A Multi-Tier Framework

for

Understanding Spoken Language

Steven Greenberg

http://www.icsi.berkeley.edu/~steveng
steveng@cogsci.berkeley.edu

Acknowledgements and Thanks

Research Funding

U.S. Department of Defense (before Iraq)

U.S. National Science Foundation

Research Collaborators

Hannah Carvey, Shawn Chang, Ken Grant, Leah Hitchcock, Joy Hollenback, Rosaria Silipo

For Further Information

Consult the web site:

www.icsi.berkeley.edu/~steveng

Simplify, simplify

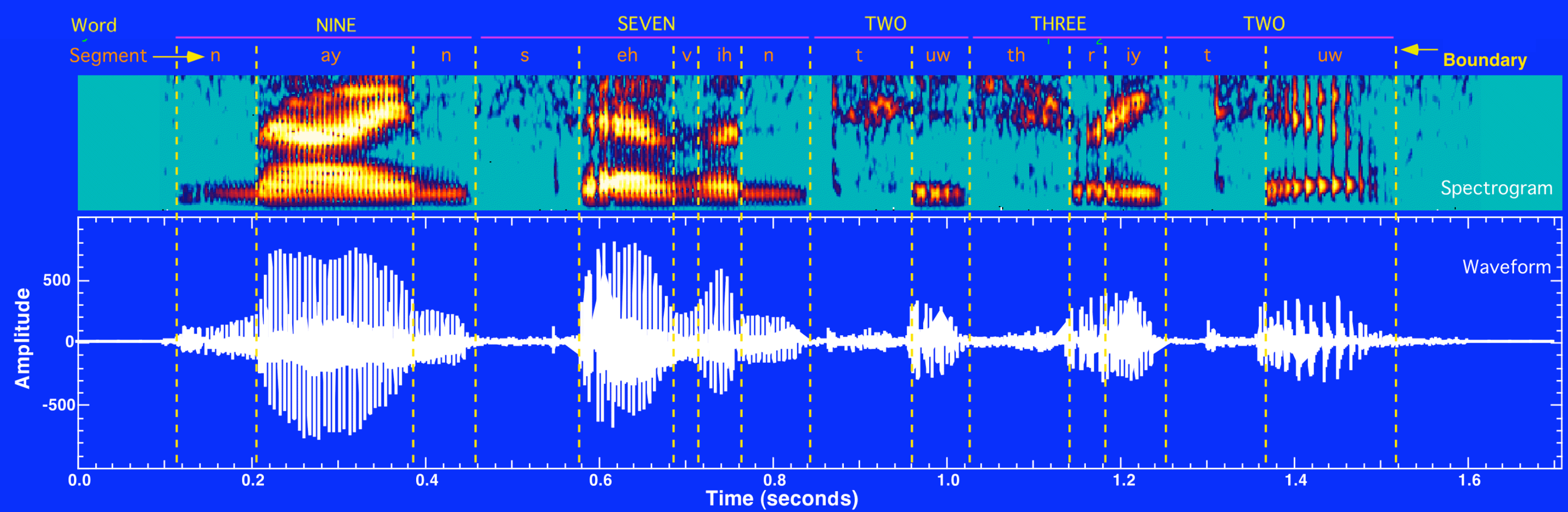
"Make things as simple as possible – but no simpler"

Albert Einstein

Speech Analysis – The Traditional Perspective

Traditionally, spoken language has been analyzed as a sequence of words, each containing a set of phonemes, organized like "beads on a string"

Such a "linear" structure provides a seemingly transparent means with which to analyze and characterize the speech signal, as shown below

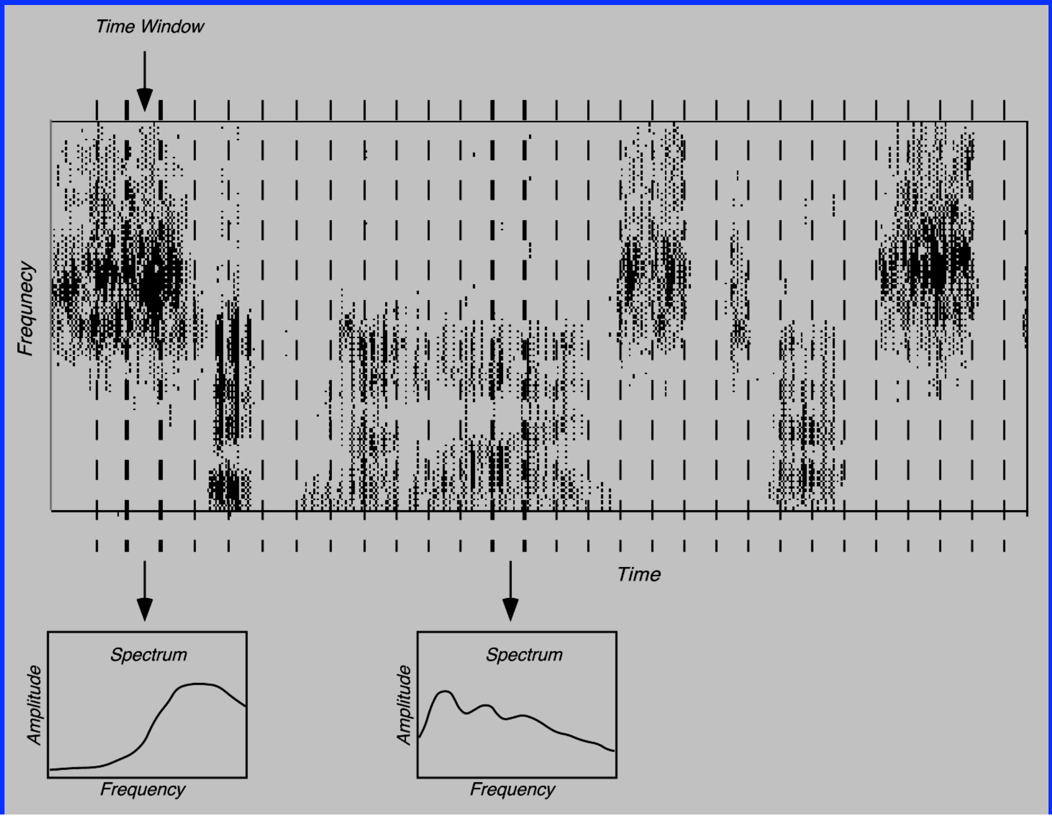


The Serial Frame Analysis Perspective

Within this serial framework, the signal is spectrally analyzed in an "egalitarian" manner

All time frames are created equal (usually 25 ms long, with 10-ms slide intervals)

This method of analysis is relatively transparent to perform, as it requires no <u>a priori</u> knowledge of the signal

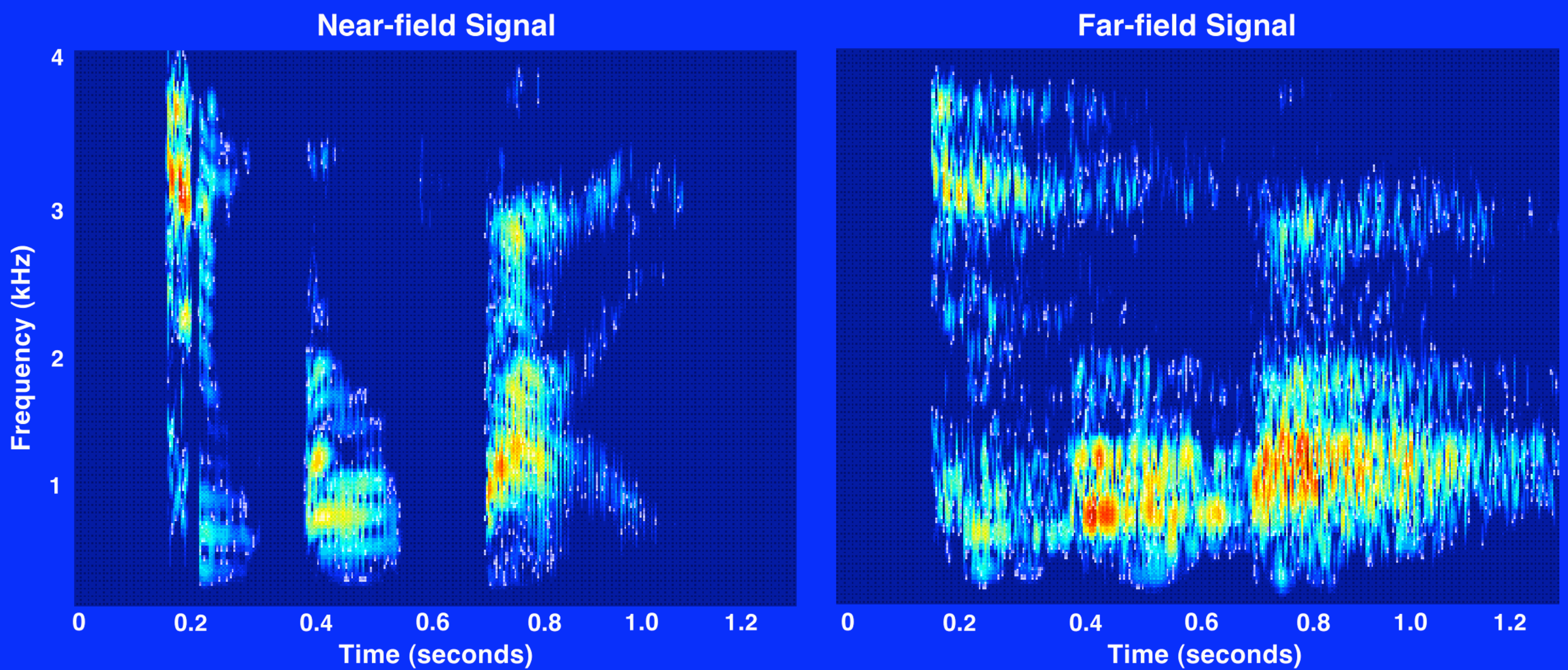


Challenge # 1 – Environmental Variability

As seductive as this egalitarian framework may be, there are four principal problems with this approach

First, the spectro-temporal properties of speech are highly variable

This variability reflects the specific nature of the acoustic environment, an example of which is shown below for a speech signal recorded at two different microphone positions in the same room



Challenge #2 – Pronunciation Variation

Second, the pronunciation of words varies A LOT, with many canonical phones (a.k.a. phonemes) "deleted," as in the word "and" (Switchboard)

N	Pronunciation
82	ae n
63	eh n
45	ix n
35	ax n
34	en
30	n *Canonical pronunciation*
20	ae n dcl d
17	ih n
17	q ae n
11	ae n d
7	q eh n
7	ae nx
6	ae ae n
6	ah n
5	eh nx
4	uh n
4	ix nx
4	q ae n dcl d
3	eh n d
3	q ae nx

N	Pronunciation
3	eh
2	ae n dcl
2	ae
2	ax m
2	ax n d
2	ae eh n dcl d
2	eh n dcl d
2	ax nx
2	q ae ae n
2	q ix n
2	ix n dcl d
2	ih
2	eh eh n
2	q eh nx
2	ix d n
1	eh m
1	ax n dcl d
1	aw n
1	ae q
1	eh dcl

Pronunciation Variation is Common

The variability observed occurs in most words spoken, and is not confined to just a few variants, as shown in this table pertaining to Switchboard material

Rank	Word	N	#Pron	MCP %Total	Most Common Pronunciation
1	I	649	53	53	ay
2	and	521	87	16	ae n
3	the	475	76	27	dh ax
4	you	406	68	20	y ix
5	that	328	117	11	dh ae
6	a	319	28	64	ax
7	to	288	66	14	tcl t uw
8	know	249	34	56	n ow
9	of	242	44	21	ax v
10	it	240	49	22	ih
11	yeah	203	48	43	y ae
12	in	178	22	45	ih n
13	they	152	28	60	dh ey
14	do	131	30	54	dcl d uw
15	so	130	14	74	s ow
16	but	123	45	12	bcl b ah tcl t
17	is	120	24	50	ih z
18	like	119	19	46	l ay kcl k
19	have	116	22	54	hh ae v
20	was	111	24	23	w ah z

Greenberg (1999)

The 20 most frequency words account for 35% of the lexical occurences

Challenge #3 – Variation in Time and Spectrum

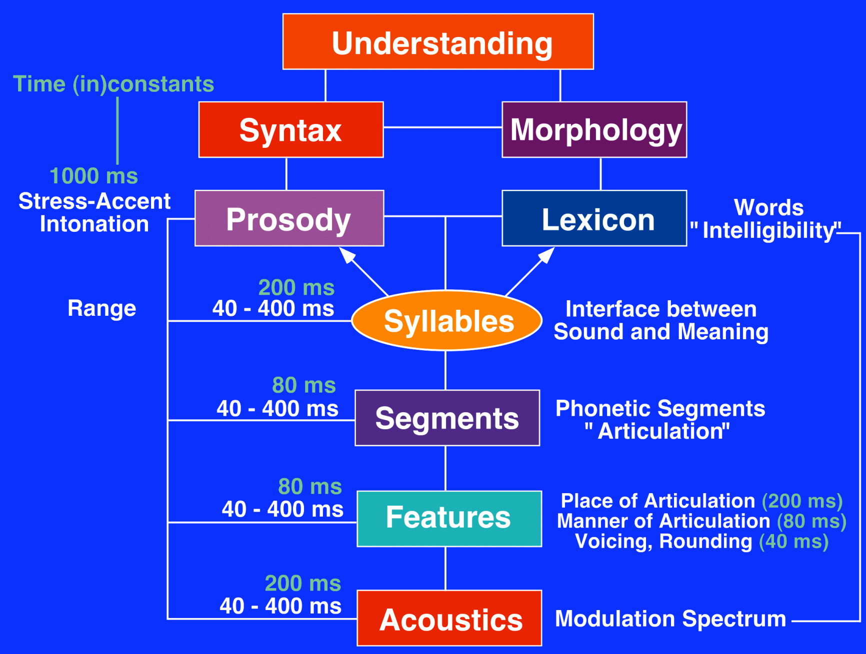
Third, the "units" of spoken language vary with respect to duration, frequency and space, thus

Certain properties are inherently SHORT in duration, or require FINE TEMPORAL RESOLUTION to adequately characterize – e.g., VOICING

Others are inherently of LONGER duration, such as PROSODIC elements

While others are INTERMEDIATE in length, such as PHONETIC SEGMENTS

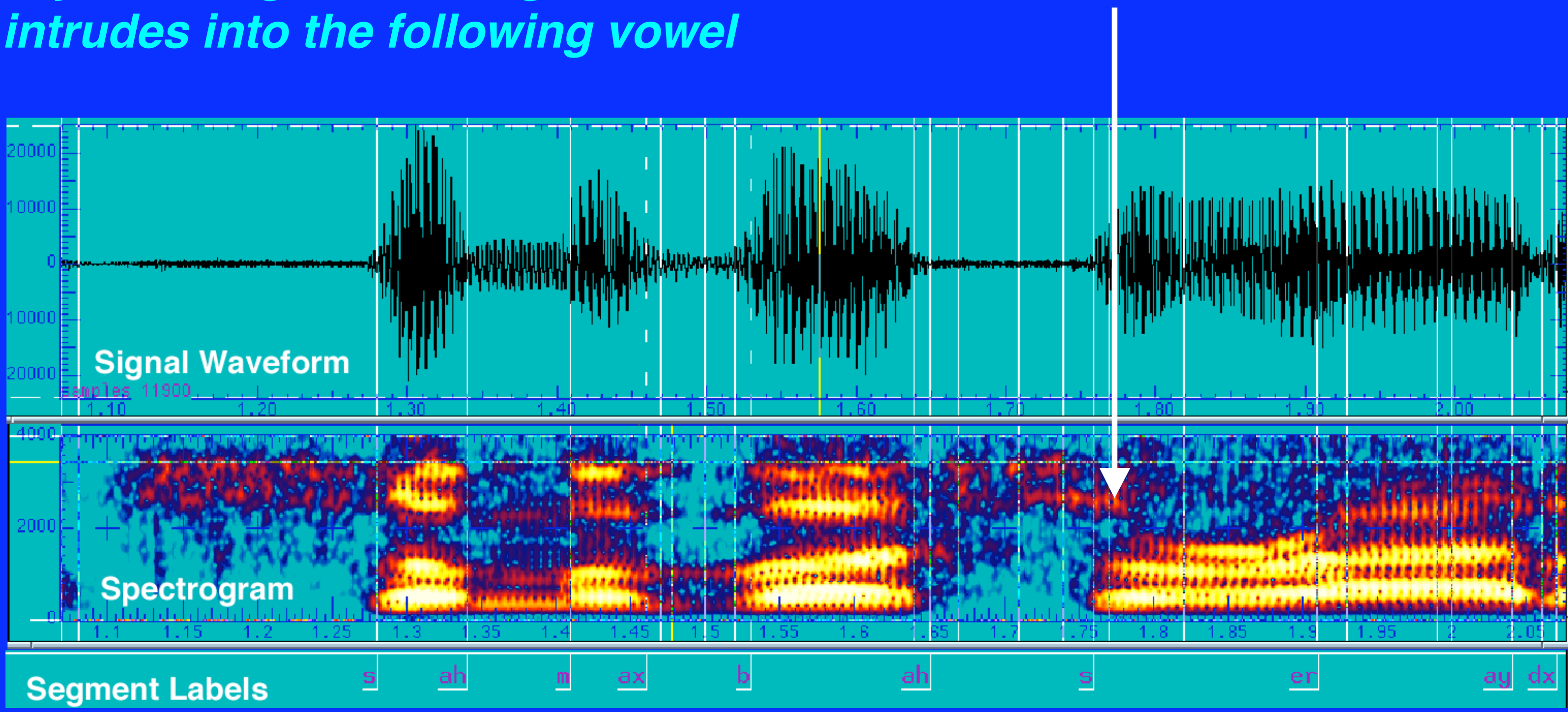
Hence, THERE IS NO SINGLE TIME INTERVAL that adequately captures all of the important acoustic and linguistic properties of spoken language



Challenge #3 – Variation in Time and Spectrum

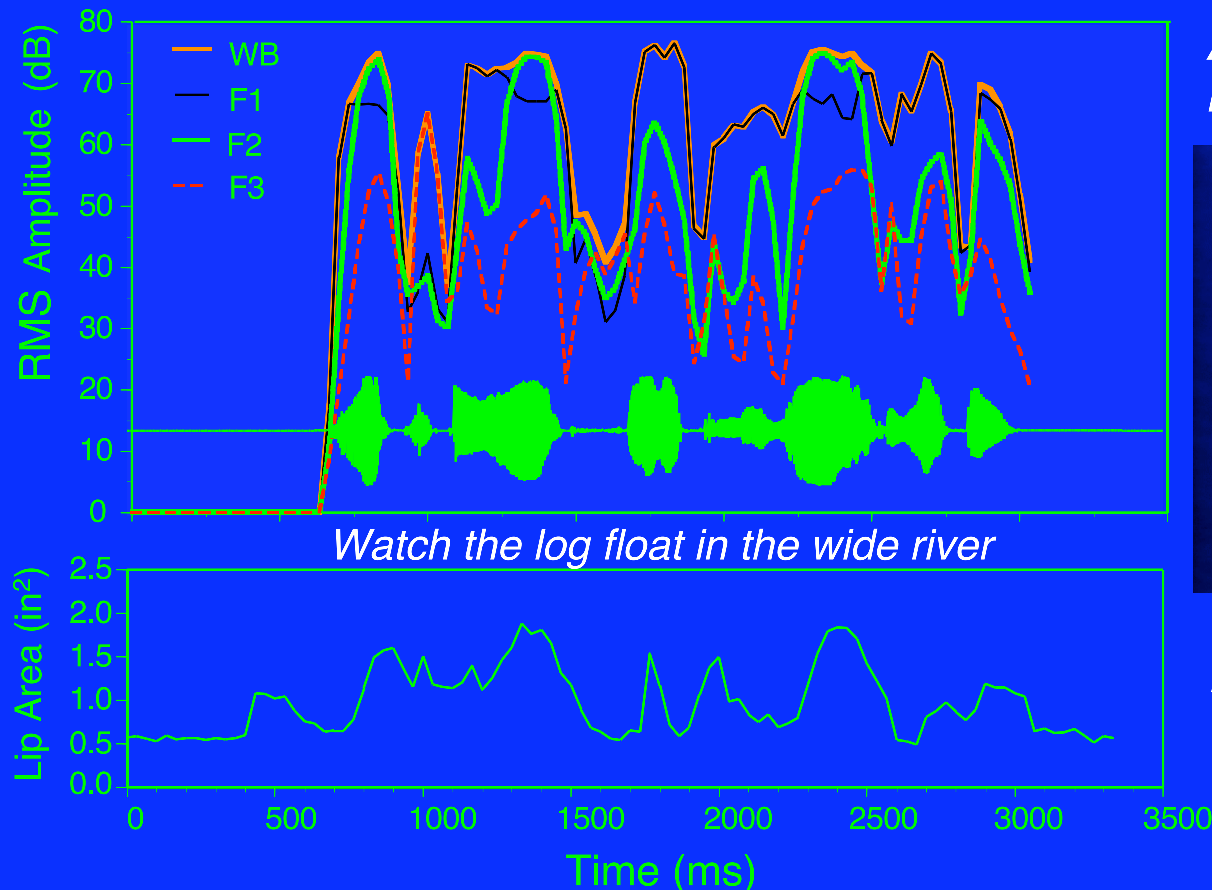
Moreover, the manner in which linguistic information is distributed across (spectral) frequency and time is NON-UNIFORM

Some of the acoustic properties associated with a phone "bleed" into adjacent segments – e.g., note the frication of the second [s] below, which intrudes into the following vowel



Challenge #4 – Importance of Vision

Further complicating the picture is the importance of visual information derived from movement of the lips, jaw and tongue, as well as other facial features – such information serves to constrain and enhance the interpretation of the acoustic signal



Amplitude Fluctuation in Different Spectral Regions

Lip Aperture Variation

Data courtesy of Ken Grant

What to Do?(with respect to speech robustness)

In the remainder of this talk, I shall focus on describing a multi-tier framework for spoken language

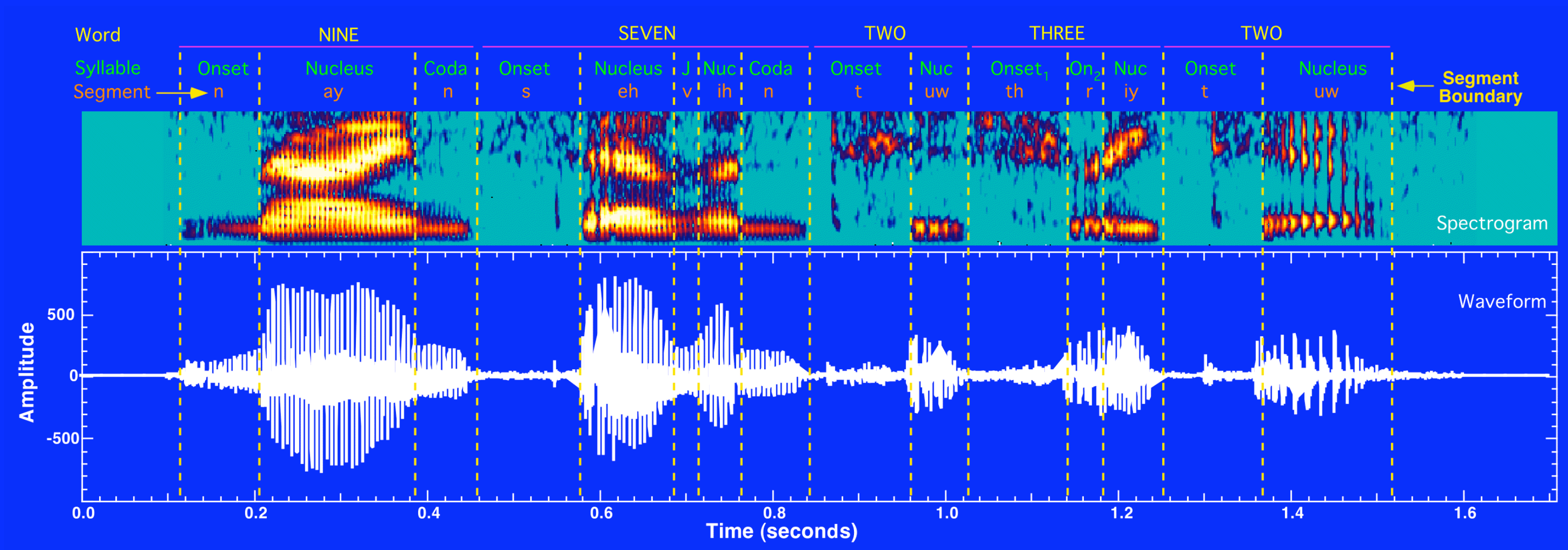
This framework is intended to explain how spoken language is processed by the (human) brain

And to use such knowledge (and insight) for developing noise-robust methods in speech technology

The following slides summarize the essence of my presentation ….

