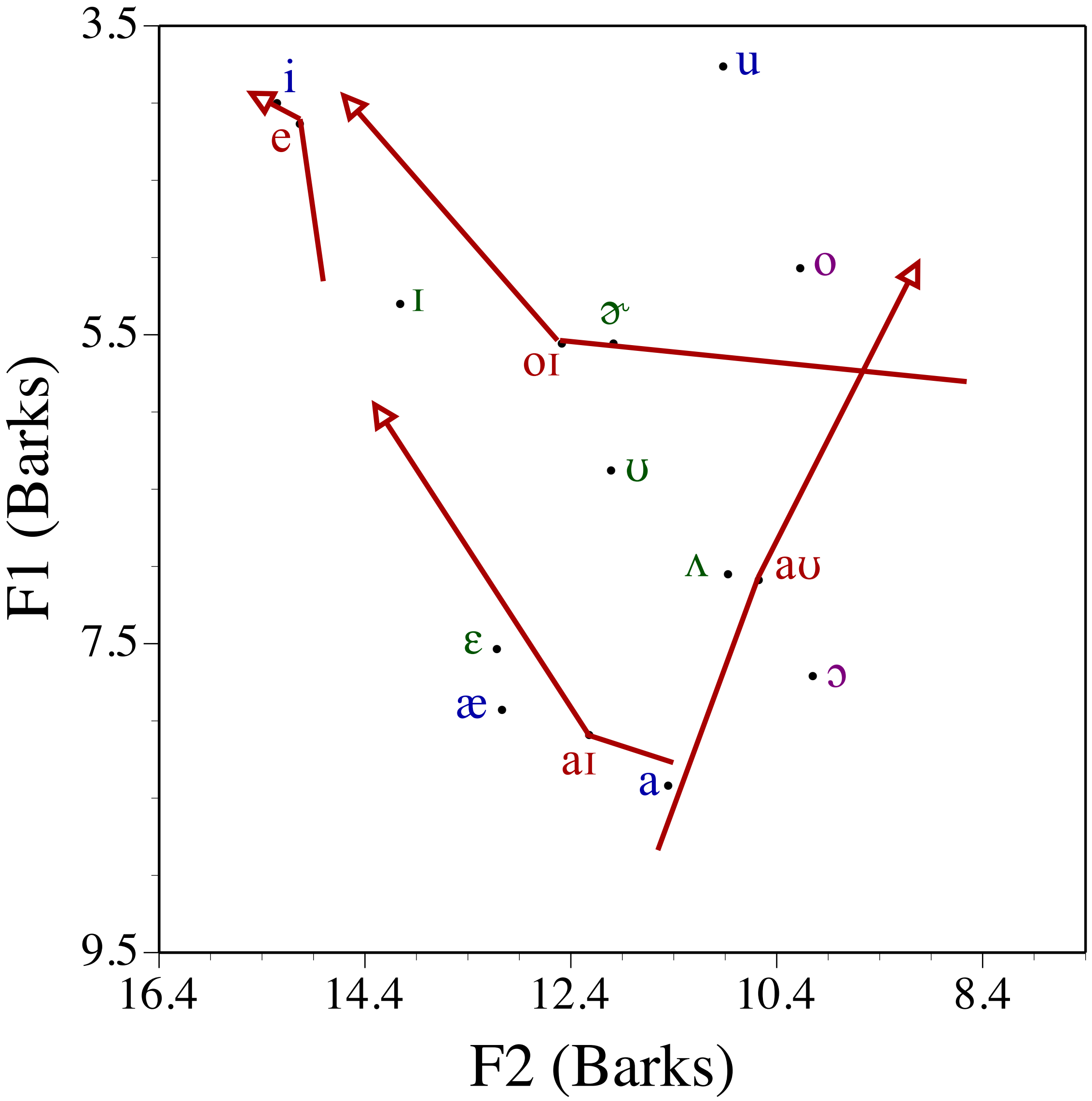
Vowel Space for CTM



Vowel Space for KMM



Vowel Space for LKZ



Acoustic Correlates to Vowel Identity

1) Static formant frequencies (F1, F2, F3 at vowel center)
2) Dynamic formant frequencies (direction and extent)
3) Duration
4) F0 (harmonic of F0 which aligns with F1)
5) Spectral Shape
6) Change in spectral shape

Nearey & Assman Results

Recorded isolated vowels. Edited short segments of beginning and end and played to listeners in different temporal arrangements.