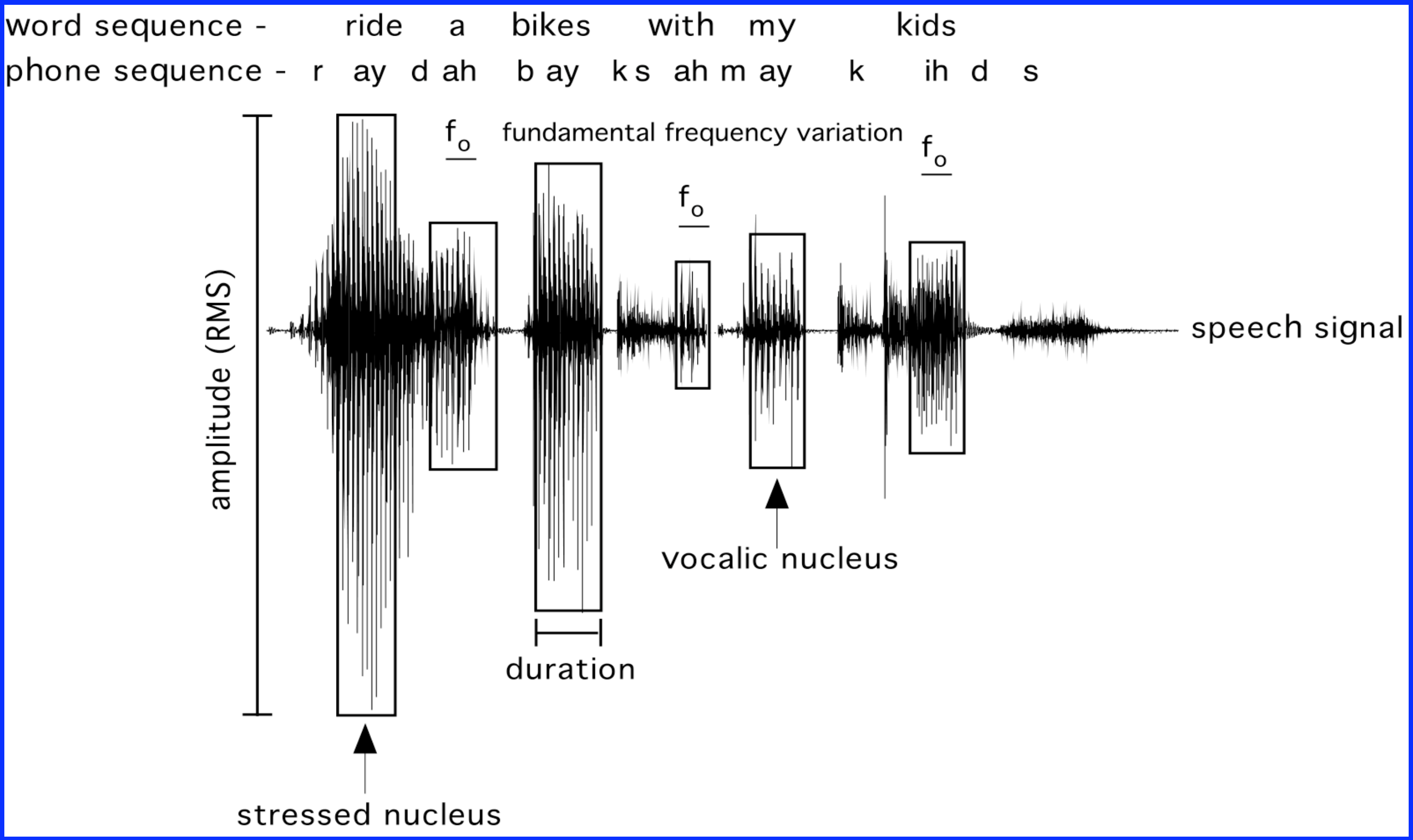
Take Home Messages

The SYLLABLE, rather than the PHONE, is the most basic organizational unit of spoken language – the patterns of pronunciation variation observed are incompatible with phonetic segment-based models



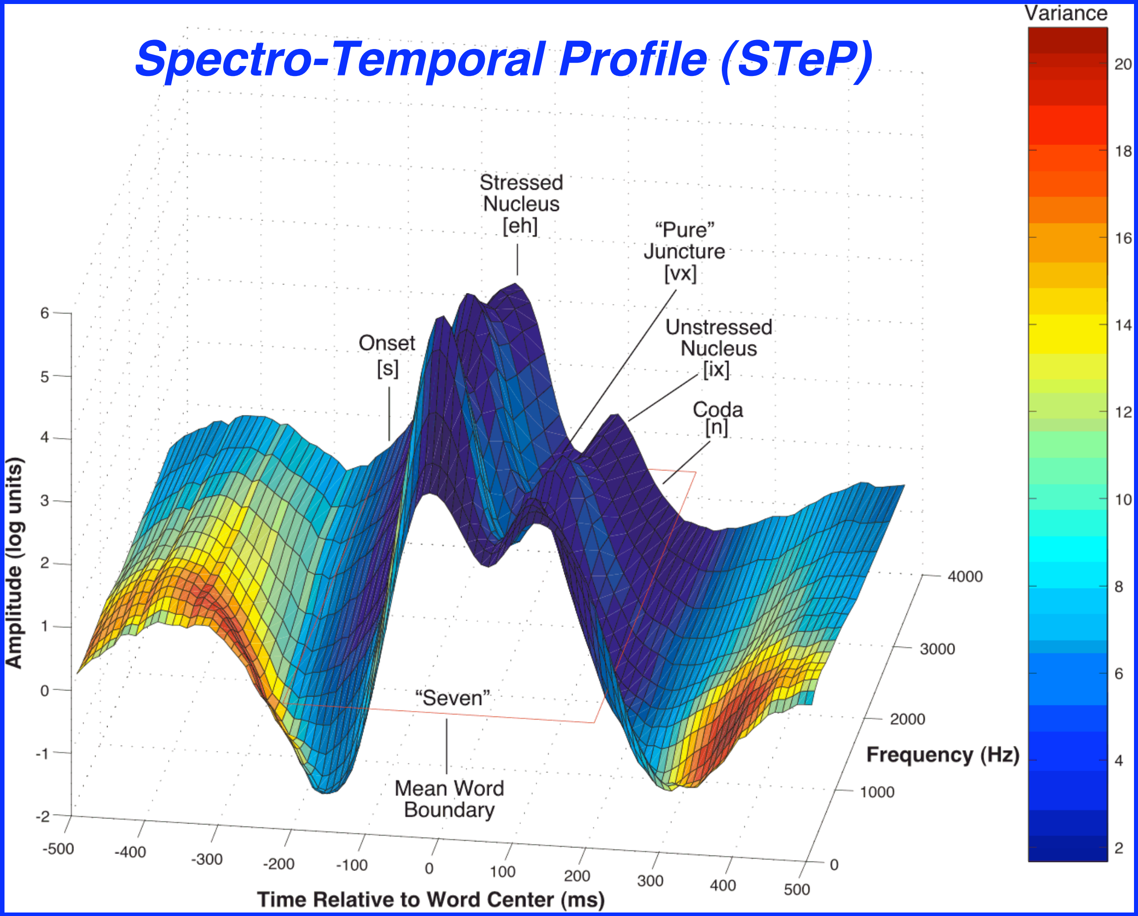
Take Home Messages

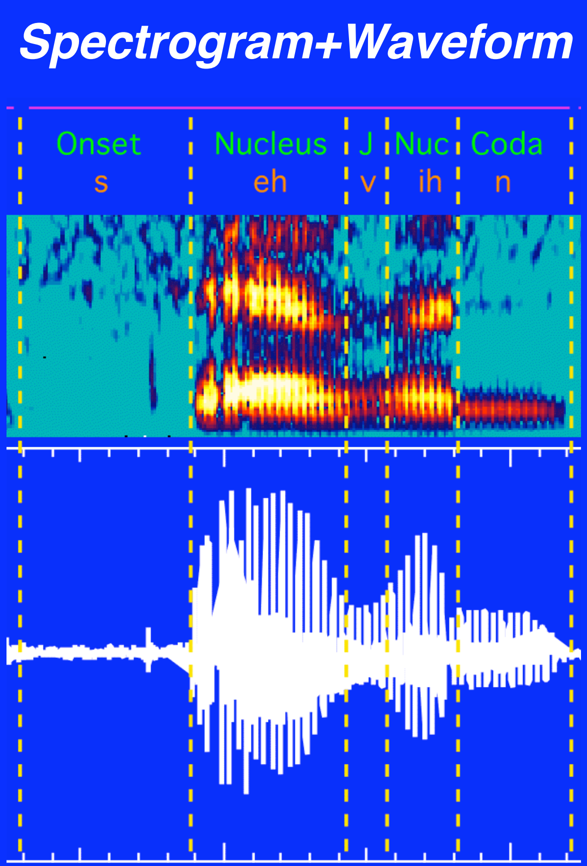
The syllable carries prosodic weight (a.k.a. "accent" or "prominence") that affects the manner in which its constituents are phonetically realized



Take Home Messages

- ***The behavior of these syllabic constituents (a.k.a. "ONSET," "NUCLEUS" and "CODA") differ dramatically from each other, and influence the phonetic character of the syllable***
- ***Syllable position is probably as important as segmental identity for characterizing pronunciation***

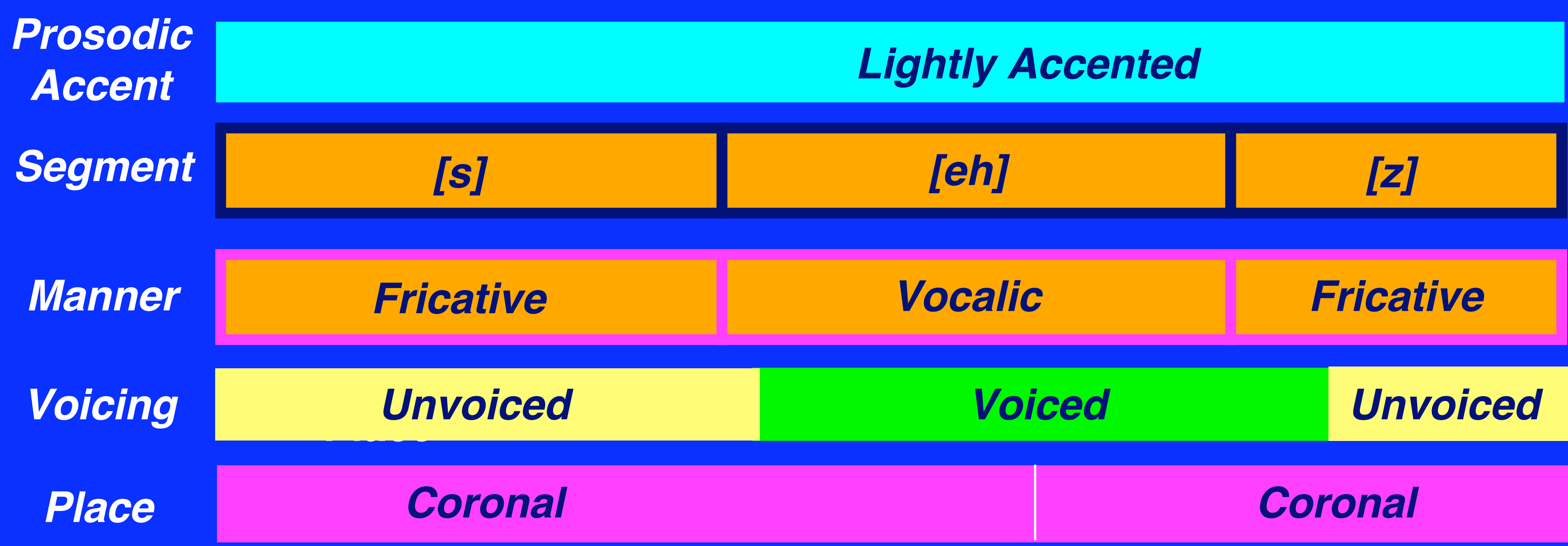




Greenberg et al. (2003)

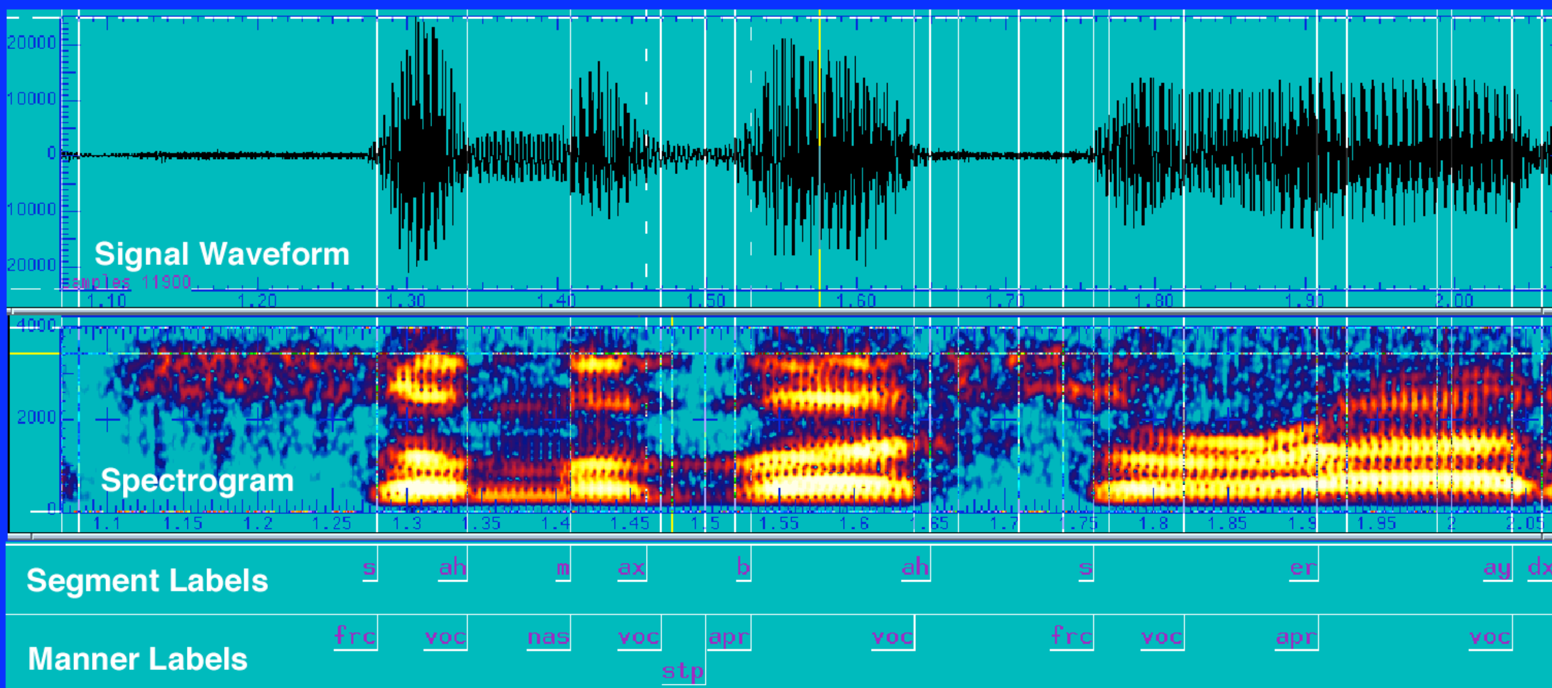
Take Home Messages

The MICROSTRUCTURE of the syllable can be delineated in terms of articulatory-acoustic features (e.g., voicing, articulatory manner and place)



Take Home Messages

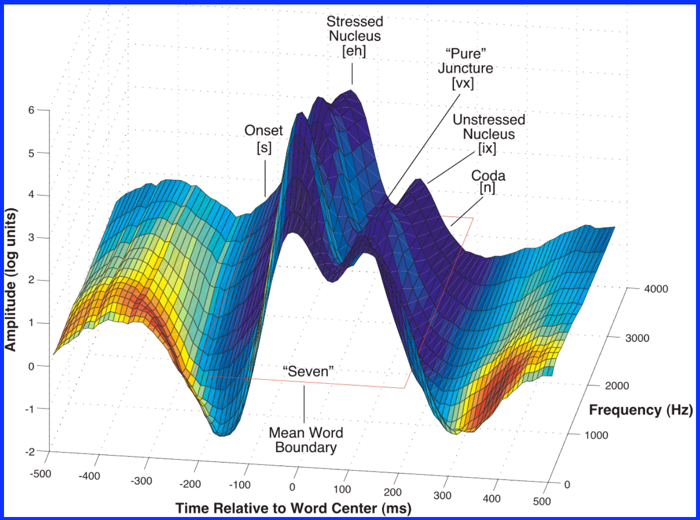
***MANNER* of articulation most closely parallels (in time and behavior) the classical concept of the *phonetic segment* and sets the basic intensity mode for the sequence of syllabic constituents (a.k.a. the "ENERGY ARC")**



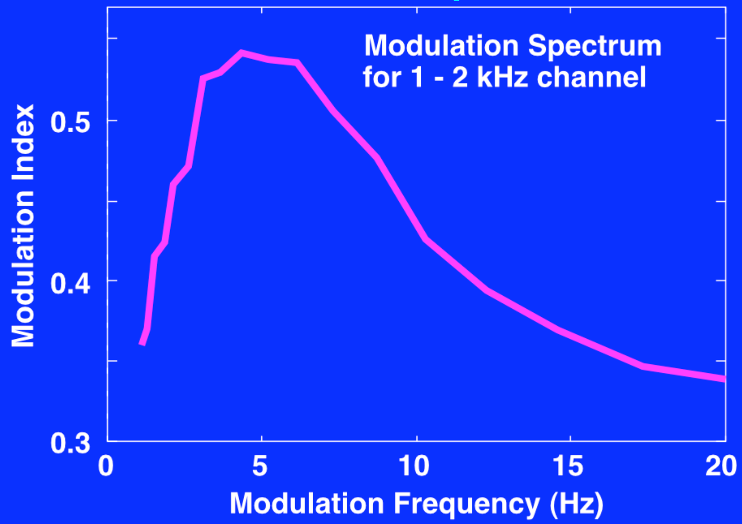
Take Home Messages

The ENERGY ARC reflects cortical processing constraints on the acoustic (and visual) signal associated with the MODULATION SPECTRUM

Spectro-temporal Profile



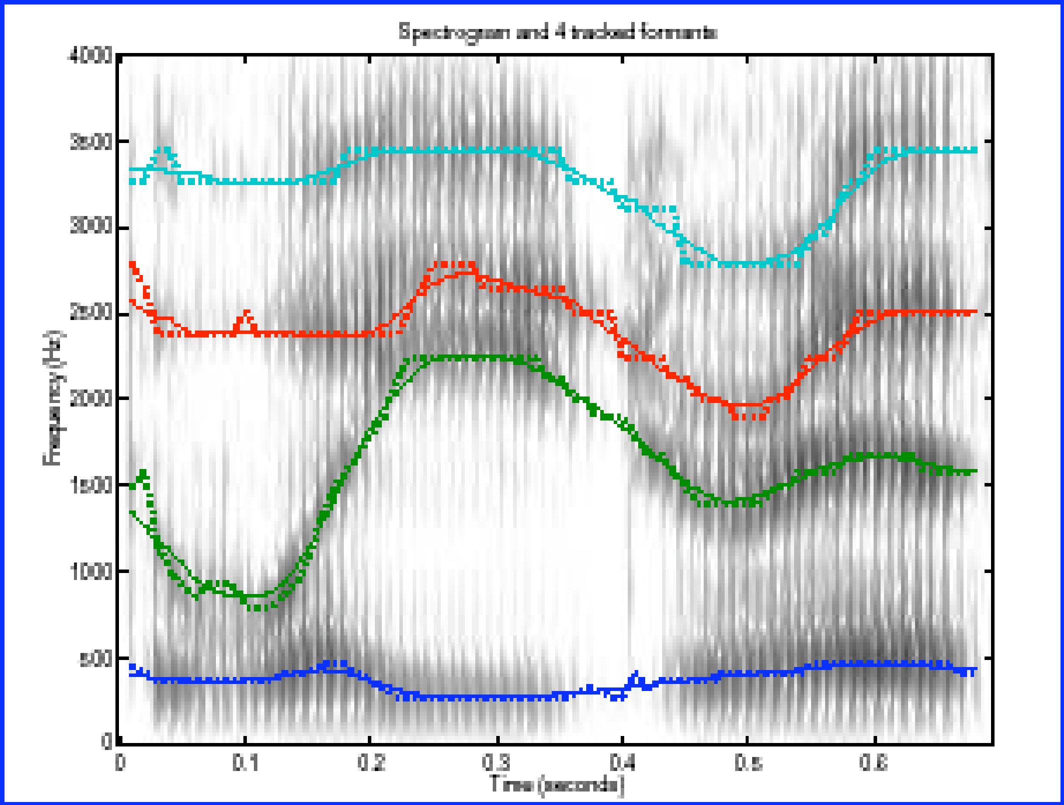
Modulation Spectrum



Take Home Messages

***PLACE* of articulation is the most information-laden articulatory feature dimension in speech, and is inherently *TRANS-SEGMENTAL*, binding vocalic nuclei with preceding and following consonants**

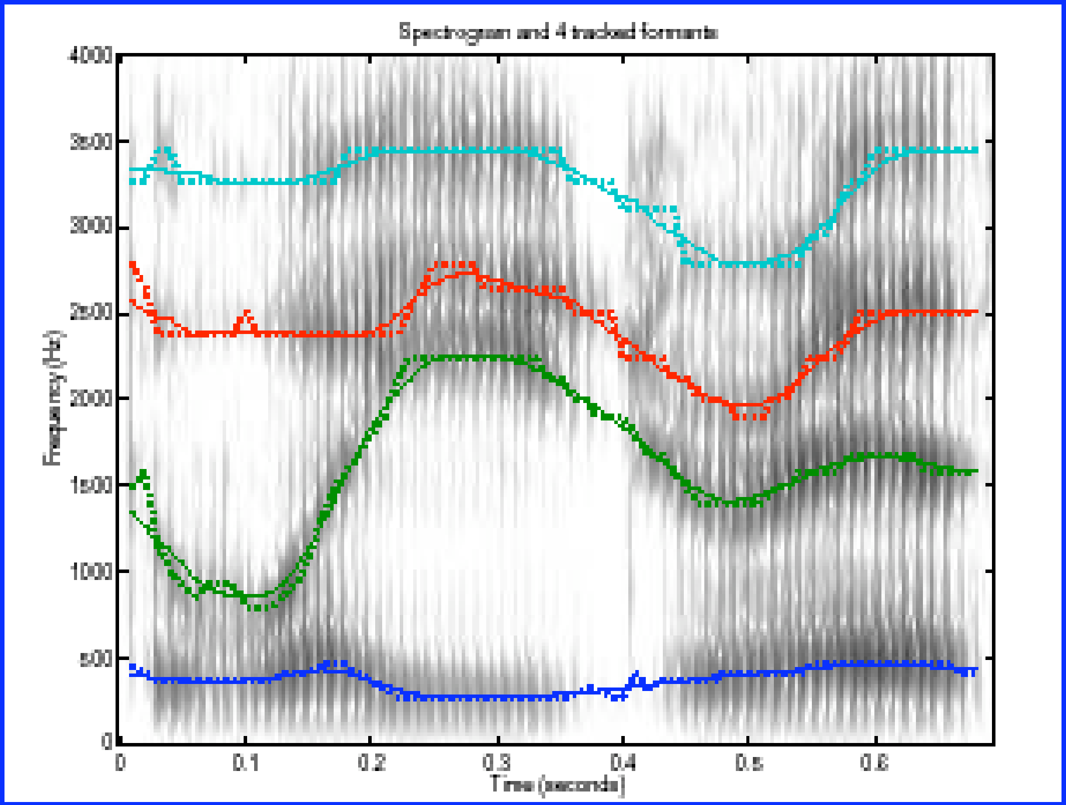
It is also the most stable phonetic dimension linguistically, although it is extremely vulnerable to acoustic interference when presented solely in the acoustic modality



Take Home Messages

The acoustic vulnerability of place of articulation cues implies that the classically cited basis for this information – formant transitions – provide inherently weak cues and often do not play a decisive role

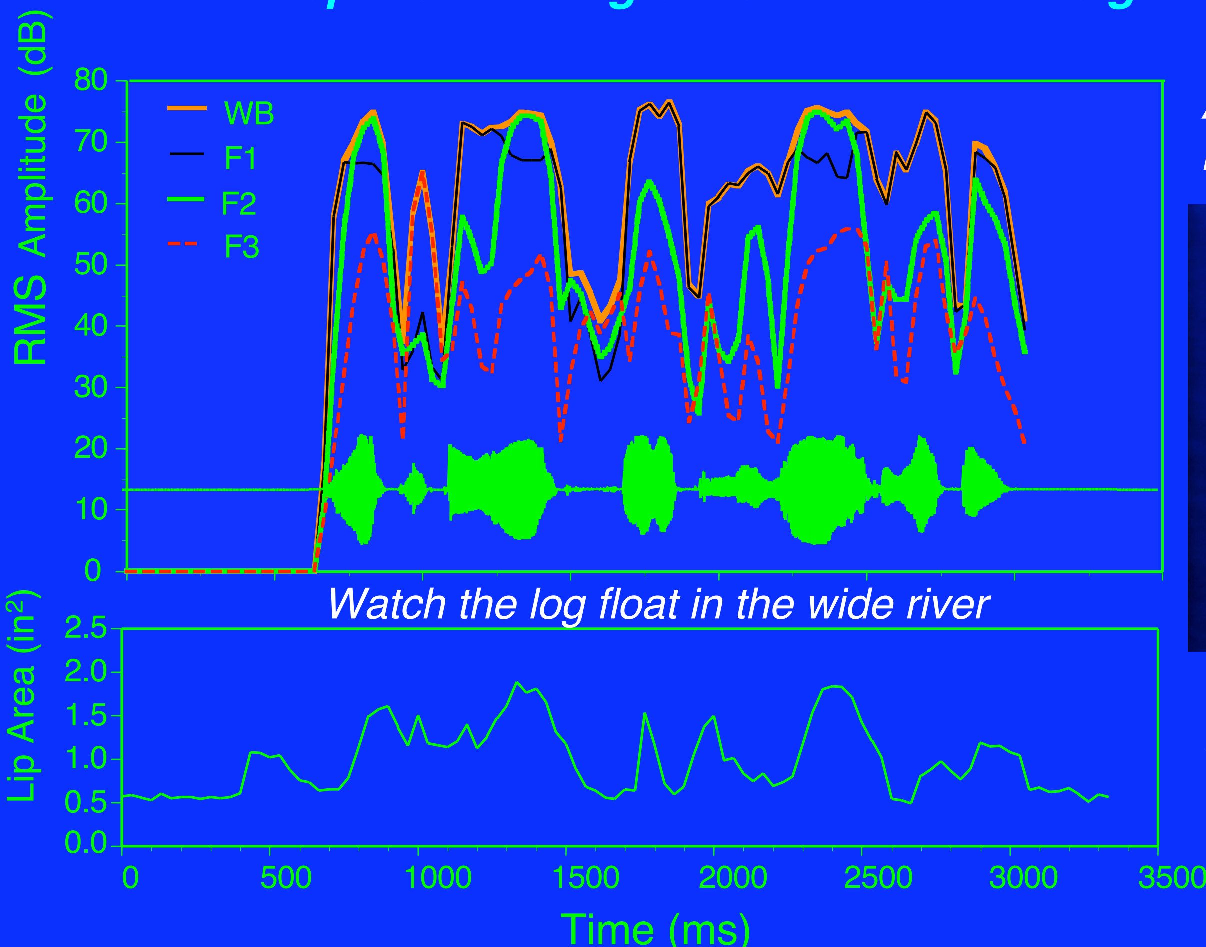
This counterintuitive observation implies that some other information source is often decisive for decoding articulatory place information, particularly in noisy environments



Take Home Messages

***PLACE* of articulation information is derived in part from *visual cues* associated with the movement of the lips, tongue and jaw during face-to-face interaction**

The robustness of articulatory place cues largely stems from its bi-modal nature –speechreading cues enhance the signal-to-noise ratio by ca. 10 dB



Amplitude Fluctuation in Different Spectral Regions

Lip Aperture Variation

Data courtesy of Ken Grant

Take Home Messages

Articulatory PLACE provides the primary discriminative (entropic) basis for lexical identity, and is therefore important to model accurately

(which means that the visual, speechreading cues can not be neglected)

Take Home Messages

VOICING ***emanates from the nucleic core of the syllable and spreads both forward (toward the coda) and backward (toward the onset), the degree of temporal spreading reflecting the magnitude of*** *prosodic prominence* ***– in this sense, VOICING is a*** *SYLLABIC* ***rather than a phonetic-segment feature, in that it is sensitive to the prominence of the syllable***

voiced *voiced* *voi* *voiced*

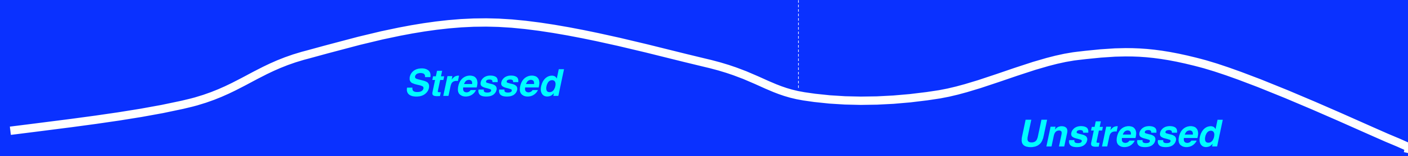
Take Home Messages

It is the PATTERN of INTERACTION among articulatory-feature dimensions across time that imparts to the syllable its specific phonetic identity

WORD – "Strengthen"

	SYLLABLE – "streng"					SYLLABLE – "then"		
	ONSET			NUCLEUS	CODA	ONSET	NUCLEUS	CODA
Segment	s	t	r	ε	ŋ	θ	ɪ	n
Manner	Fric	Stop	Rhotic	Vowel	Stop	Fric	Vowel	Nasal
Place	ø	Central	ø	Front	Back	Central	Front	Central
Height	ø	ø	ø	Mid	ø	ø	High	ø
Voicing	–	–	+	+	+	–	+	+
Duration	170 (ms)			80	60	60	30	50

Energy Contour

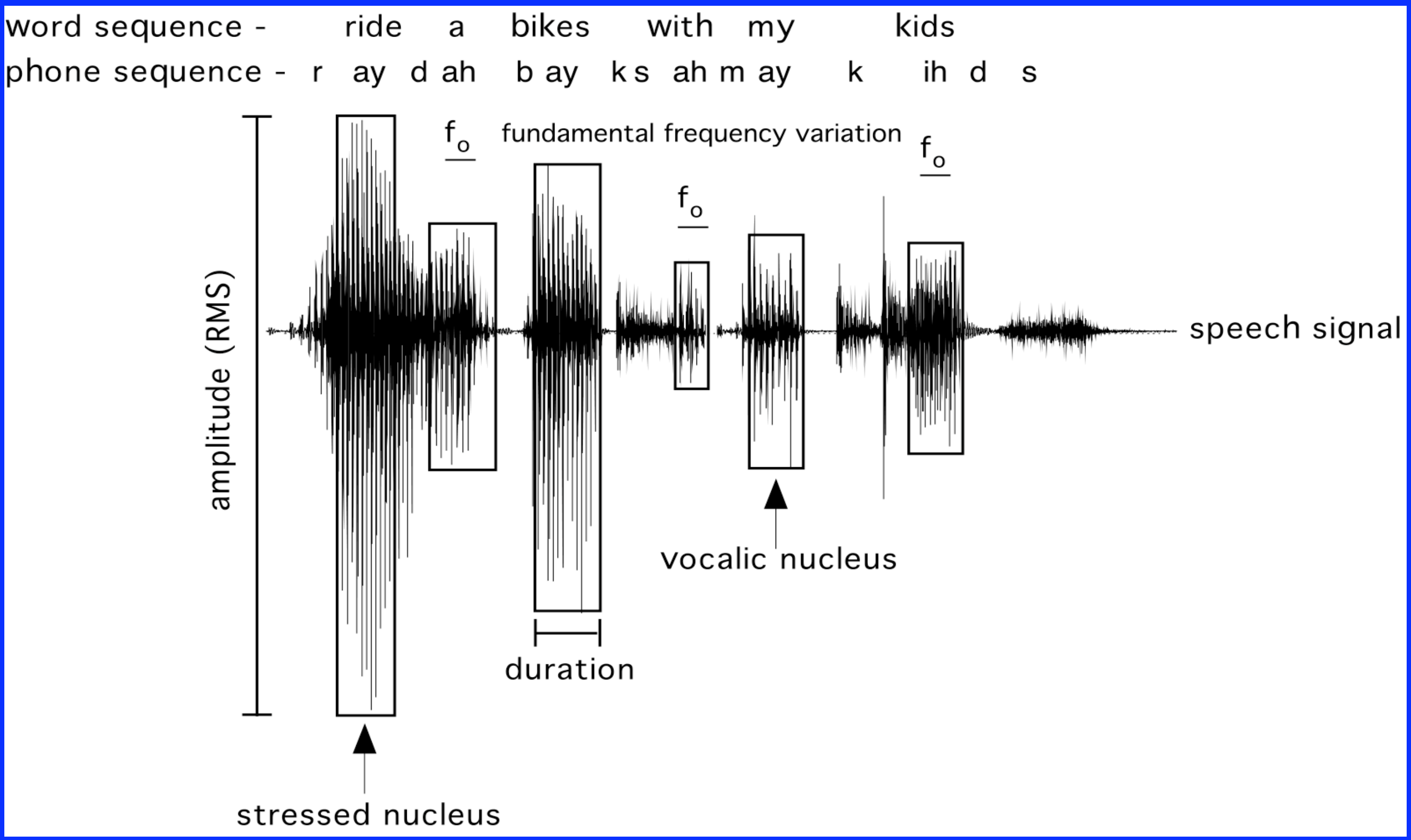


Take Home Messages

The specific REALIZATION of ARTICULATORY FEATURES is governed by prosodic PROMINENCE, as well as their POSITION within the SYLLABLE

Take Home Messages

The PROSODIC pattern reflects INFORMATION contained within the utterance



Take Home Messages

Therefore, it is ultimately INFORMATION (and lexical distinctiveness) that governs the detailed phonetic properties of spoken language

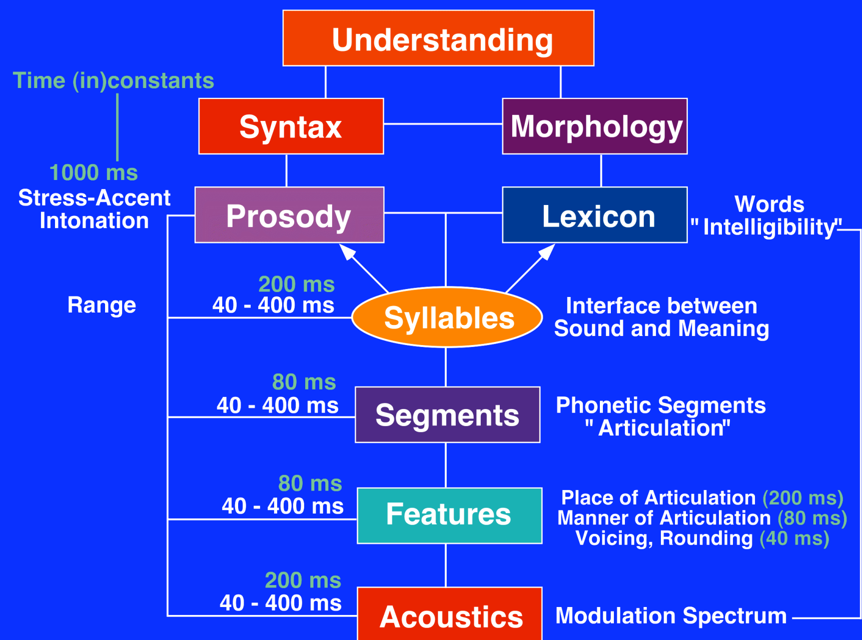
A Vision of the Future

A Vision of the Future for Speech Technology

Multi-tier, entropy-based analysis

Unification of linguistic tiers into an overarching, coherent representation

Incorporating acoustics, phonetics, phonology, prosody,visemes, lexemes, pragmatics, grammar and (ultimately) understanding



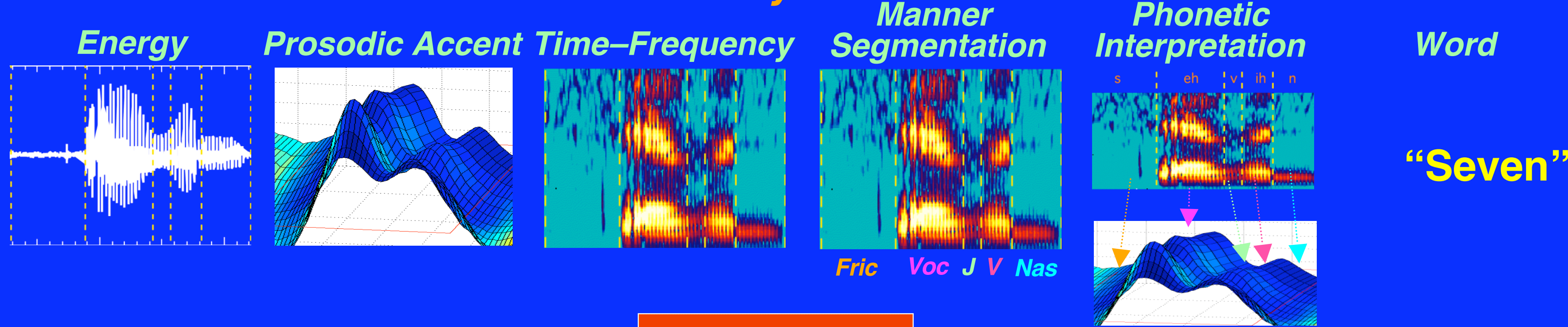
The Path to Utopia

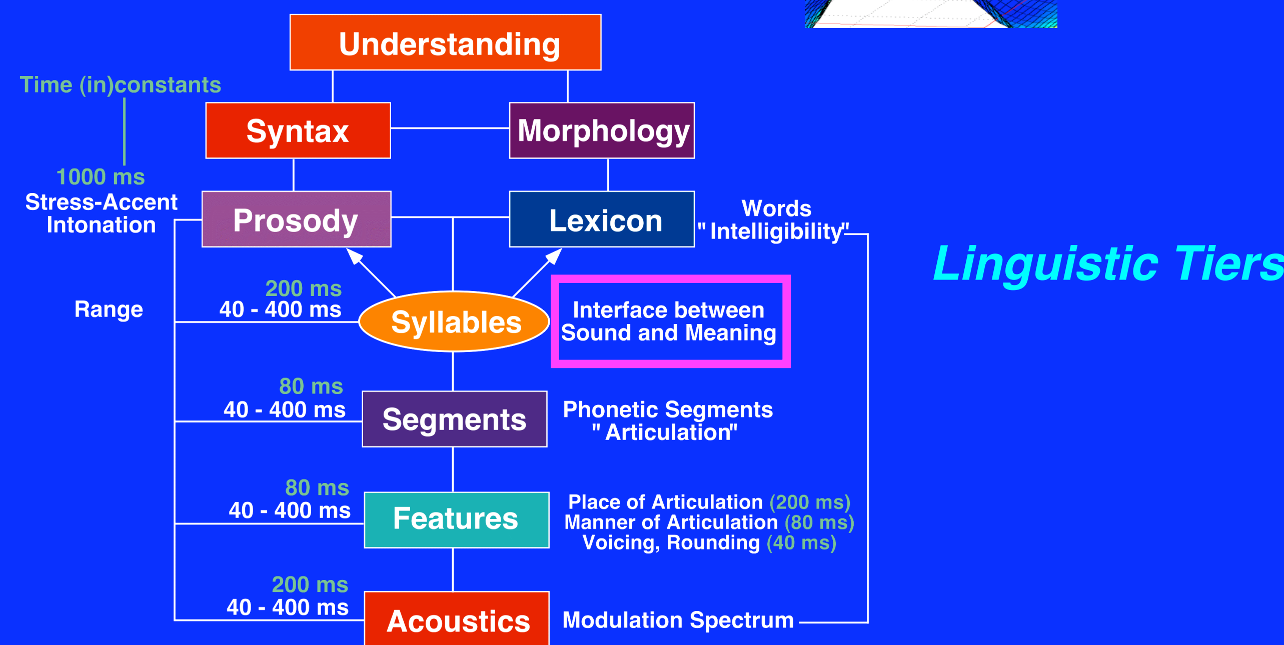
Where should we go?

Language – A Syllable-Centric Perspective

An empirically grounded perspective of spoken language focuses on the SYLLABLE and PROSODIC ACCENT as the interface between "sound" and "meaning" (or at least lexical form)

Modes of Analysis





The Importance

of

The Energy Arc

for

Understanding Spoken Language

The Energy Arc

Syllables are characterized by rises and falls in energy (see below, left)

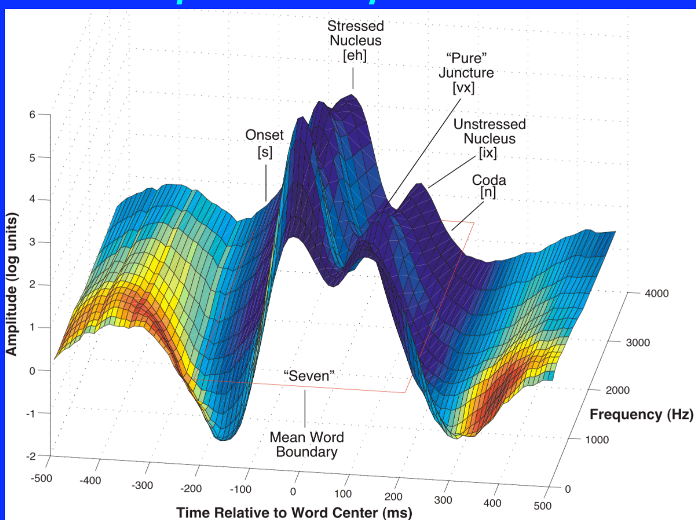
The "energy arc" can be considered to reflect both production and perception

From production's perspective, the arc reflects the articulatory cycle from closure to maximally open aperture and back again (in crude terms)

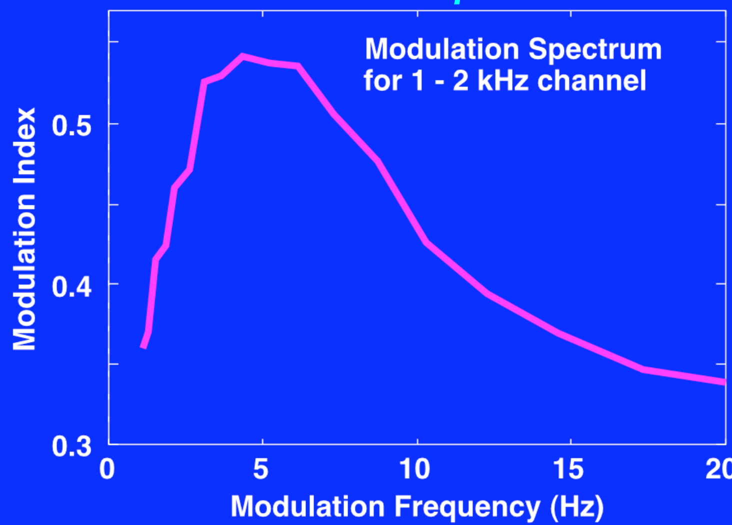
From the ear's perspective, the energy arc reflects the packaging of information within the temporal limits that the auditory system (and other sensory organs) has evolved to process

This temporal dimension is reflected in the modulation spectrum of spoken language (below, right)

Spectro-temporal Profile



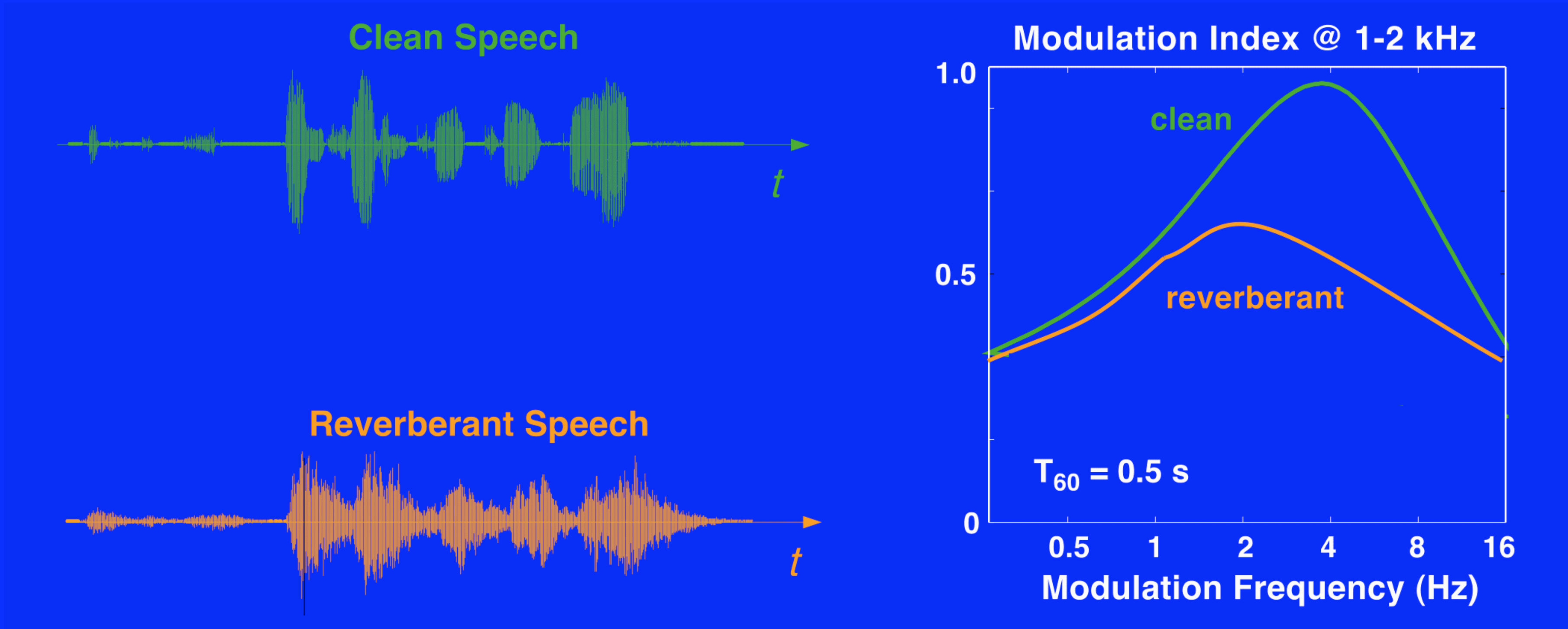
Modulation Spectrum



Importance of the Arc for Intelligibility

We know from perceptual studies that distortion of this energy arc (in the form of low-pass filtering the modulation spectrum or highly reverberated speech) destroys the intelligibility of speech

Preservation of syllable boundary information appears to be important for understanding spoken language



Based on a figure by Hynek Hermansky

The Arc's Relation to the Syllable

But what does the energy arc reflect linguistically?

And why is it so important for understanding speech?

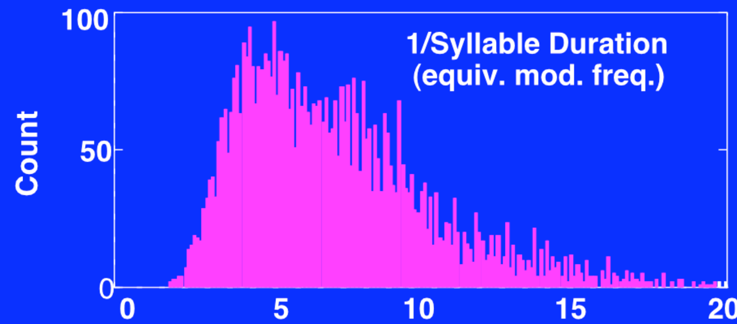
The concept of "sonority hierarchy" (Jespersen, 1899) is a (not very satisfactory) descriptive framework for specifying the order of segments within the syllable

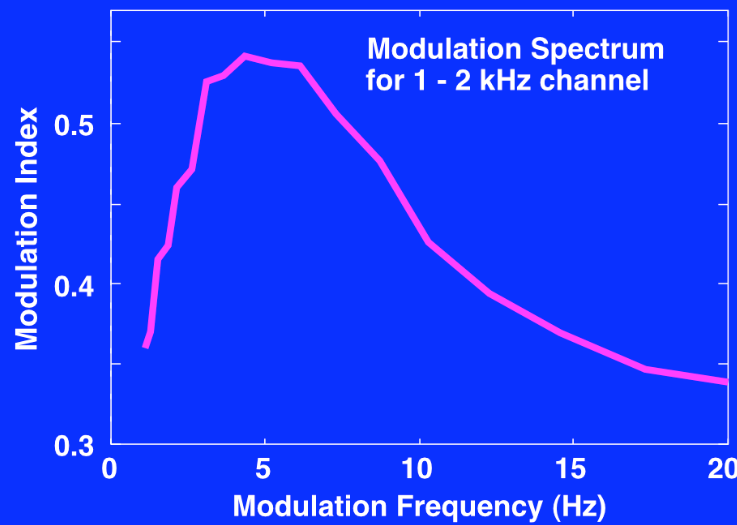
The energy arc provides a more principled (and accurate) framework for understanding why segments occur in the order they do within the syllable

Because the auditory system (and brain) requires that acoustic energy be packaged in oscillations of ca. 3 - 10 Hz, and because syllables are the linguistic manifestation of the modulation spectrum (see right)

The ARC essentially represents SYLLABLE structure

But HOW is this instantiated from the perspective of the vocal apparatus?





The Arc's Relation to Syllable Phonotactics

If we return to the basic question – WHY are syllables realized as rises and falls of energy ….