- Beginning followed by end best identified.
- Just beginning or end intermediate.
- End followed by beginning *miss-identified most*.

Results show that listeners exploit dynamic information such as the extent and direction of formant movement.

Strange et al. Results

Recorded CVC syllables. Edited to produce vowel centers (removed initial and final transition portions) and “centerless” syllables (removed center, keeping transitions in their original temporal relationship).

Results show that listeners identified the vowel in the “centerless” syllables as well or better than the vowel centers.

Dynamic information appears critical to vowel identification.

Duration Revisited

Some speakers of Midwestern American English produce the vowels /ɛ/ and /æ/ such that the formant frequencies at the center of the vowel are nearly identical.

These same talkers produce the vowels so that there is a reliable difference in duration with the /æ/ longer and the /ɛ/ shorter.

F1, F2 and F3 for Talker M20

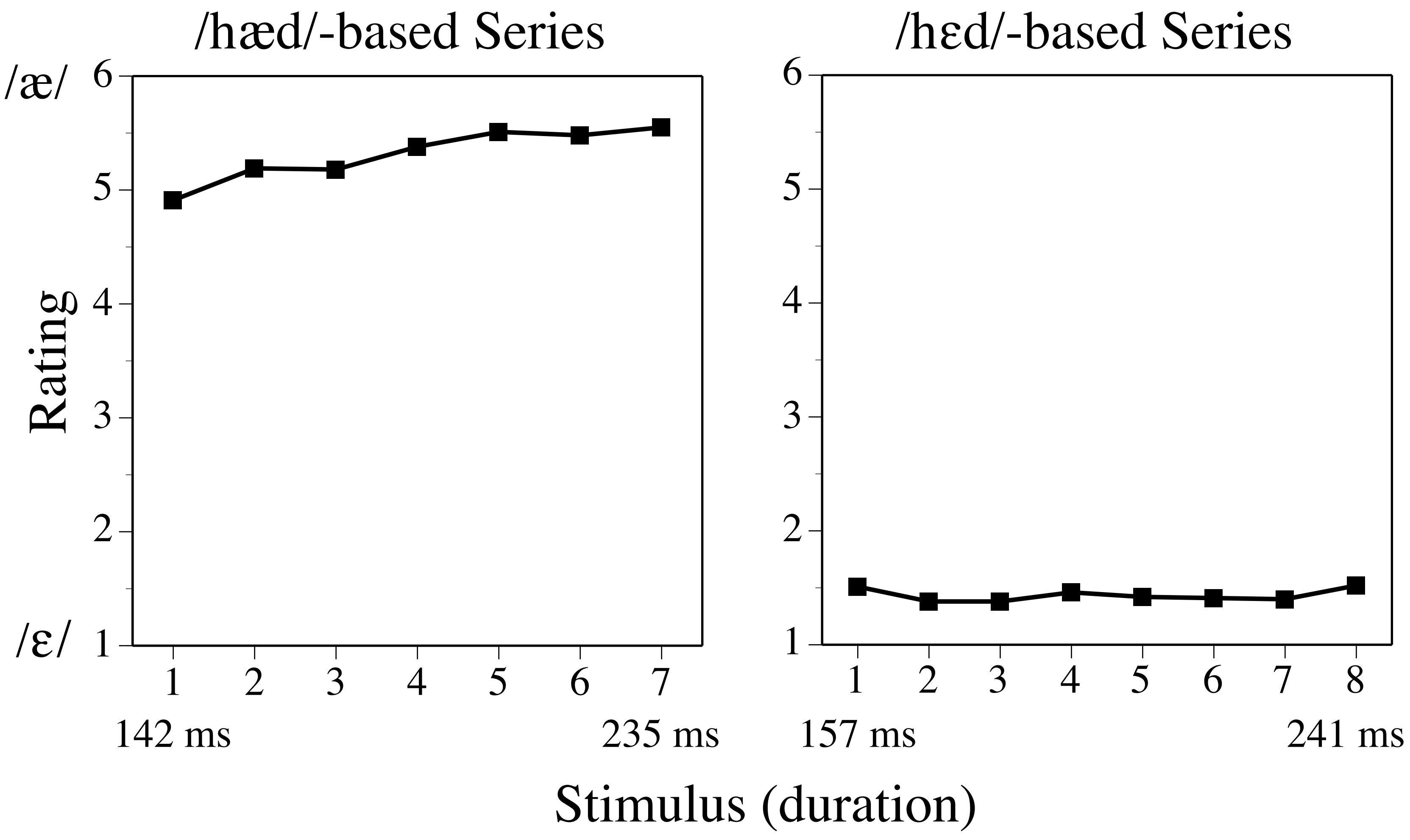
Token	Formant	Onset	Middle	End
/hæd/	F1	594	608	645
/hɛd/		601	607	565
/hæd/	F2	1850	1745	1618
/hɛd/		1860	1740	1690
/hæd/	F3	2360	2271	2360
/hɛd/		2400	2340	2440