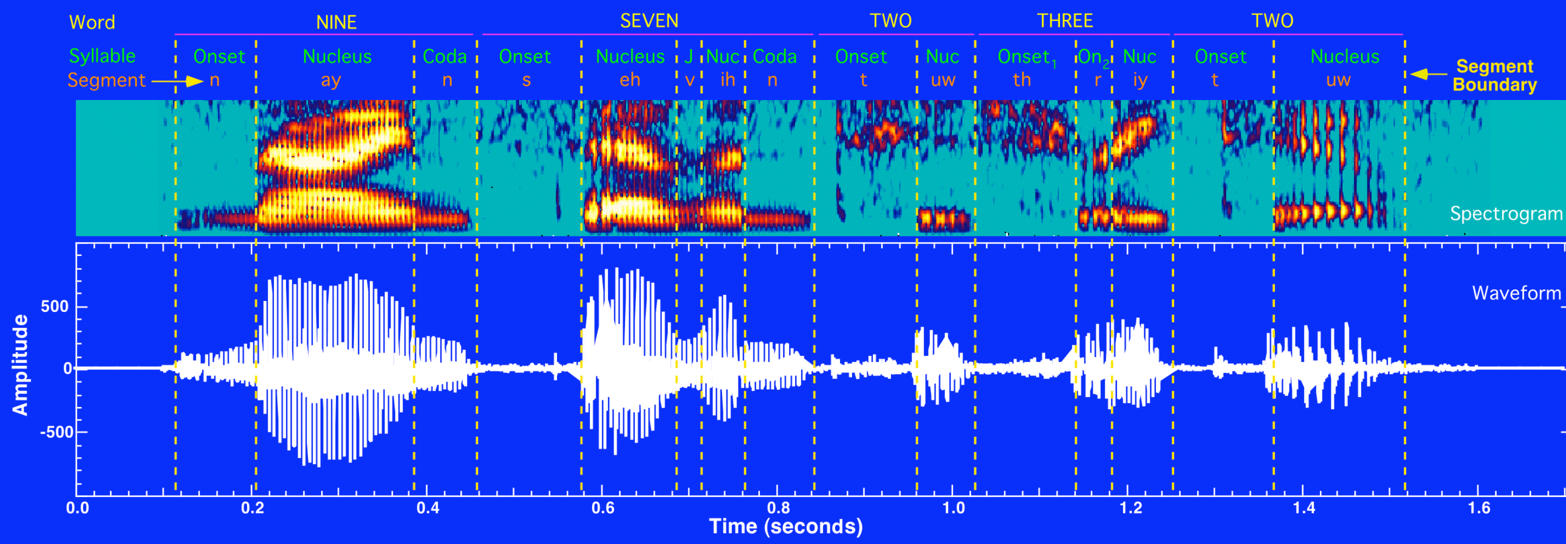
And we make the simple assumption that each manner of articulation – vowel, fricative, nasal, etc. – is associated with a specific energy level

Vowels being highest

Stops and fricatives lowest

With nasals, liquids and glides in between

Then we gain some insight as to why the segments occur in the order they do within the syllable



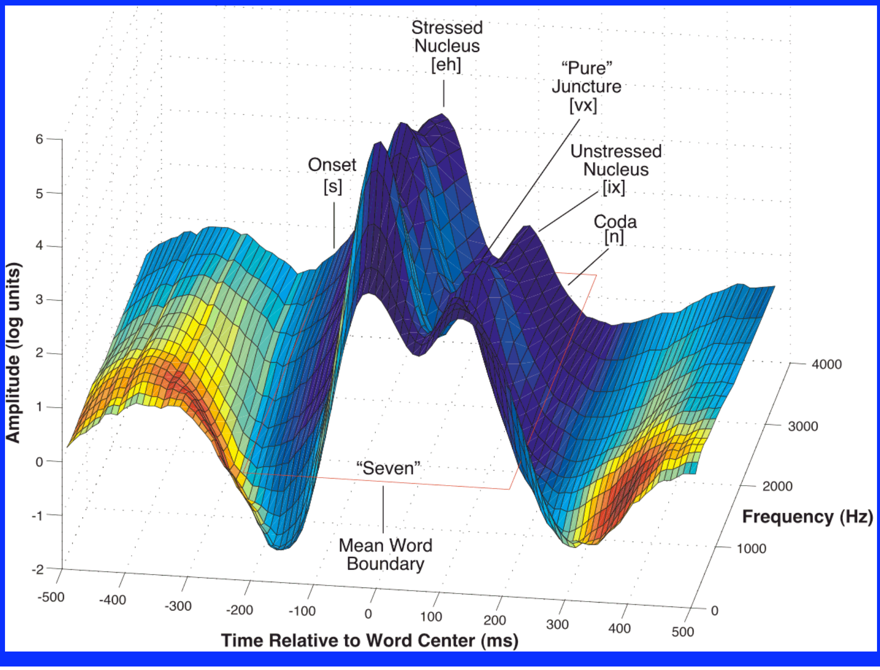
The Arc's Relation to Syllable Phonotactics

In effect, the segments reflect various manners of production, which are associated with different energy levels

From the perspective of "command and control" the relation between syllable production and the energy arc is automatic and unconscious

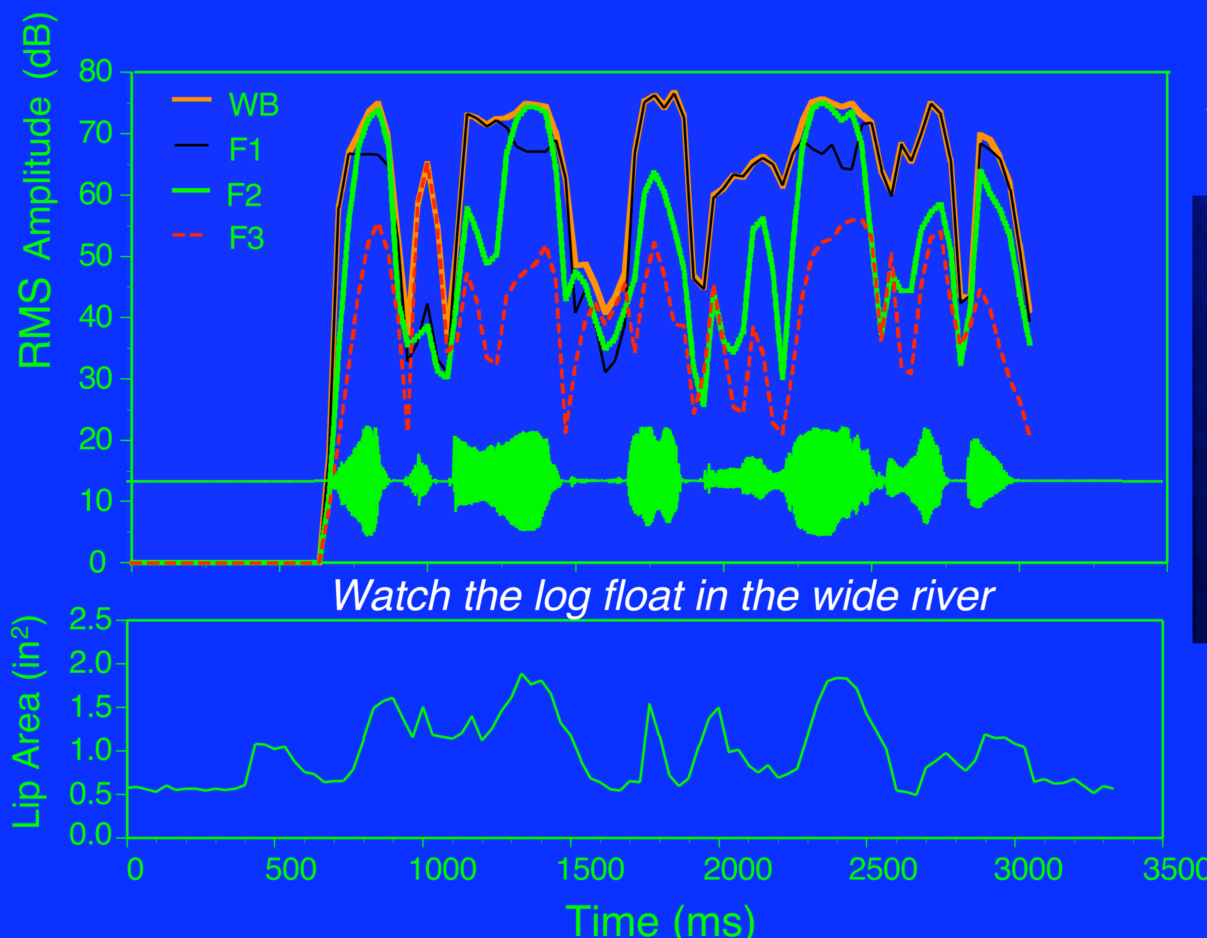
Syllables are intrinsically arcs that are readily digested by the auditory system and the brain

This may account for why it is possible to articulate (and perceive) in terms of syllables, but not in terms of isolated phones (unless they are syllables themselves)



The Arc's Relation to Visible Speech

- ***The energy arc's modulatory properties may also provide a basis for binding speechreading information with acoustic cues***
- ***This association with visual, speechreading cues could be important, as it provides up to a 10-dB SNR gain under noisy conditions***



Amplitude Fluctuation in Different Spectral Regions

Lip Aperture Variation

Data courtesy of Ken Grant

The Energy Arc and Voicing

Within the traditional framework, voicing is considered a segmental property

A segment is either voiced or not

However, we know that this segmental perspective on voicing is only a crude caricature of the acoustic properties of speech

Many theoretically voiced segments are at least partially unvoiced

For example, in Am. English it is common for [z] to be unvoiced – particularly in syllable-final position in unaccented syllables

The so-called voiced obstruents ([b], [d], [g]) are usually realized as partially unvoiced (this is what voice-onset-time refers to), with various languages differing with respect to the specific values of VOT

This sort of behavior implies that voicing is NOT a segmental feature, but rather one that is under SYLLABIC control and actually reflects prosodic factors (which is WHY languages vary with respect to VOT)

How can this be so?

The Syllabic Control of Voicing

- Recall, that the core of the syllable – the nucleus – is almost always voiced
- The nucleus is usually a vowel and contains the peak energy in the syllable
- Voicing spreads from the nucleus forward in time to the coda, as well as backward to the onset
- Voicing is continuous in time, and is associated with the higher-energy parts of the syllable
- The lower-energy components of the syllable may or may not be voiced
- But where the signal is unvoiced, the associated constituents reside in the "tails" of the syllable – the onset and/or coda
- It is probably not a coincidence that the most linguistically informative components in speech are NOT associated with voicing

voiced voiced voi voiced

The Syllabic Control of Voicing – Significance

The most energetic components of the speech signal are usually voiced

Voicing helps to build up energy in the syllable

Voicing provides implicit structure for the syllable

This structure could be extremely important in decoding the speech signal in noisy environments

Recall the importance of fundamental-frequency information for separating concurrent talkers or distinguishing speech from a noisy background

Pitch-related cues could only play such an important role if the speech signal is largely voiced

voiced ***voiced*** ***voi*** ***voiced***

The Relation Between Voicing and Manner

- Thus, voicing appears to cut across segmental boundaries
- It only APPEARS to be associated with individual segments
- Voicing serves to bind the segments into a syllabic whole through its temporal continuity
- It is probably not coincidental that 80% (or more) of the speech signal is voiced
- And that relatively few manner classes (usually stops, afficates, fricatives) can be realized as unvoiced (except in whispered or exaggerated speech)
- Voicing is indirectly related to the energy arc, in that it is associated with the most intense components of the syllable and is most robust to noise and reverberation
- Thus, it is extremely important for decoding speech in noisy environments

The Relation Between

The Energy Arc

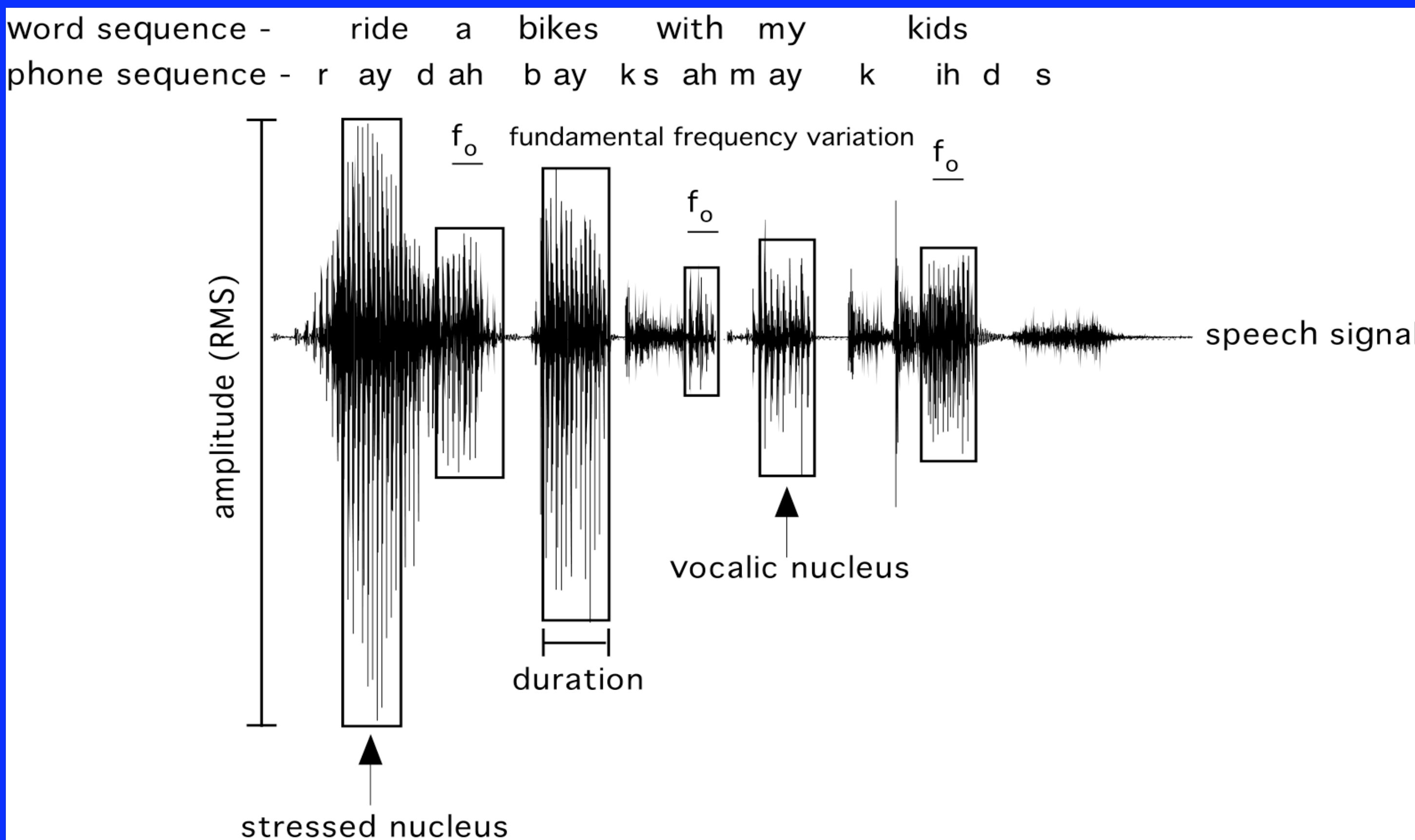
and

Prosody

Prosody is Related to the Energy Arc

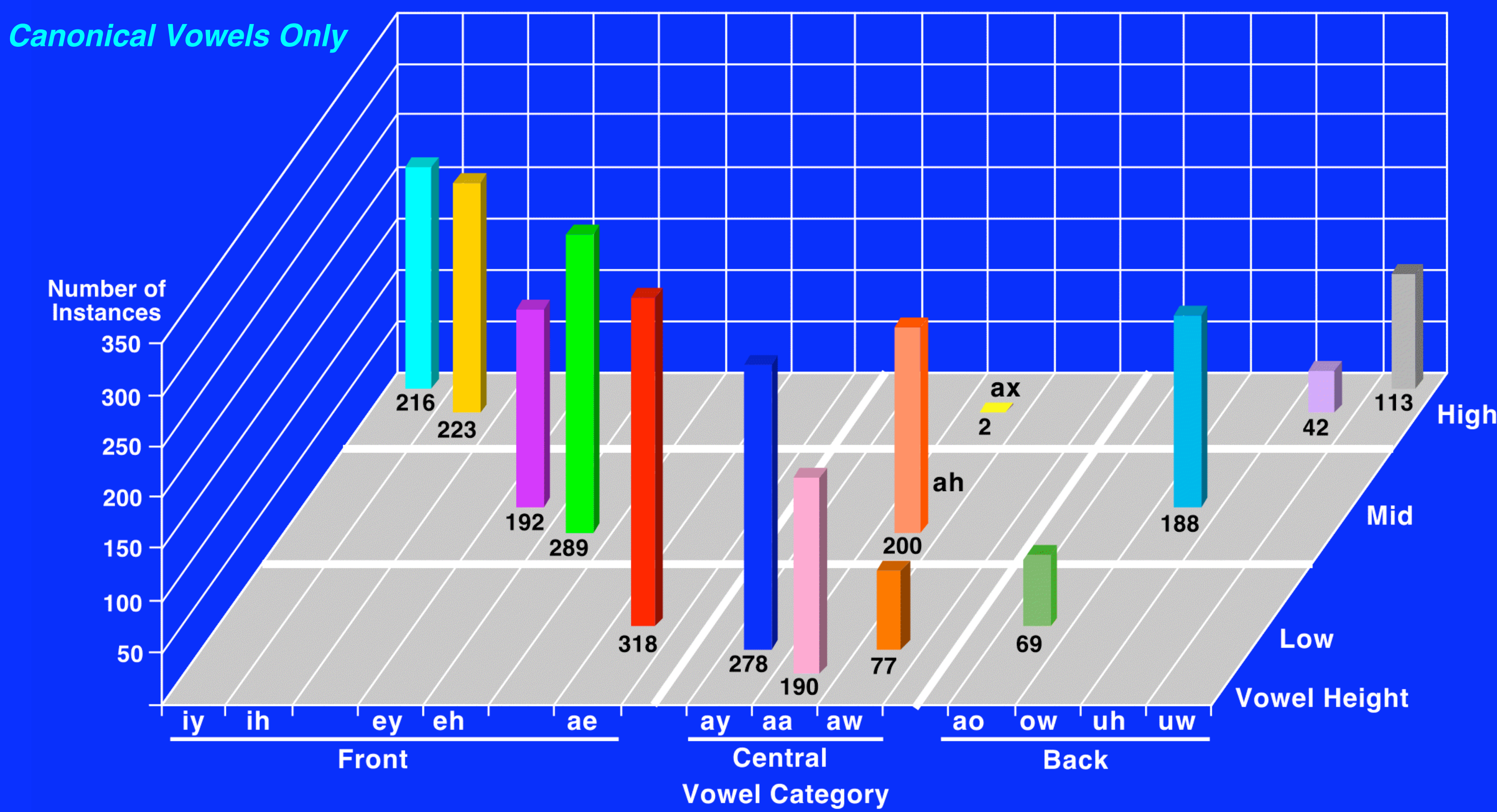
Utterances are composed of syllables of variable prominence

The vowels in the heavily accented syllables tend to differ from those in unaccented syllables



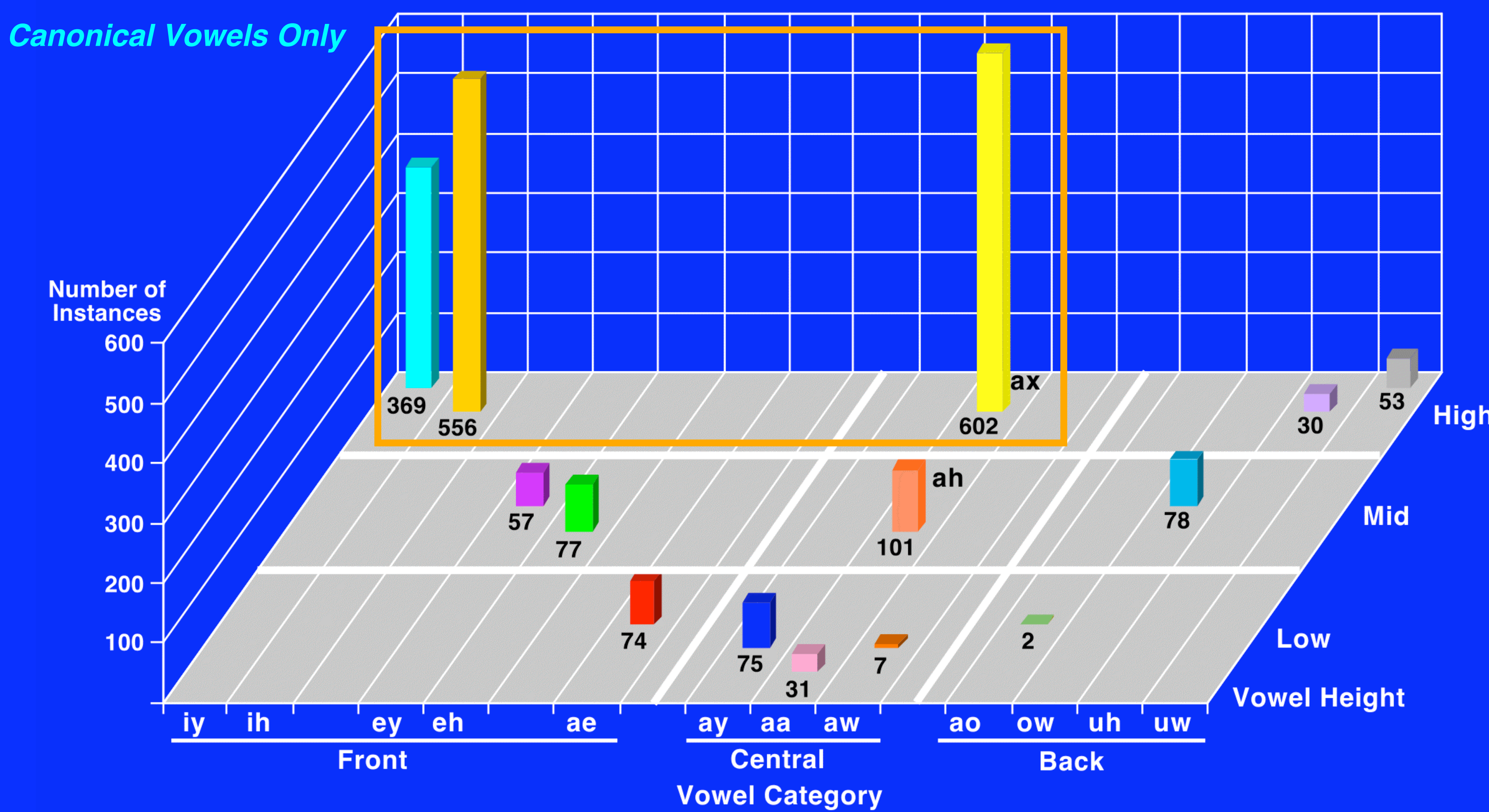
The Vowel System Under (Full) Stress (Accent)

In HEAVILY ACCENTED nuclei there is a relatively even distribution of segments across the vowel space, with a slight bias towards the front and central vowels



The Vowel System Without (Stress) Accent

In UNACCENTED syllables vowels are confined largely to the high-front and high-central sectors of the articulatory space

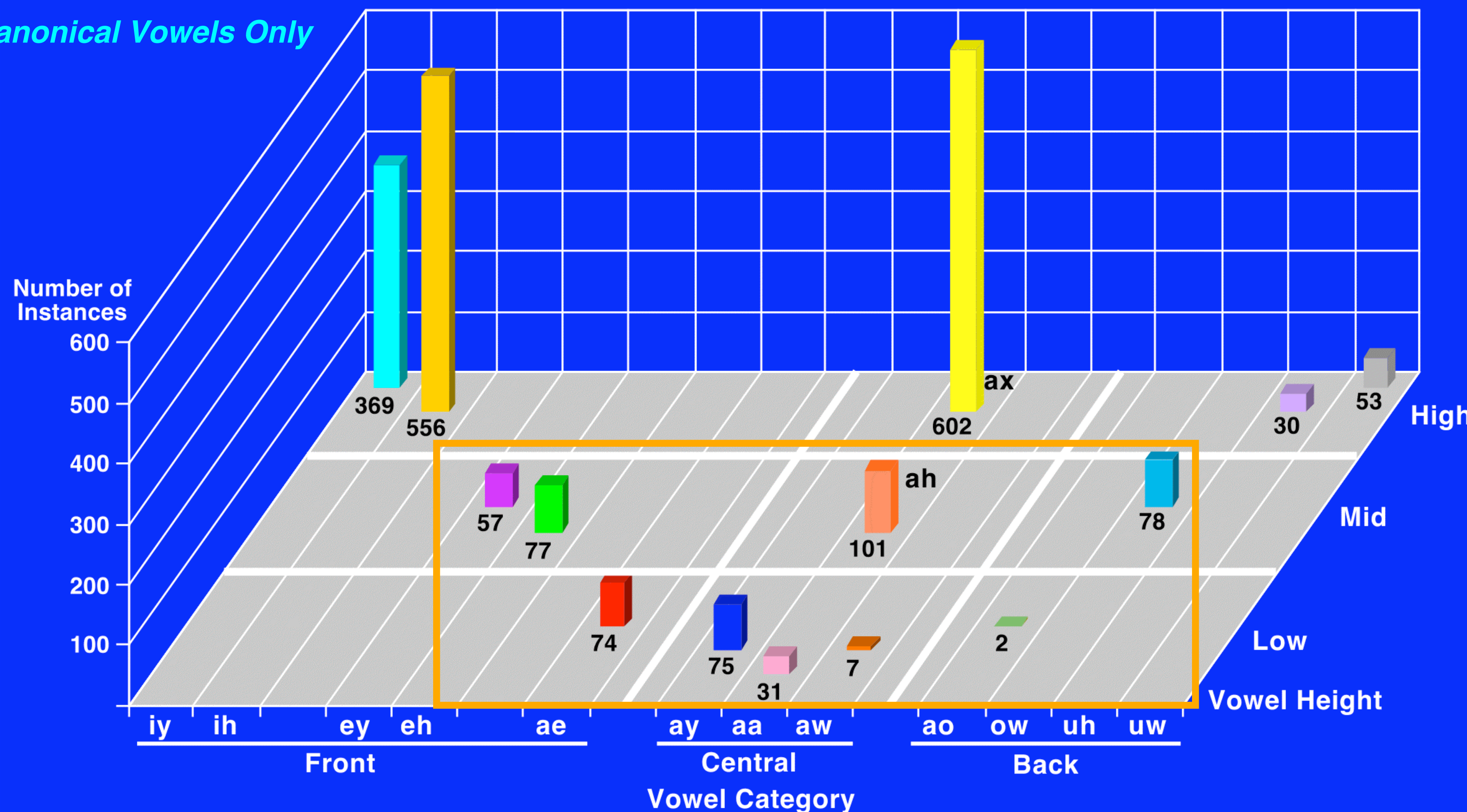


The Vowel System Without (Stress) Accent

In unaccented syllables vowels are confined largely to the high-front and high-central sectors of the articulatory space

The low and mid vowels "get creamed"

Canonical Vowels Only



The Vowel Systems Compared

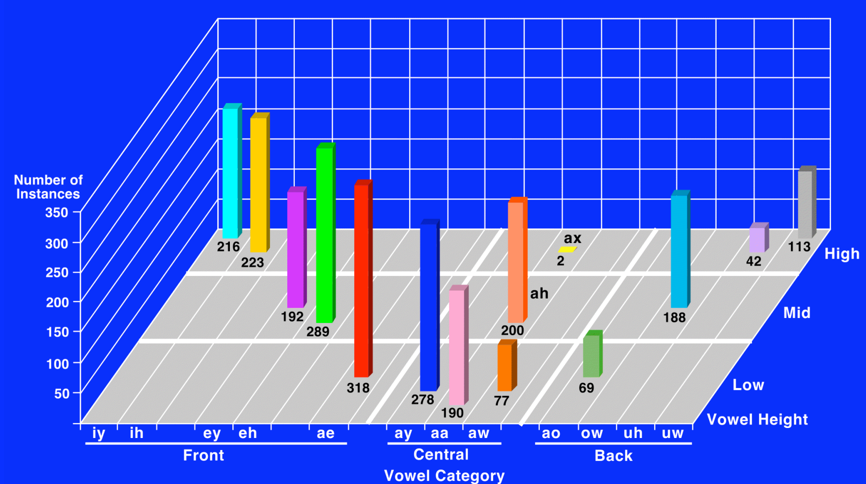
Stress accent exerts a profound effect on the character of the vowel space

High vowels are largely associated with unaccented syllables

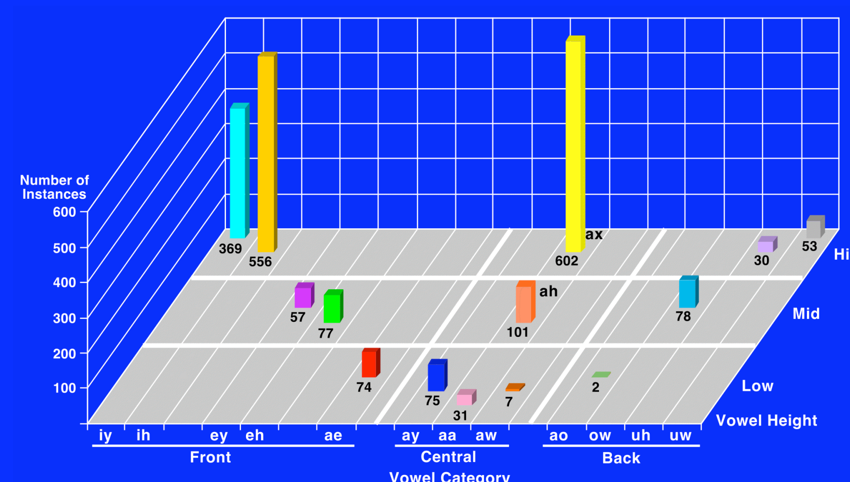
Low vowels are mostly found in accented syllables

This distinction between accented and unaccented syllables is of profound importance for understanding (and modeling) pronunciation variation

Heavily Accented



Unaccented

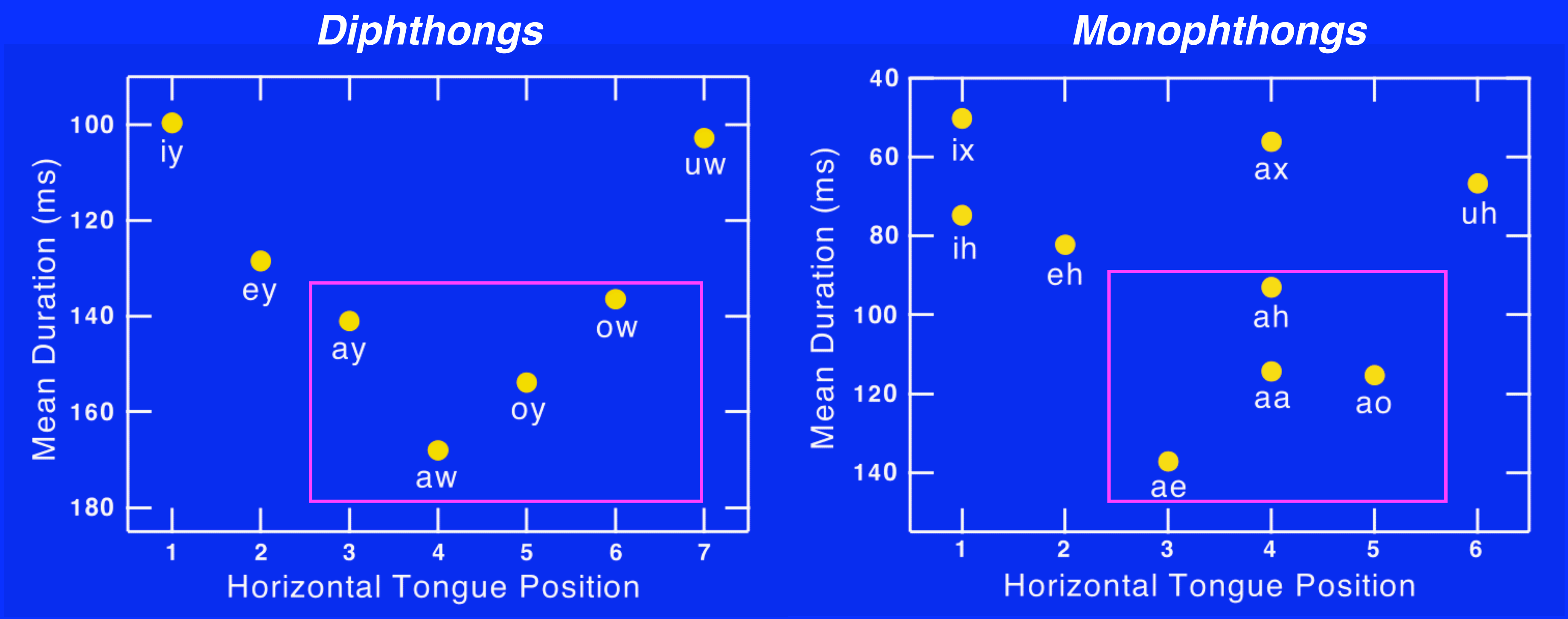


Canonical Vowels Only

Vowels as Carriers of Prosody

Vowels are an intricate component of the prosodic system

It is not coincidental that "low" vowels tend to be longer in duration and are more intense than "high" vowels



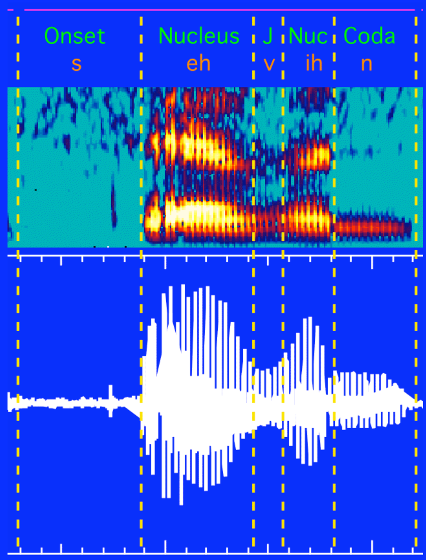
Vowels as Carriers of Prosody

Important words (and syllables) in an utterance tend to contain "low" and "mid" vowels

Frequent (function) words tend to contain "high" vowels

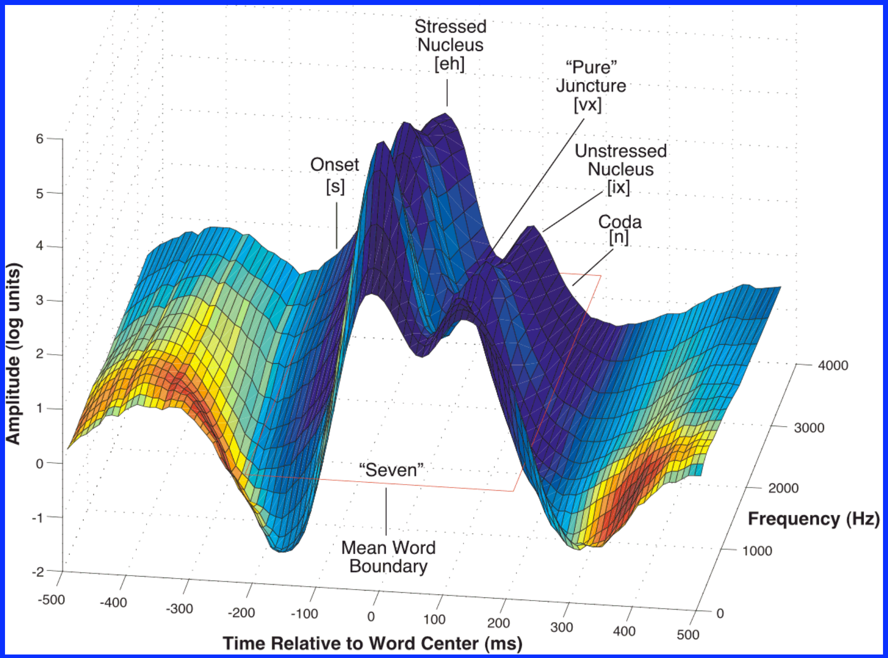
Thus, vowels (and hence syllables) with more energy and longer duration tend to carry more information than their shorter, less intense counterparts

Spectrogram + Waveform



"seven"

Spectro-temporal profile (STeP)



The Micro-Structure

of

The Syllable

(and why it matters)

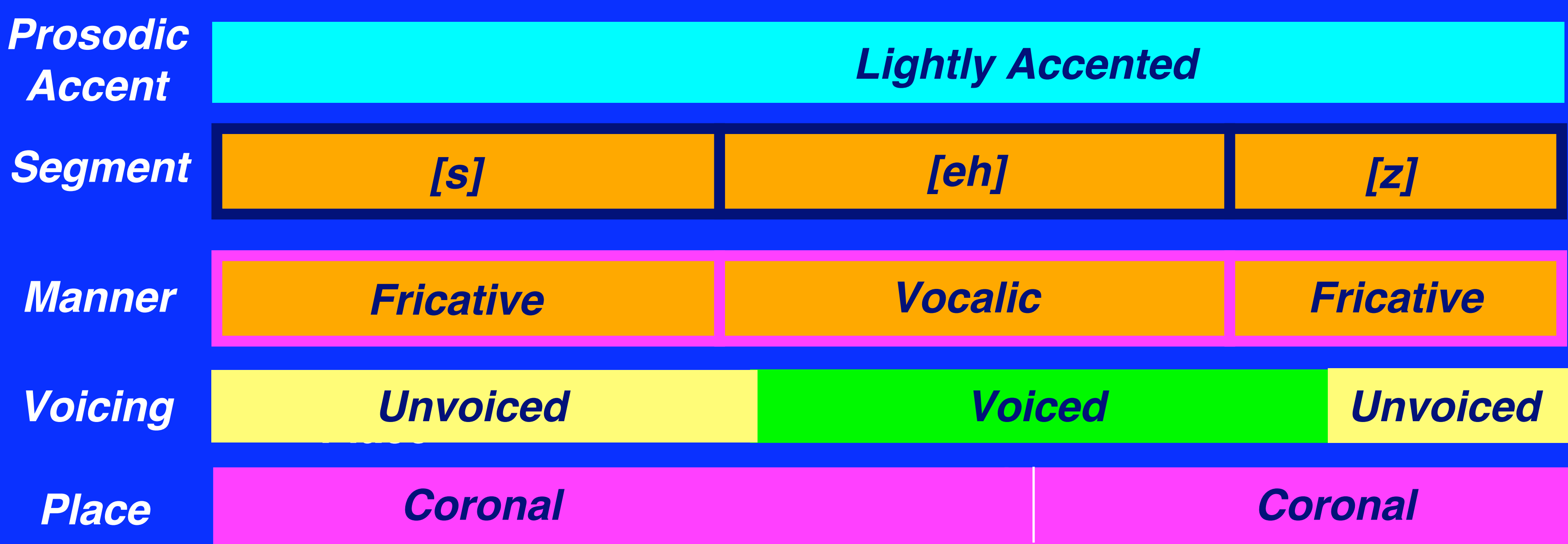
Micro-Structure of the Syllable

We now delve into the syllable's micro-structure to delineate the interaction among the phone(me), articulatory features, prosody and lexical identity

Three principal articulatory dimensions are distinguished (among others) – VOICING, MANNER and PLACE of articulation

Each articulatory dimension plays a specific functional role and is associated with a different time constant

Each dimension is sensitive to prosody, but in different ways



Lexical Structure

There are certain patterns to the phonetic-prosodic properties of words in terms of:

Voicing

Order of manner of articulation within the syllable

Articulatory place

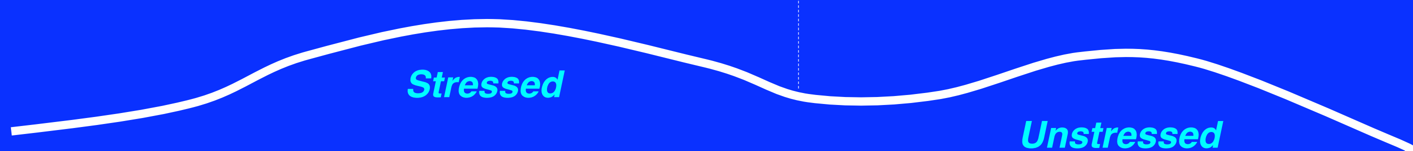
Energy contour

And so on …. (let's focus on place of articulation for the moment)

WORD – "Strengthen"

	SYLLABLE – "streng"					SYLLABLE – "then"		
	ONSET			NUCLEUS	CODA	ONSET	NUCLEUS	CODA
Segment	s	t	r	ε	ŋ	θ	ɪ	n
Manner	Fric	Stop	Rhotic	Vowel	Stop	Fric	Vowel	Nasal
Place	ø	Central	ø	Front	Back	Central	Front	Central
Height	ø	ø	ø	Mid	ø	ø	High	ø
Voicing	–	–	+	+	+	–	+	+
Duration	170 (ms)			80	60	60	30	50

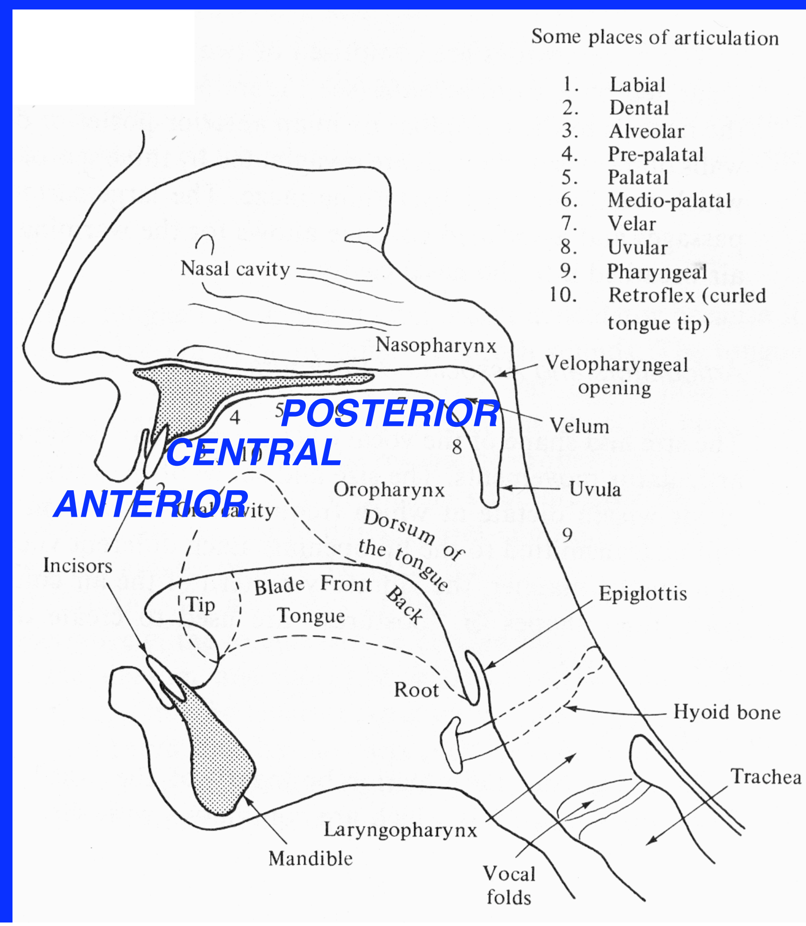
Energy Contour



Place of Articulation

- ***Articulatory place information is important for distinguishing among syllables and words (particularly for consonants)***
- ***The distinction among [b], [d] and [g], and [p], [t] and [k] is primarily one of "place," in that the location of maximum articulatory constriction varies from front to back***

- ***Generally, there are only three distinct loci of constriction for any single manner class***
- ***Hence, the problem of determining articulatory place is greatly simplified if the manner of production is known***
- ***Manner-dependent place of articulation classifiers have been successfully applied in automatic phonetic transcription***

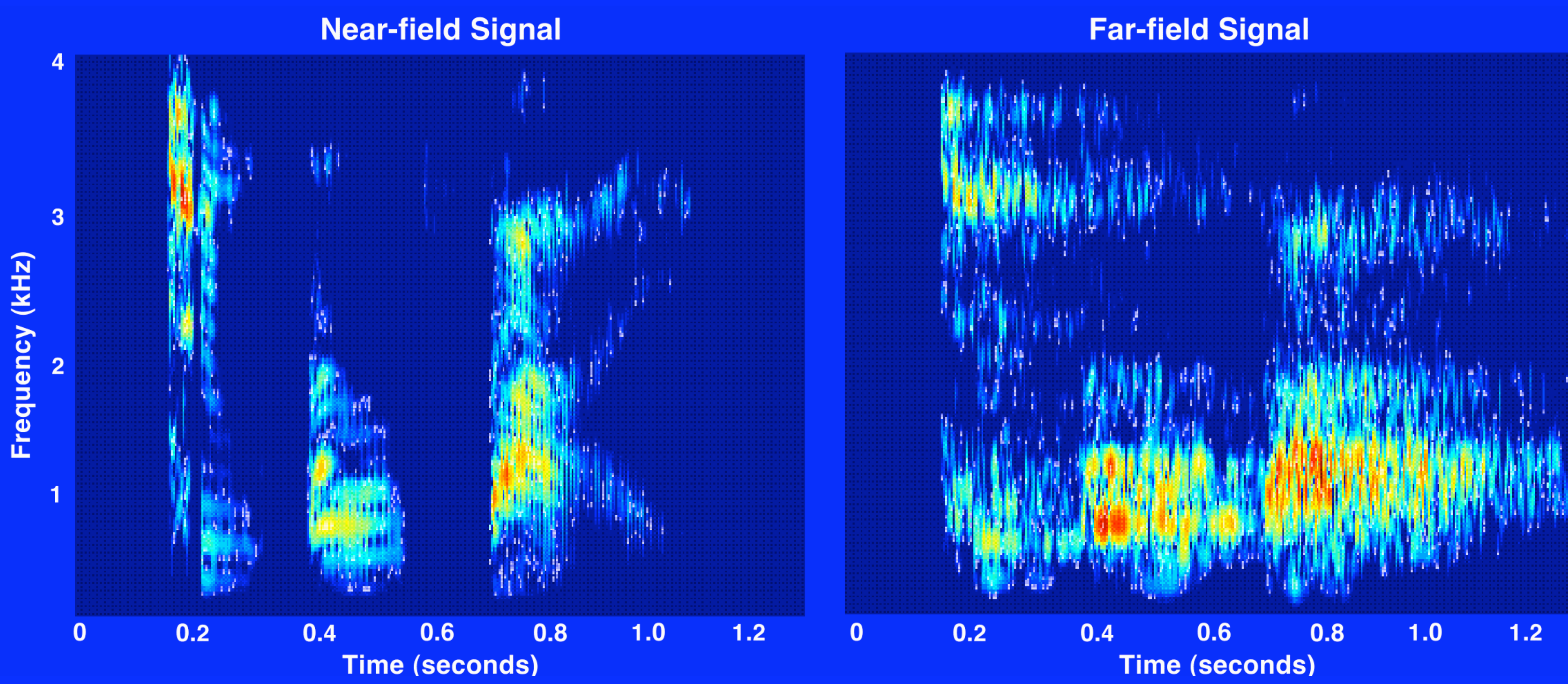


Place of Articulation

The formant patterns associated with place of articulation cues vary broadly over frequency and time

When speech is described as "dynamic" it is usually such formant patterns that are meant (this is a little misleading, in that syllable cues are also highly dynamic, but this is a separate story ….)

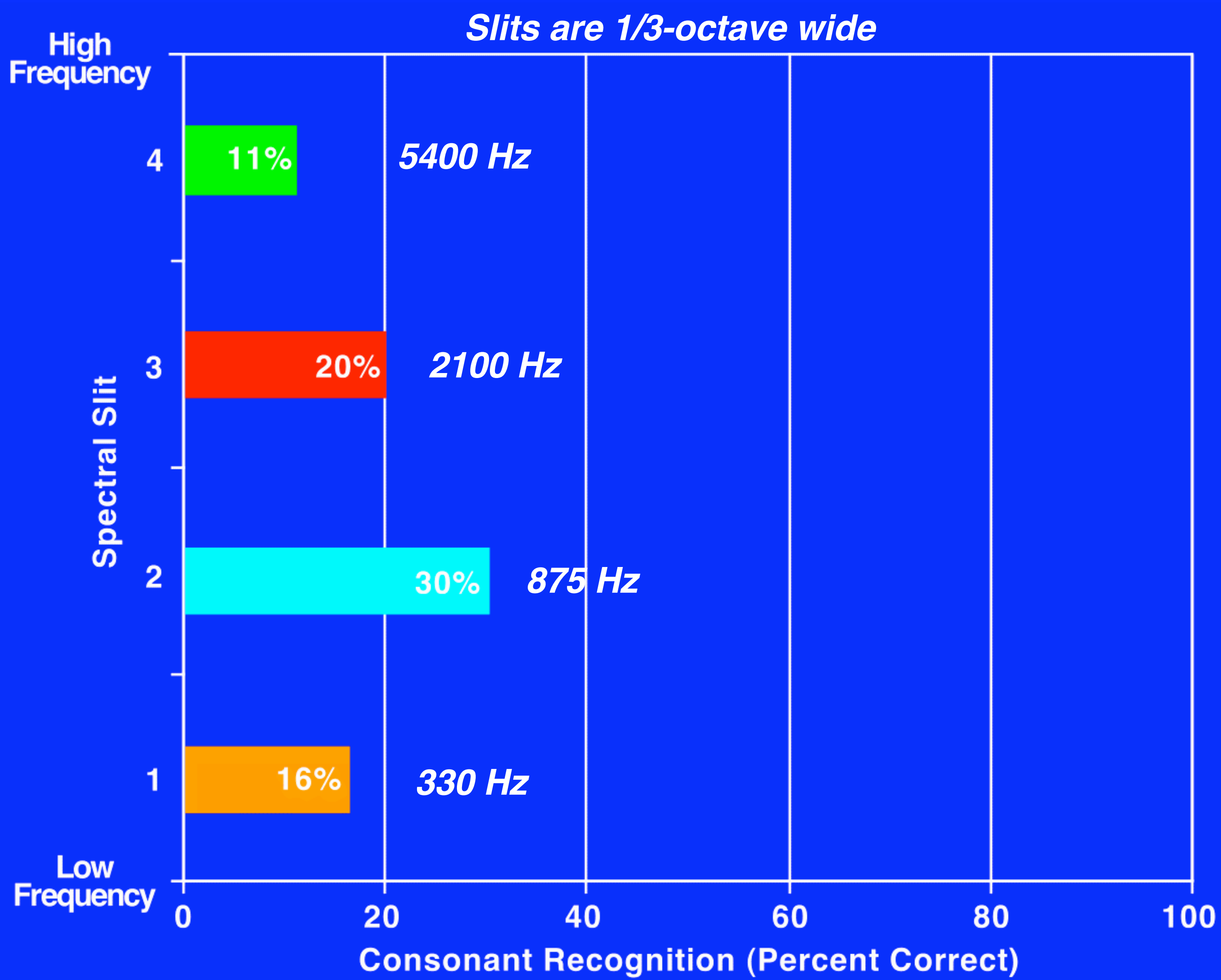
In low signal-to-noise ratio conditions and among the hearing impaired, place-of-articulation cues are usually among the first to degrade