Data Summary

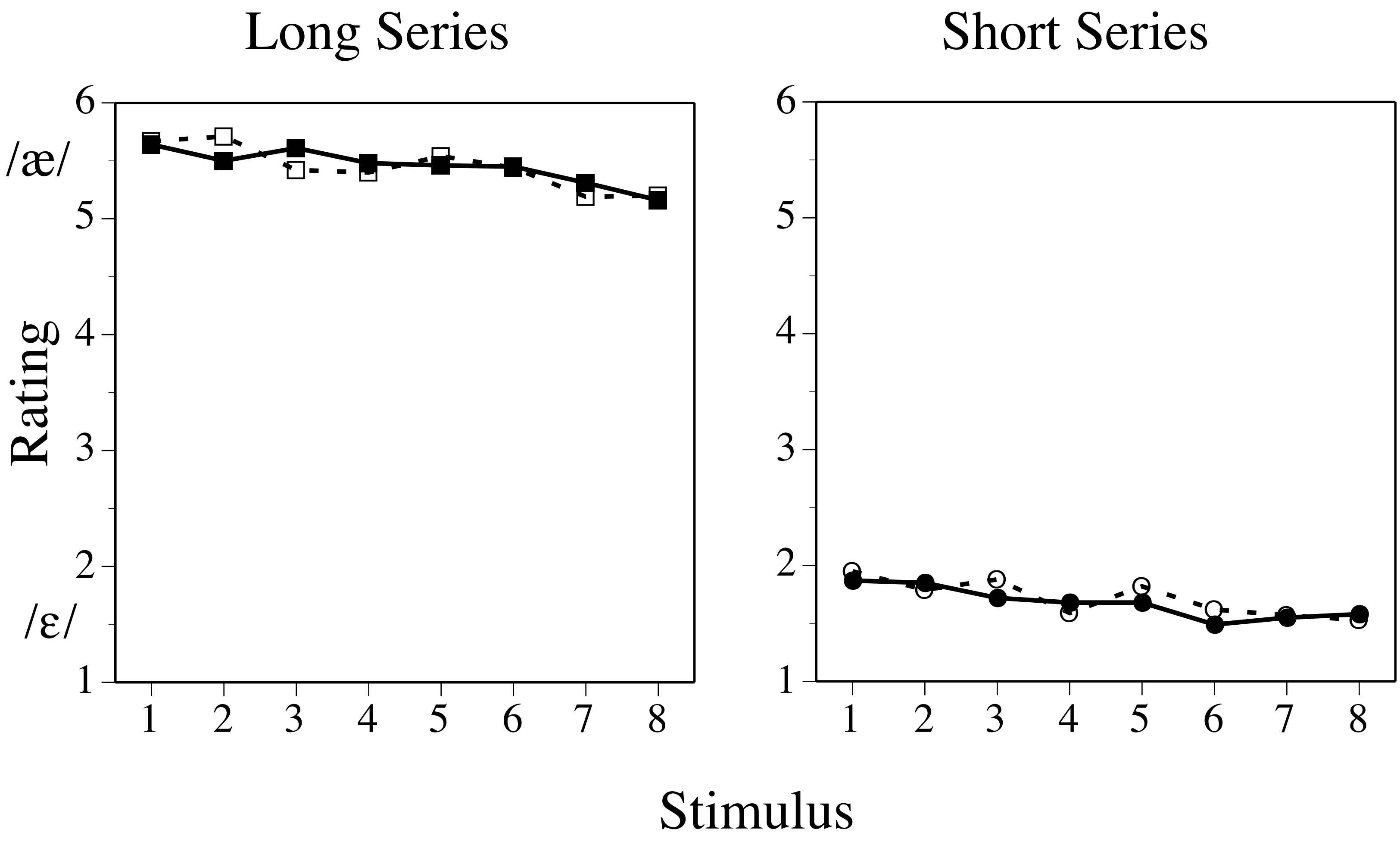
If the syllable /hæd/ is shortened, listeners continue to label it as /hæd/. If the syllable /hɛd/ is lengthened, listeners continue to label it /hɛd/. Even though the duration is distinctive in the talker's articulation, listeners do not use it.