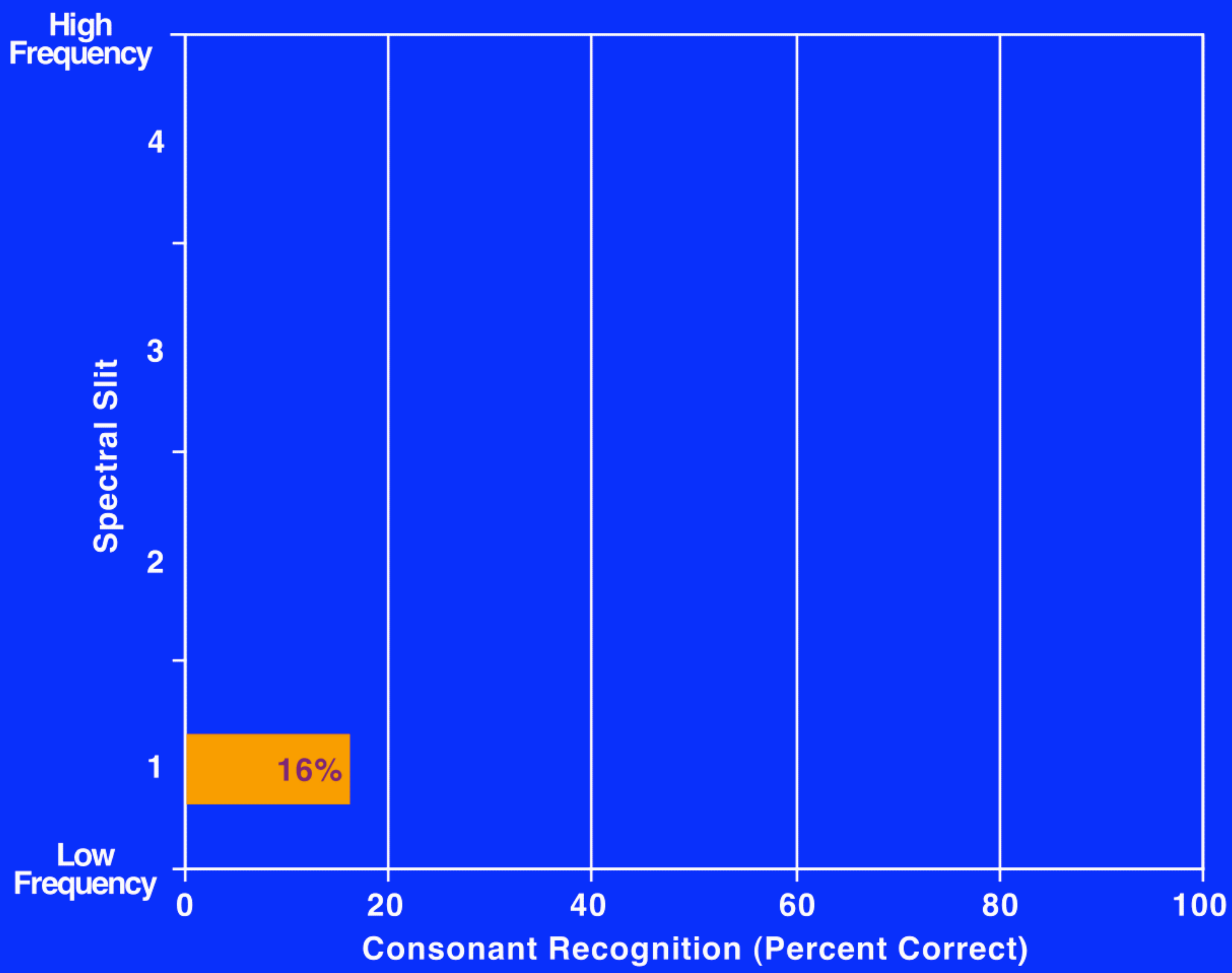
Place of Articulation

- ***The reasons for this seeming vulnerability are controversial, but can be understood through analysis of data shown on the following slides***
- ***In this experiment, nonsense VC and CV syllables were presented to listeners, who were asked to identify the consonant***
- ***The syllables were spectrally filtered, so that most of the spectrum was discarded***
- ***The proportion of consonants correctly recognized was scored as a function of the number of spectral slits presented and their frequency location, as shown on the next series of slides***
- ***The really interesting analysis comes afterwards ….***

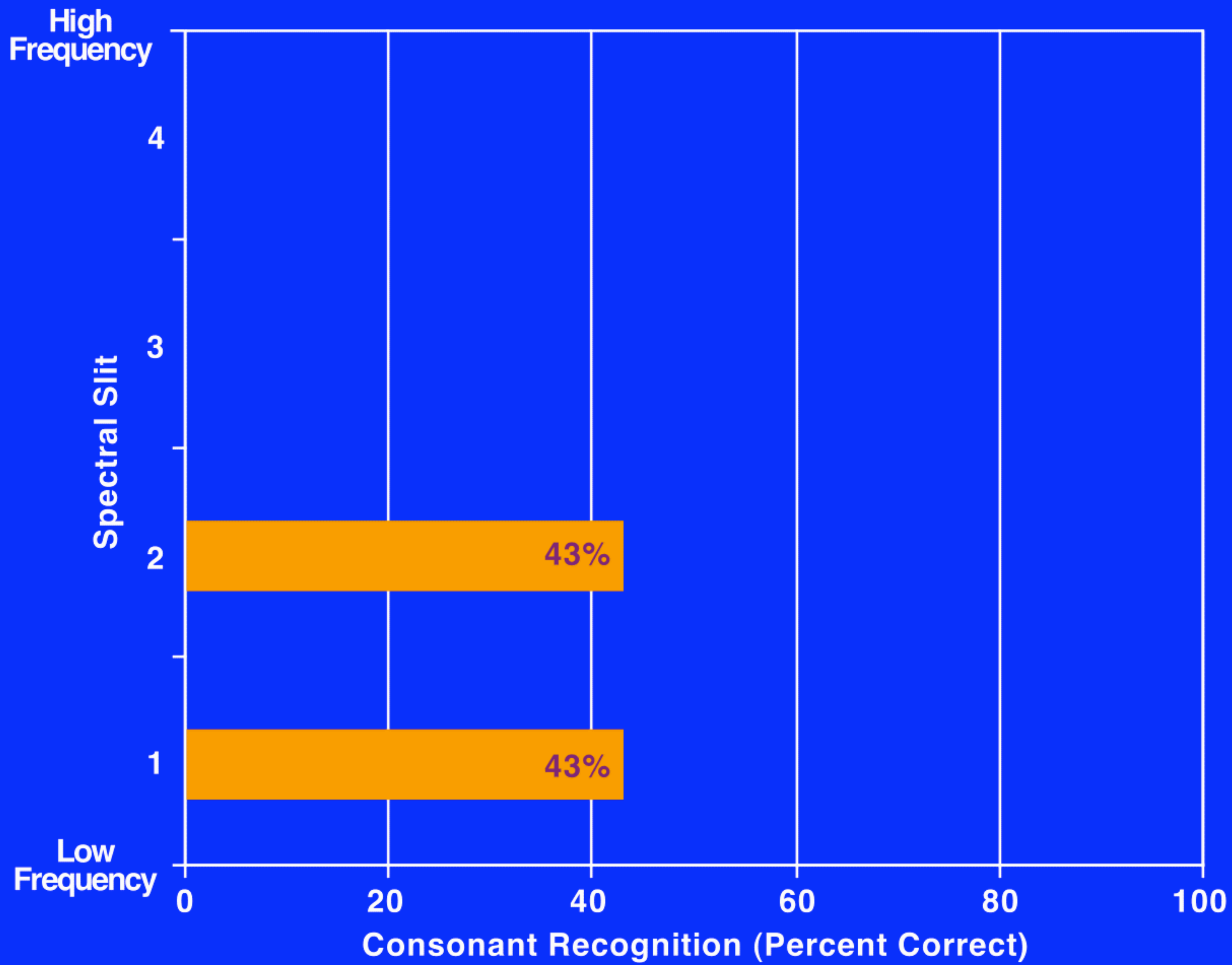
Consonant Recognition - Single Slits



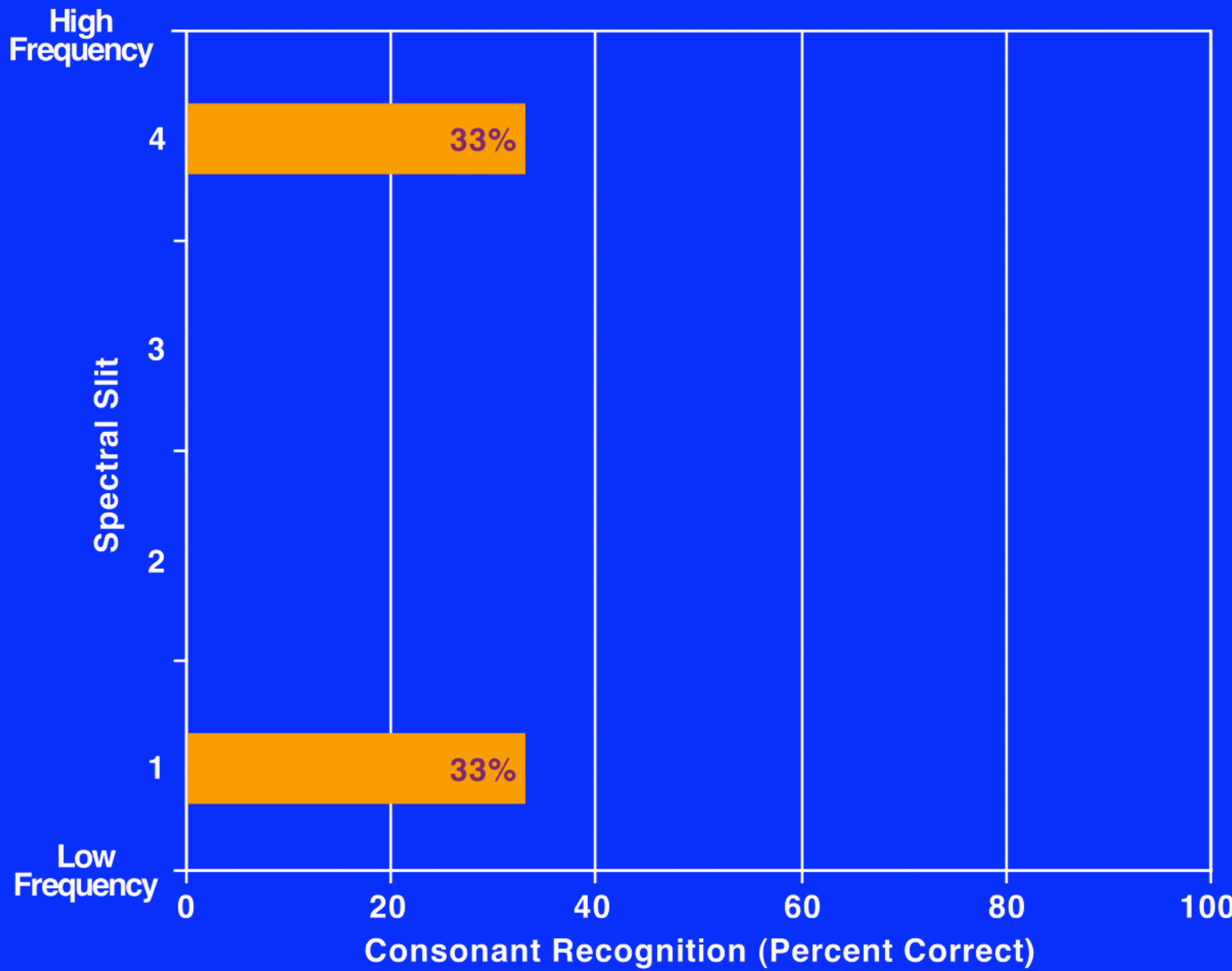
Consonant Recognition - 1 Slit



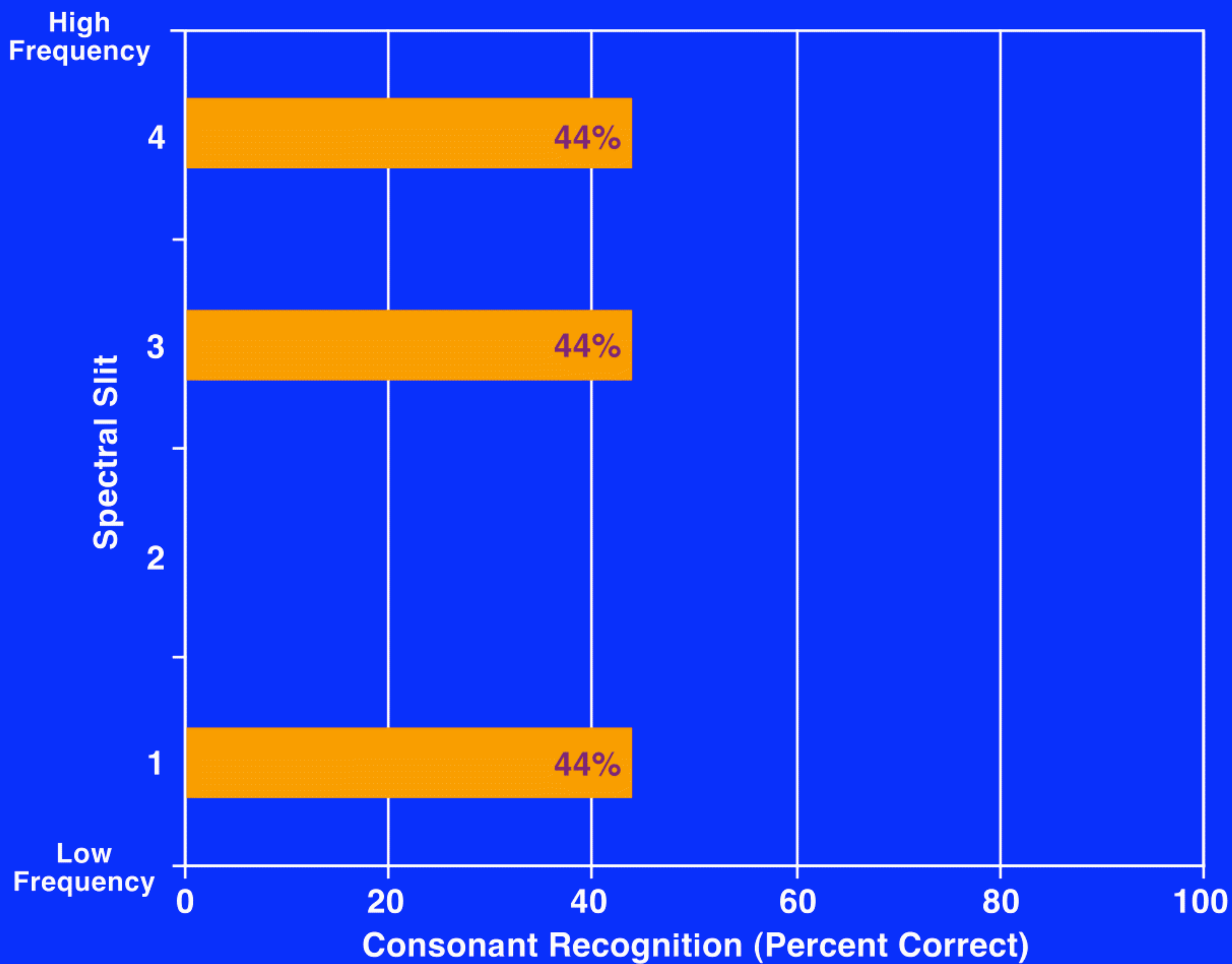
Consonant Recognition - 2 Slits



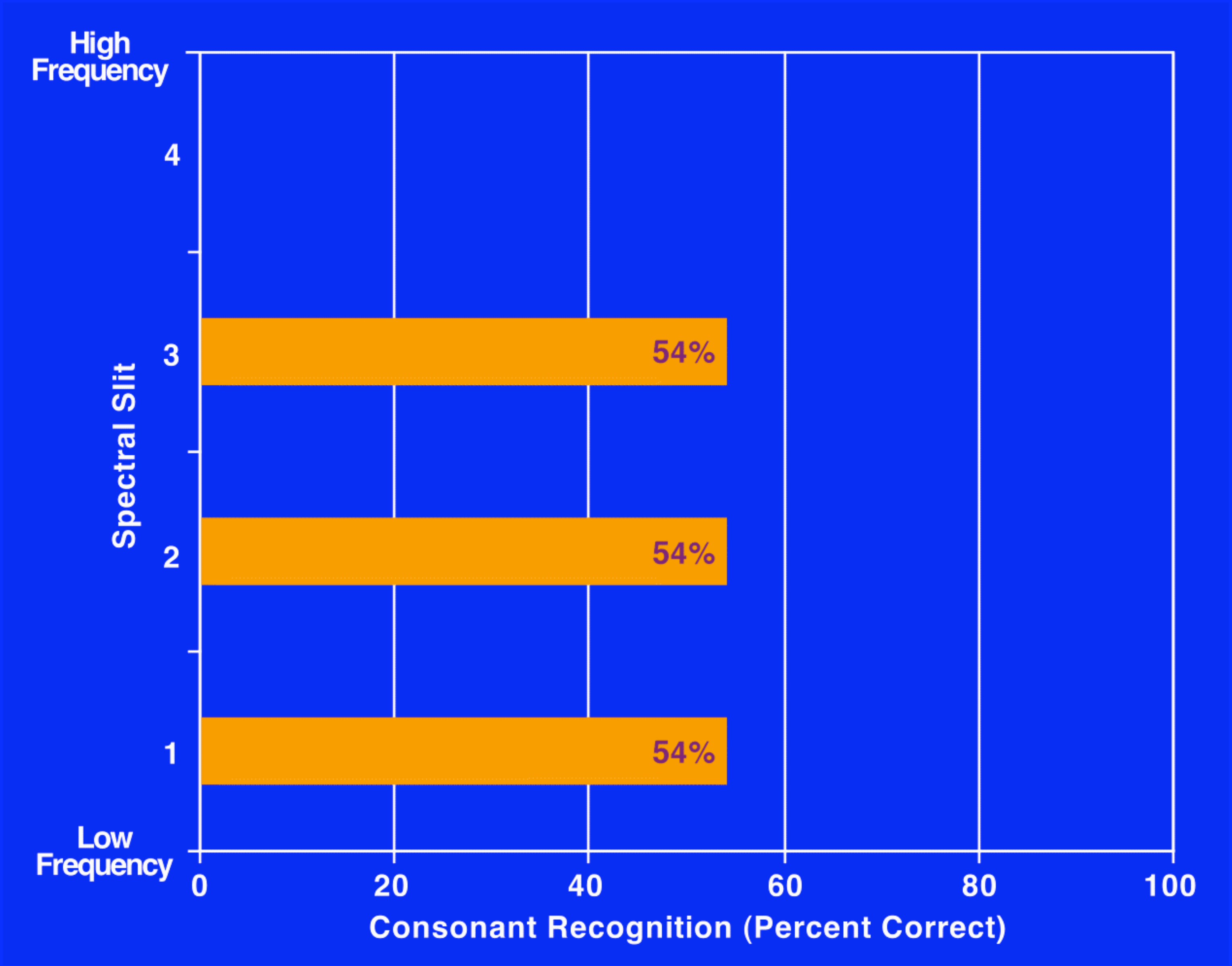
Consonant Recognition - 2 Slits



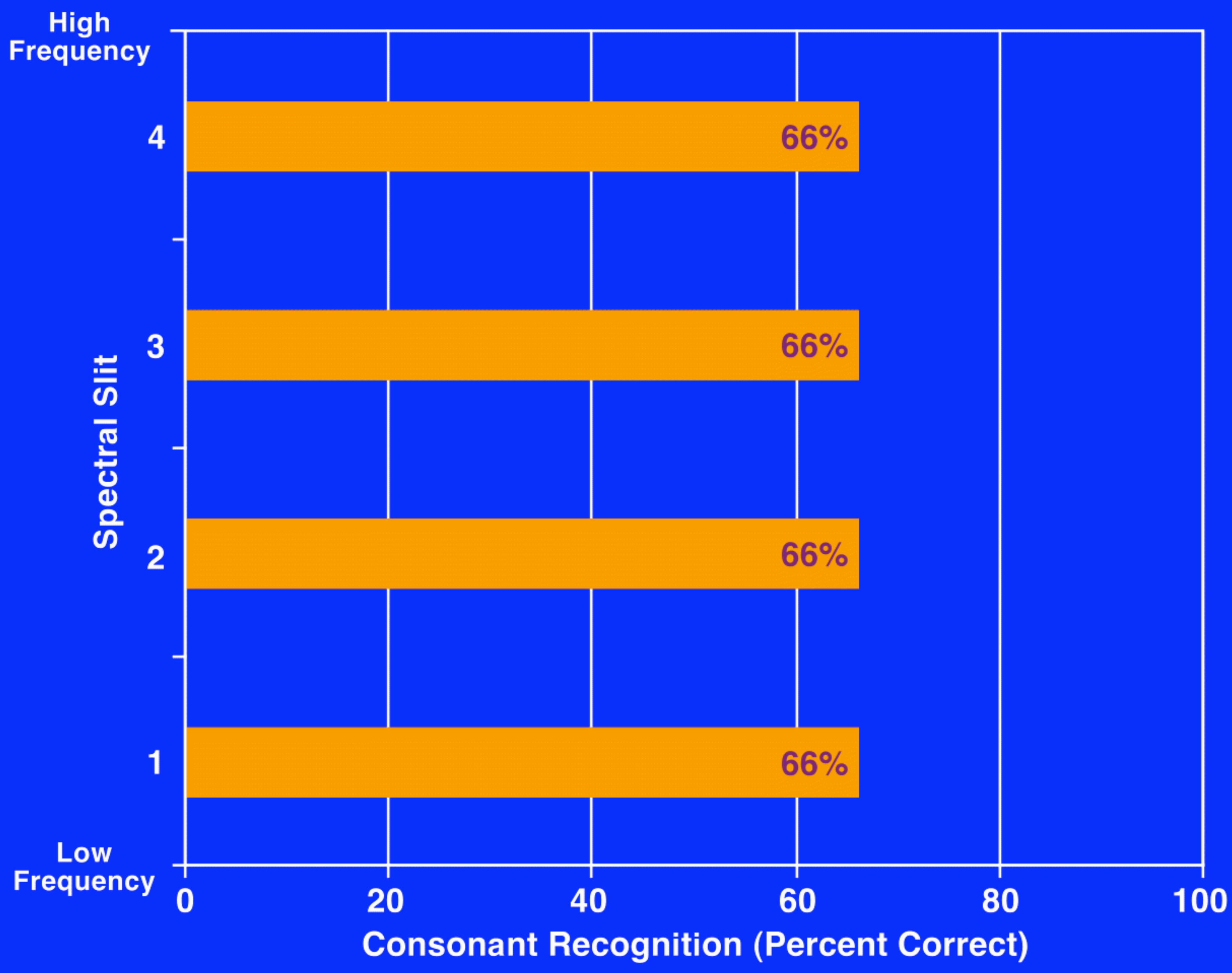
Consonant Recognition - 3 Slits



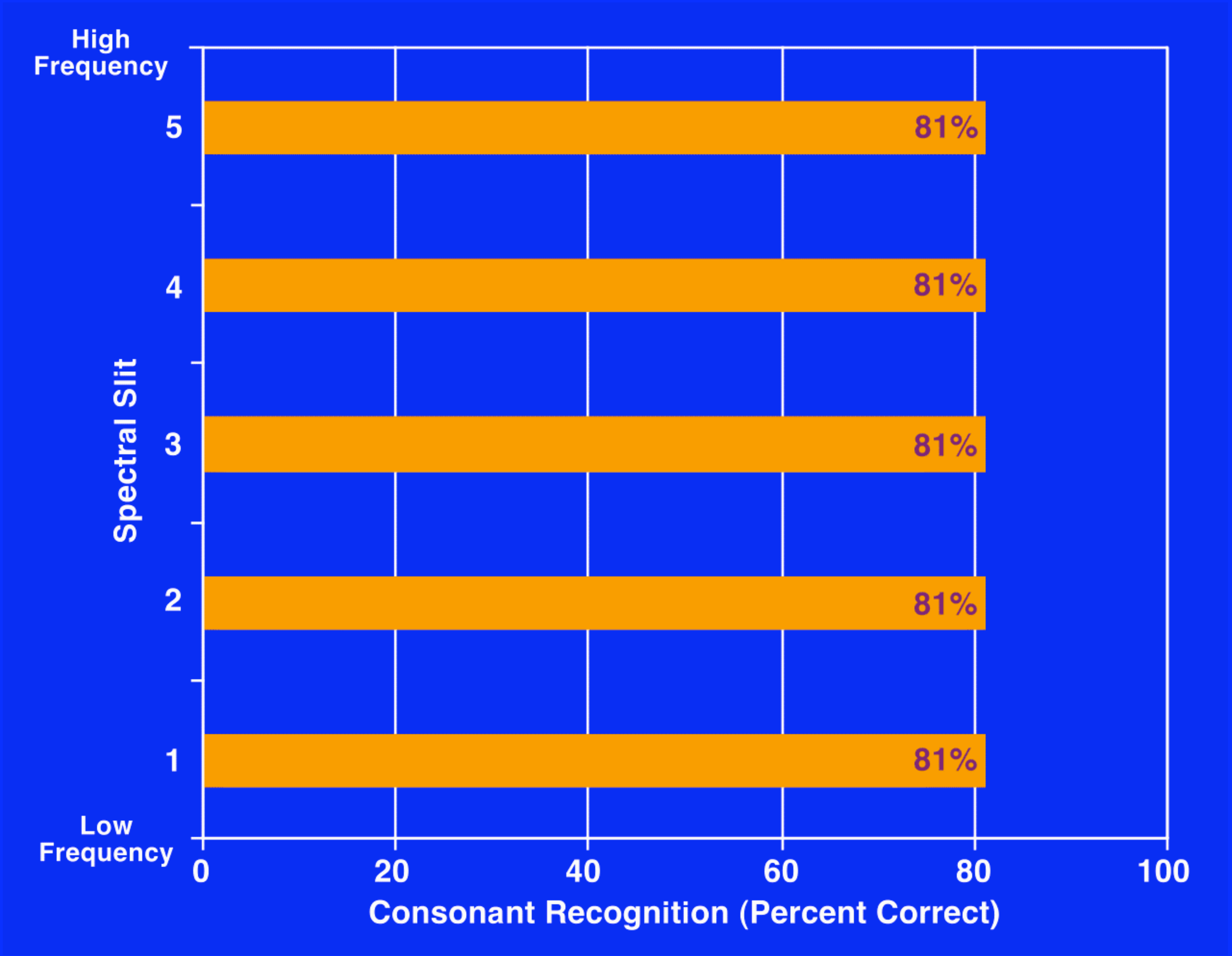
Consonant Recognition - 3 Slits



Consonant Recognition - 4 Slits



Consonant Recognition - 5 Slits



Articulatory-Feature Analysis

The results, as scored in terms of raw consonant identification accuracy, are not particularly insightful (or interesting) in and of themselves

They show that the broader the spectral bandwidth of the slits, the more accurate is consonant recognition

Moreover, a more densely sampled spectrum results in higher recognition

However, we can perform a more detailed analysis by examining the pattern of errors made by listeners

From the confusion matrices we can ascertain precisely WHICH ARTICULATORY FEATURES are affected by the various manipulations imposed

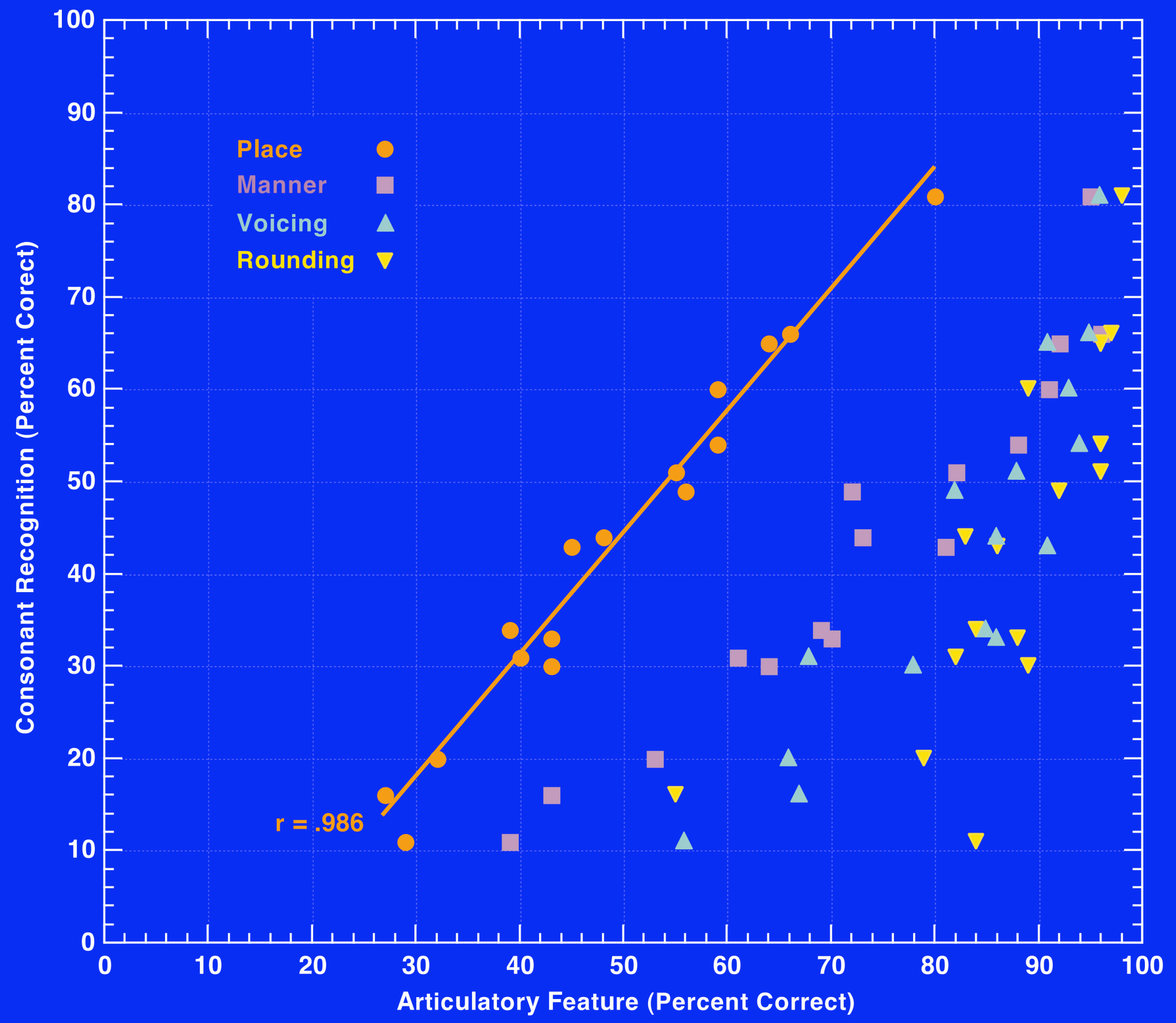
And from this error analysis we can make certain deductions about the distribution of phonetic information across the tonotopic frequency axis potentially relevant to understanding why speech is most effectively communicated via a broad spectral carrier

Correlation - AFs/Consonant Recognition

Consonant recognition is almost perfectly correlated with place-of-articulation performance

This correlation suggests that PLACE features are based on cues distributed across the entire speech spectrum, in contrast to features such as voicing and rounding, which appear to be extracted from a narrower span of the spectrum

MANNER is also highly correlated with consonant recognition, implying that such features are extracted from a fairly broad portion of the spectrum as well



Importance of Place Cues for Speechreading

The significance of these results is apparent when we consider cross-modal integration of speech information

Speechreading cues can provide extremely important information for understanding spoken language in noisy and reverberant conditions, as well as for the hearing impaired and non-native speakers of a language

It is estimated that 94% of the information provided by the visible articulators pertains to PLACE of articulation

PLACE cues are broadly distributed across the spectrum, with particular emphasis above 800 Hz, consistent with speechreading studies

Place information also appears to be crucial for lexical discrimination

And visual cues can play a crucial role in place decoding, particularly in noise and among the hearing impaired

Speech is likely to have evolved in face-to-face settings where the visual cues render place information inherently robust

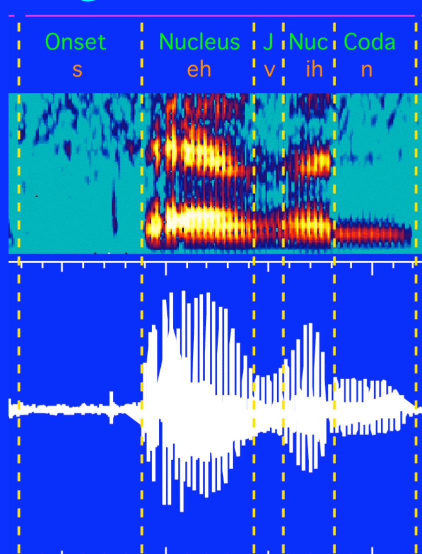
Time Course of Place Cues

Place of articulation is an inherently trans-segmental feature that effectively binds the syllabic nucleus with either preceding or following consonant(s)

The cues for place are distributed across segmental boundaries, even though there are cues within the segment that can be used to identify place under certain conditions

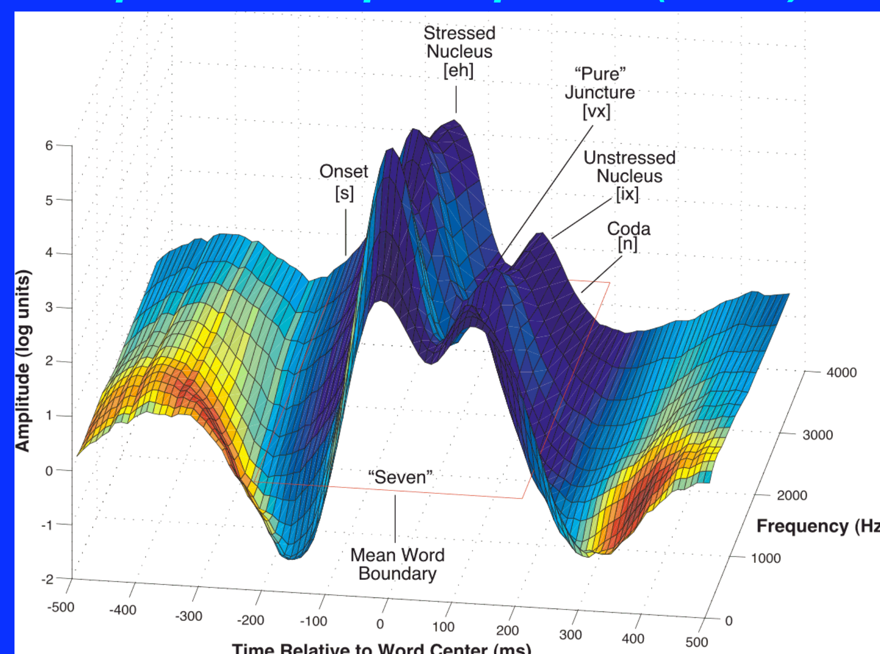
The acoustic "attack" (i.e., velocity and acceleration of the energy rise) may be an important cue for place of articulation, and is consistent with voice onset time varying with place (short for labials, long for velars, etc.)

Spectrogram + Waveform



"seven"

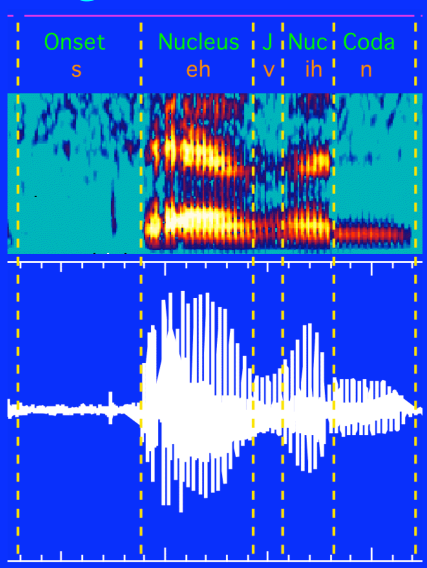
Spectro-temporal profile (STeP)



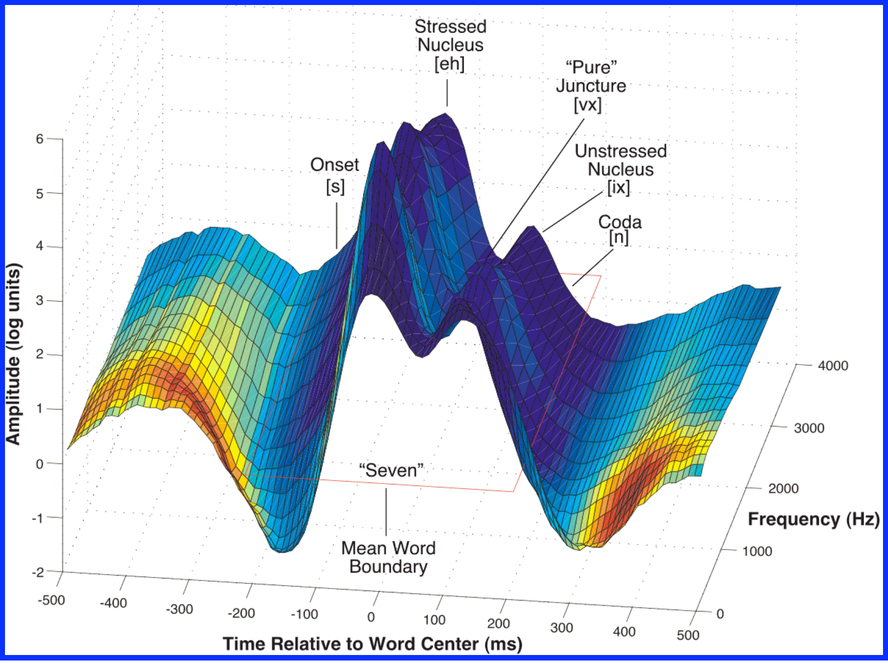
Place of Articulation

The cues for place information ride on top of the coarser syllable dynamics cues. As the syllable rises (or falls) in energy there is a slightly higher rise in energy that carries the place information

Spectrogram + Waveform



Spectro-temporal profile (STeP)



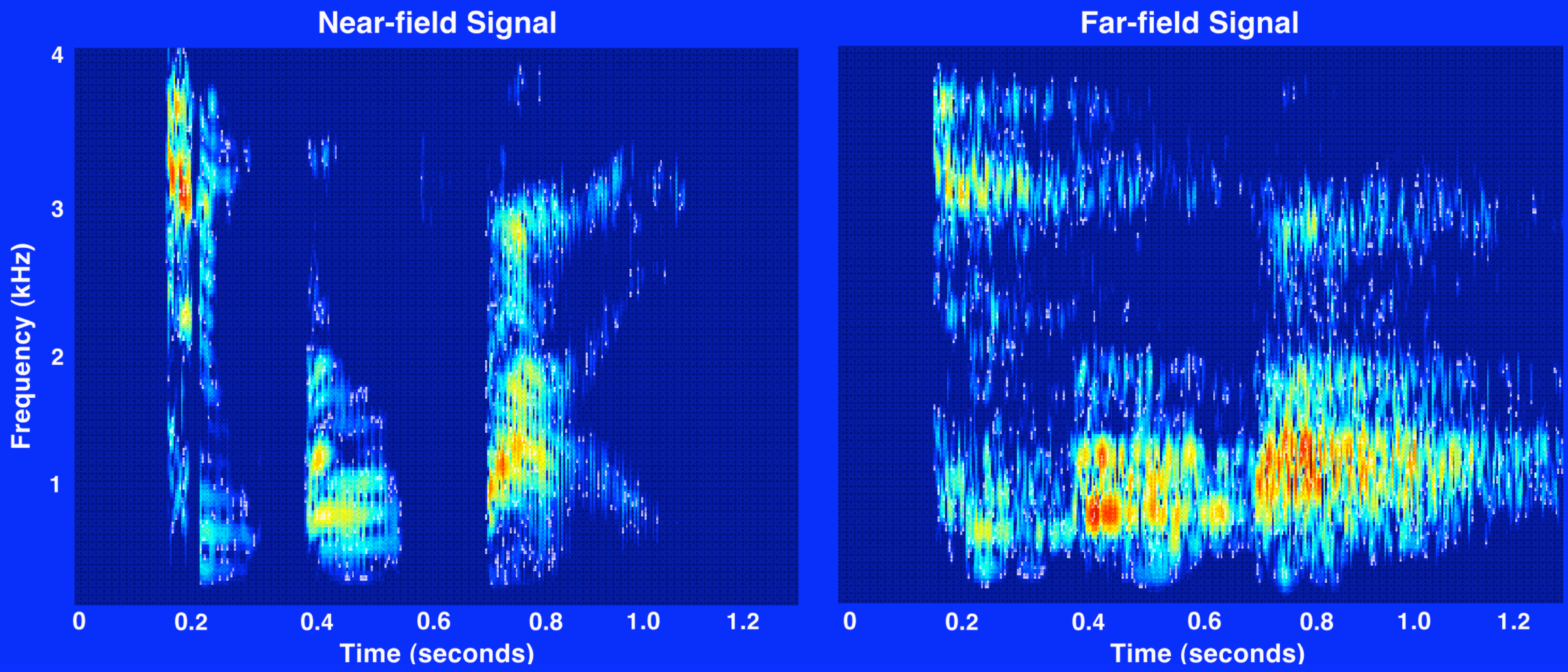
"seven"

Place of Articulation

This additional information is more vulnerable to extraneous background noise and reverberation

Place cues may also be extracted from the initial 20 ms of a stop burst, but this information can be lost or distorted in real-world speaking conditions

Thus, the primary ACOUSTIC cues for place of articulation only APPEAR to be formant transition patterns pointing to a specific locus region



Place of Articulation

These cues are reinforced by the visual, speechreading information (as mentioned earlier)

The inherent robustness of place cues may be largely due to their bi-modal nature

In the absence of visual cues, place information is extremely vulnerable to background noise

And in the presence of incongruent visual cues (i.e., the McGurk effect) the percept is often governed or influenced by non-acoustic information

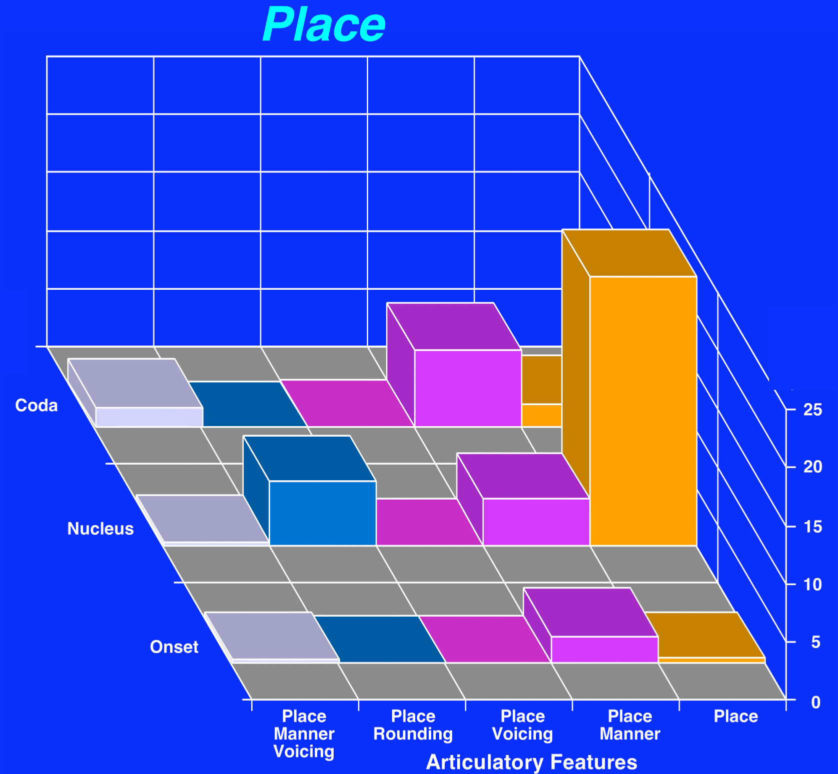
Suggesting that the ACOUSTIC cues associated with place are fragile and ambiguous

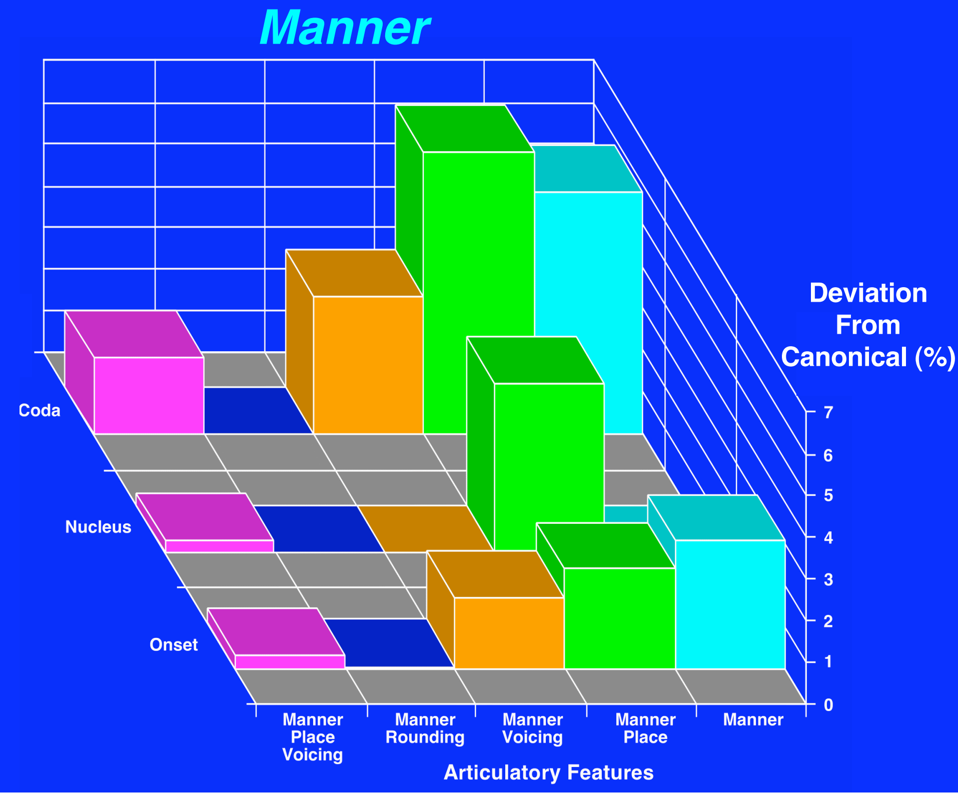
Place of Articulation

Ironically, place information is historically more robust than manner information – cognates and genetically related lexical forms are usually closer in (functional) place than in manner

This is reflected in the frequency with which pronunciation in spontaneous speech deviates from the canonical form (in terms of articulatory features)

Place cues are much less likely to deviate from the canonical than manner (in onsets and codas) (recall that prosody doesn't affect place realization)

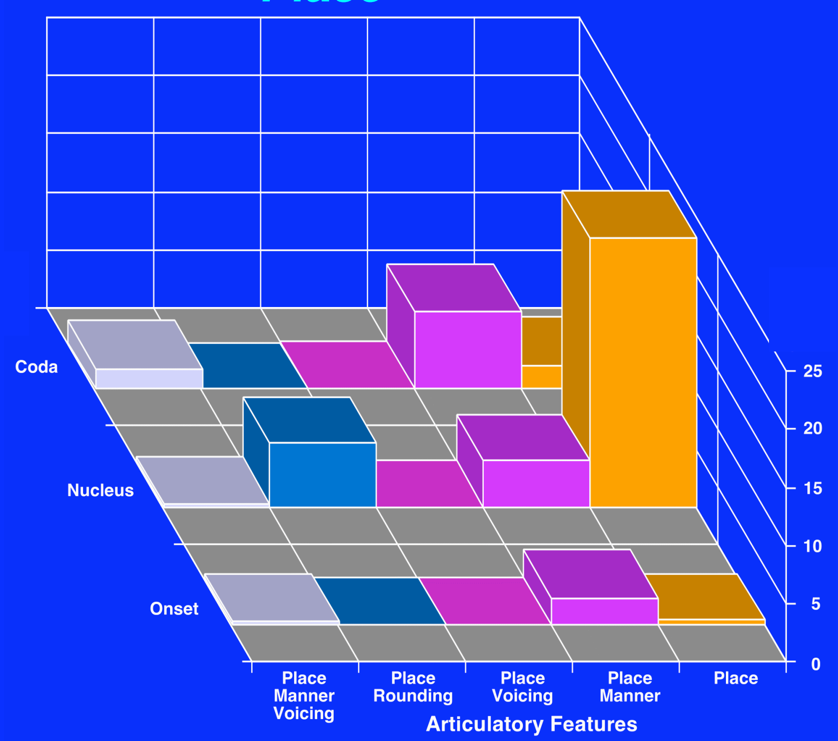




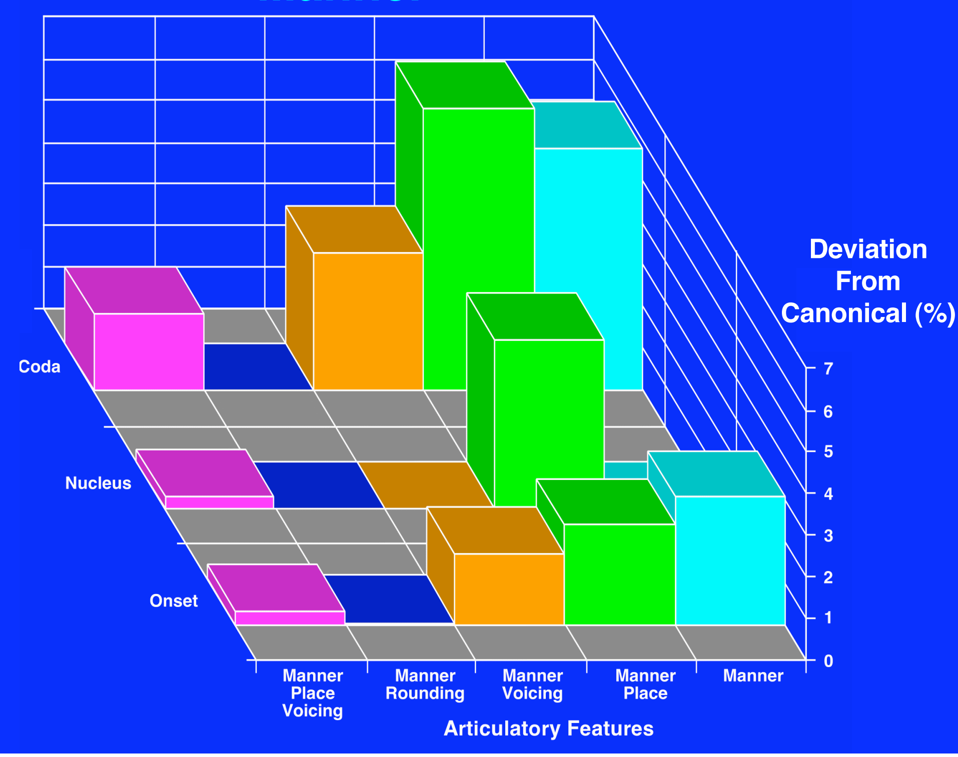
Manner vs Place Stability Across Time

- ***Manner information is much less stable historically over time – consistent with the greater likelihood of deviation from canonical pronunciation***
- ***It seems likely that manner of articulation is largely under prosodic control given its association with the fine details of onset and coda contours (e.g., stops becoming fricated, nasals dropping in coda position in favor of nasalization of the preceding vowel, etc.) and is consistent with sound change being the product mostly of prosodic forces***

Place



Manner



Pronunciation Variability

of

“Real” Speech

(and why it matters)

Pronunciation Variability of Real Speech

The specific ways in which words (particularly common ones) are pronounced provide important clues about the distribution of entropy in the speech signal

Such "entropy" patterns can be observed in terms of which segments are commonly deleted in spontaneous speech

As shown on the following slide ….