If high quality synthetic versions of /hɛd/ and /hæd/ are created (with formant movements that mimic the originals and an F0 that is the average of the two), listeners use duration. Long syllables are labeled /hæd/ and short are labeled /hɛd/.

Talker 20 Tokens
Effect of vowel duration



Synthetic Tokens
Effect of Formant Frequencies

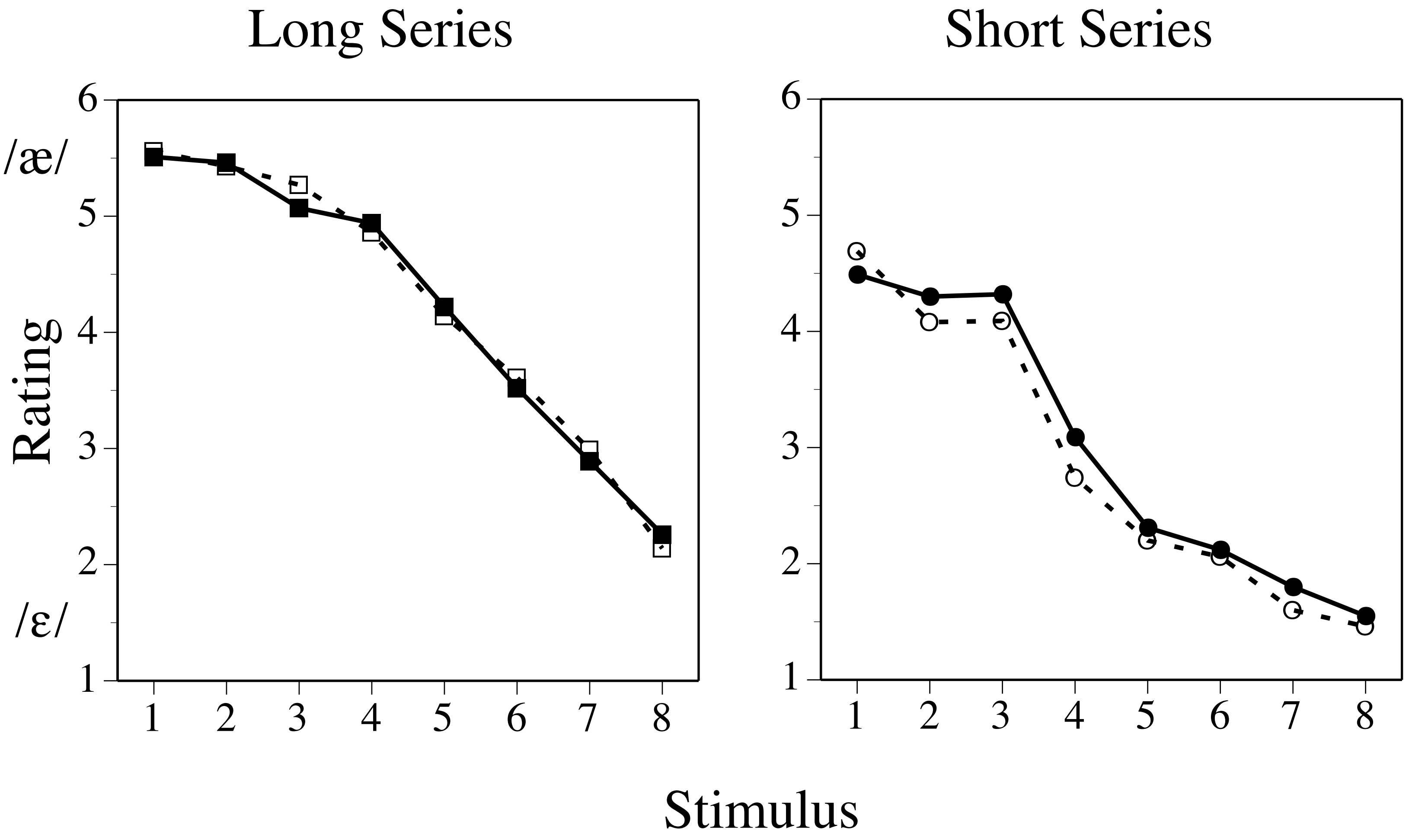


Data Summary - 2

Finally, if different F0s, mimicking the original syllables, are used, then listeners' labels follow the F0 and formants. With a low F0 (130 Hz), the stimuli are labeled /hæd/ and a higher F0 (152 Hz), they are labeled /hɛd/, regardless of duration.

Listeners appear to be using the amplitude of the harmonics of the fundamental. When the most intense harmonic is lower (3 times the fundamental), they identify the syllable as /hɛd/. When the most intense harmonic is higher (4 times the fundamental), the syllable is identified as /hæd/.

Synthetic Tokens
Effect of F0 & Formant Frequencies



Vowel Summary

Listeners use the peaks in the spectrum (the information in the signal related to the formants). Duration is a secondary cue.

The extraction of the “first formant” is probably done by resolving the most intense, low frequency harmonic of the fundamental. The other formants are probably simply peaks in the auditory spectrum.

Listeners use dynamic information and are more accurate in identifying dynamic vowels than steady-state vowels.

Vowel Summary - 2

If perception involves a speaker independent representation, then differences across talkers must be normalized using information such as F0, F4, the prior F1, F2, and F3 from the talker, and context to map the talker’ s vowel space.

Alternatively, listeners may map both the vowel and the talker simultaneously to a set of stored representations.

Models of Vowel Recognition

We will look at two different approaches:

Syrdal & Gopal - Uses F1, F2, F3, F4, and F0 at vowel midpoint.

Hillenbrand & Houde - Uses a normalized cross-section of the spectrum at multiple points (beginning, middle and end of vowel).

Syrdal & Gopal

This is a “classic” model. The acoustic cues to vowel recognition are F0, F1, F2, F3, and F4. The vowels are normalized to a fully abstract, talker independent representation.

Within each vowel, F0 and F4 represent “talker” qualities while F1, F2, and F3 are related to vowel identity. By using the distance between formants as the primary cue, talker differences are eliminated from the signal. Since this “normalizing” information is contained in each vowel, this is a form of intrinsic normalization.

Syrdal & Gopal - Processing Steps

1. Extract F0, F1, F2, F3 and F4 using a frequency scale that mimics human hearing (a bark scale is used).
2. Compute "peak differences" as F1-F0, F2-F1, F3-F2, F4-F2.
3. Map the peak differences onto stored representations of these values for each vowel (prototypes).
4. The prototype that fits best (least distance) is the vowel.

Hillenbrand

This model does not use formants and avoids problems associated with extracting formants from a speech signal.

Each vowel is represented in memory as a sequence of spectral slices (3 to 5) or sections. The auditory system computes this for an incoming signal and the sections are compared to stored representations (prototypes) for men, women and children for each vowel. That is, this model "normalizes" by using separate memory representations.

Hillenbrand - Processing Steps

1. Compute spectral cross-sections at 5 points in vowel.
2. Compute similarity (distance) to prototypes. Note that there are 3 prototypes for each vowel (men, women, children).
3. Prototype that fits best (least distance) is vowel.

Performance

Syrdal & Gopal, using only a single, static representation, is 80-85% correct for hVt syllables. Confusions are generally neighboring vowels that often differ in duration.

Hillenbrand & Houde, using only a single spectral section at the center of the vowel, is 80% correct for hVt syllables. Using 3 slices (beginning, middle, end) it is 92% correct. Confusions are generally neighboring vowels that often differ in duration plus some “odd” errors (e.g. /i/ and /u/).