How Many Pronunciations of “and”?

N	Pronunciation				
82	ae	n			
63	eh	n			
45	ix	n			
35	ax	n			
34	en				
30	n	*Canonical pronunciation*			
20	ae	n	dcl	d	
17	ih	n			
17	q	ae	n		
11	ae	n	d		
7	q	eh	n		
7	ae	nx			
6	ae	ae	n		
6	ah	n			
5	eh	nx			
4	uh	n			
4	ix	nx			
4	q	ae	n	dcl	d
3	eh	n	d		
3	q	ae	nx		

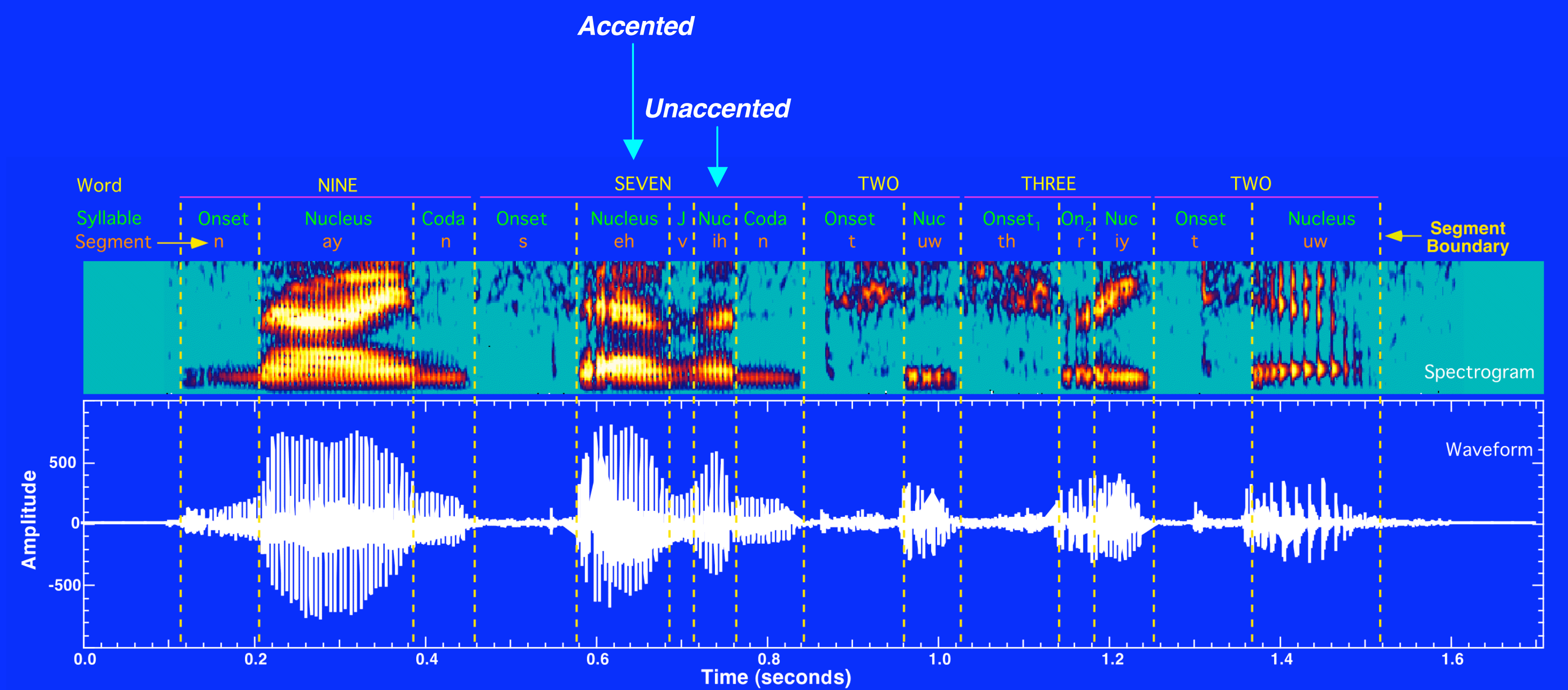
N	Pronunciation				
3	eh				
2	ae	n	dcl		
2	ae				
2	ax	m			
2	ax	n	d		
2	ae	eh	n	dcl	d
2	eh	n	dcl	d	
2	ax	nx			
2	q	ae	ae	n	
2	q	ix	n		
2	ix	n	dcl	d	
2	ih				
2	eh	eh	n		
2	q	eh	nx		
2	ix	d	n		
1	eh	m			
1	ax	n	dcl	d	
1	aw	n			
1	ae	q			
1	eh	dcl			

The Importance of Syllable Structure

The information contained in the speech signal is non-uniformly distributed

Stressed syllables contain more information than unstressed syllables

And syllable onsets are more informative than codas



The Importance of Syllable Structure

In the analyses to follow, the phonetically realized data (from the phonetic transcripts) are directly compared to the "canonical" pronunciations (from a standard recognition lexicon)

The analyses are therefore in terms of "deviation from canonical" pronunciation

Such data serve to illustrate the sort of variation observed that is conditioned by position within the syllable

(i.e., "ONSET" - "NUCLEUS" - "CODA")

As well as gauge the impact of syllable prominence on phonetic patterning

(i.e., "HEAVY" - "LIGHT" - "NONE")

Pronunciation Variation – Syllable and Accent

- **Stress accent has a direct impact on the probability of canonical pronunciation (which is related to entropy)**
- **Unaccented syllables are far more likely to be non-canonically pronounced than their accented counterparts**

All Segments

Pronunciation Variation – Substitutions

Most of the SUBSTITUTION deviations occur in the NUCLEUS

Stress accent level has a profound impact on the probability of substitutions

NUCLEUS Territory

Pronunciation Variation – Deletions

Most of the DELETION deviations occur in the CODA

Stress accent has a significant impact on the probability of coda deletion

CODA Territory

Pronunciation Variation – Summary

Different components of the syllable are "specialized" wrt to pronunciation patterns (at least with respect to deviation from the canonical form)

The NUCLEUS is associated with SUBSTITUTIONS

The CODA is associated with DELETIONS

The patterns ultimately reflect the distribution of information in speech

All Segments

Deletions

CODA
Territory

Substitutions

Insertions

NUCLEUS
Territory

ONSET
Territory

Pronunciation Patterns – Syllable Codas

The ANTERIOR and POSTERIOR codas are usually CANONICALLY realized, similar in pattern to the onsets

The CENTRAL (coronal) codas are often non-canonical articulated

The following slide illustrates (in part) why this may be so

Place of Articulation									Approximants		
Anterior			Central			Posterior			Chameleon		
SEG	Onset	Coda	SEG	Onset	Coda	SEG	Onset	Coda	SEG	Onset	Coda
p		C	t		N	k		C	l		N
b		C	d		N	g		C	lg		N
m		N^0	dx		ø	ng		N^0	r		N
f		C	n		N	sh		C	hh		ø
v		N^0	nx		ø	zh		C			
th		C	s		C	ch		C			
dh		ø	z		N	jh		N			
y		ø				w		ø			
						q		N			

C = Canonical realization
N = Non-canonical realization, N^0 = Non-canonical in unaccented syllables

Why do Coronal Coda Segments "Delete" So Often?

There is something "special" about coronal segments (in coda position)

A significant proportion of these segments are phonetically unrealized

One potential "explanation" pertains to the trajectory of the second formant (reflecting the front cavity resonance)

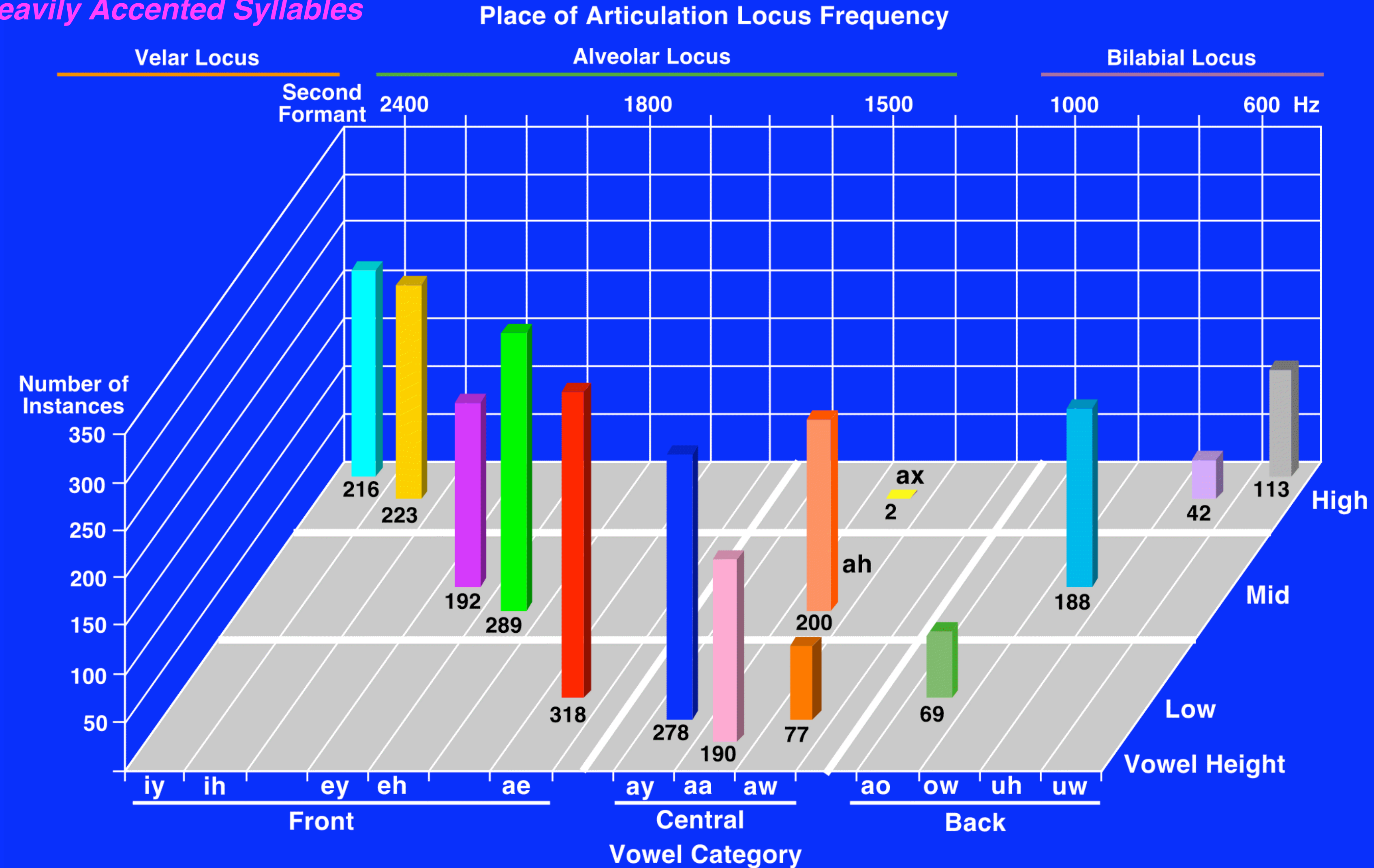
The locus (target) frequency of coronals is ca. 1500-2500 Hz, similar to the second formant of the front and central vowels

Given the preponderance of non-back vowels in the corpus, the second formant for vocalic segments preceding a coda consonant is likely to be between 1500 and 2500 Hz

Why do Coronal Coda Segments "Delete" So Often?

The absence of a coda segment points, by implication, to the coronal place of articulation under many circumstances

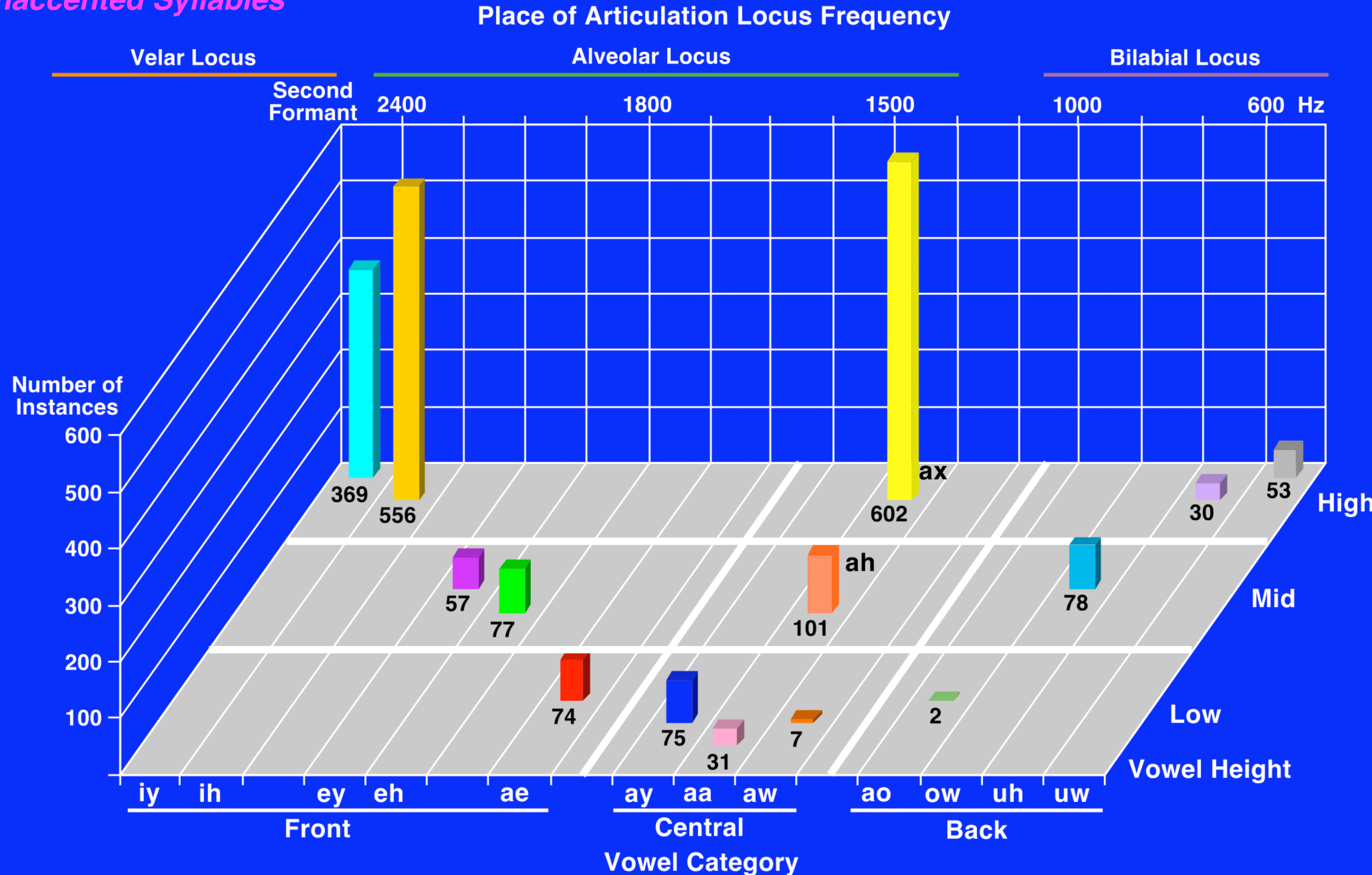
Heavily Accented Syllables



Why do Alveolar Coda Segments "Delete" So Often?

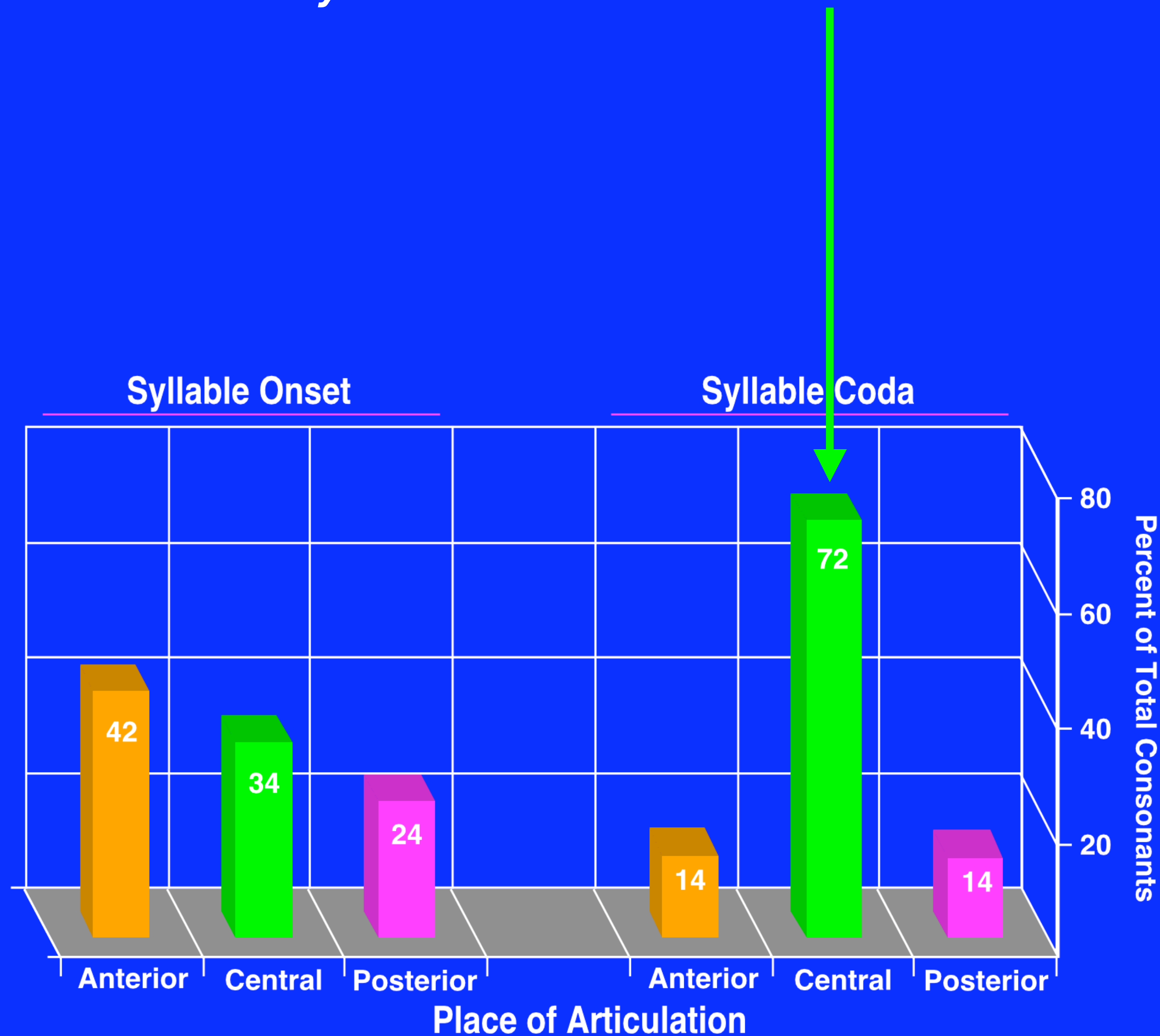
The absence of a coda segment points, by implication, to the coronal place of articulation under many circumstances

Unaccented Syllables



Preponderance of Coda Coronals

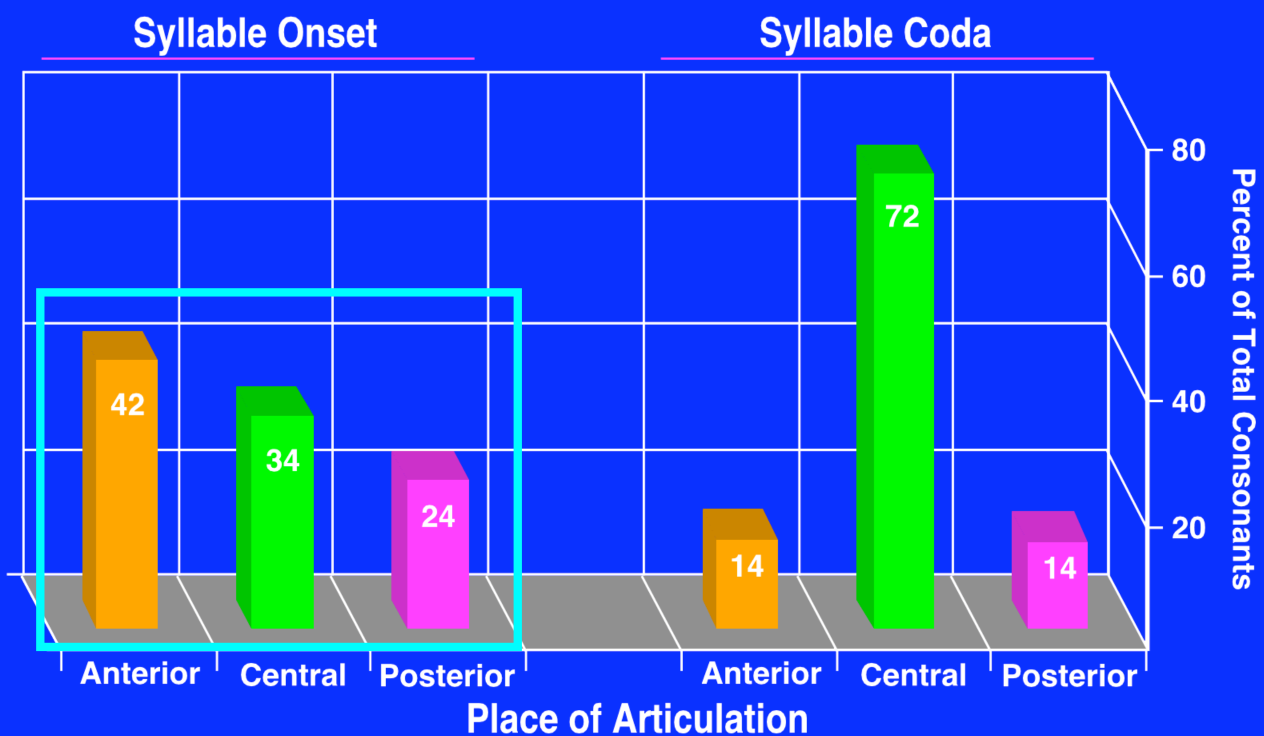
This may also account for why 75% of all coda consonants are coronals



All accent levels combined (canonical elements)

Preponderance of Coda Coronals

In contrast is a far more equitable distribution across place among onsets



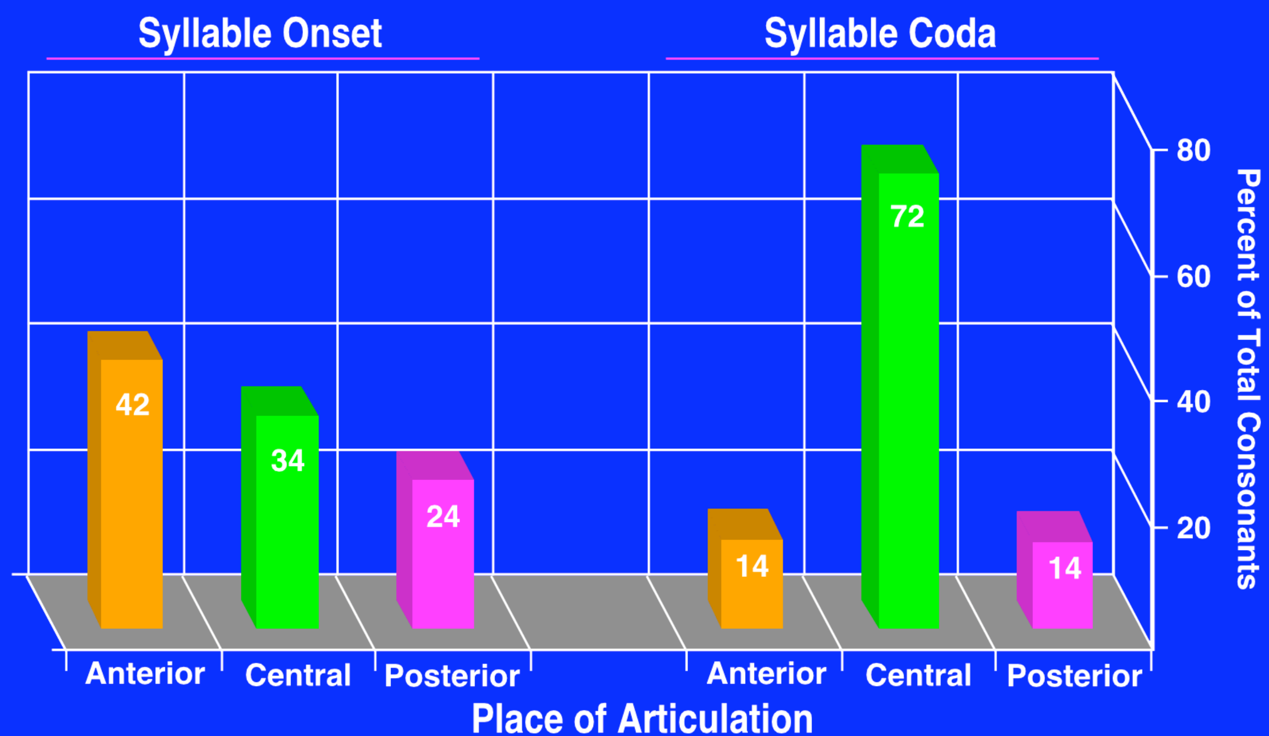
All accent levels combined (canonical elements)

Preponderance of Coda Coronals

Nearly three-quarters of the CODA consonants are CORONALS

In contrast is a far more equitable distribution across place among onsets

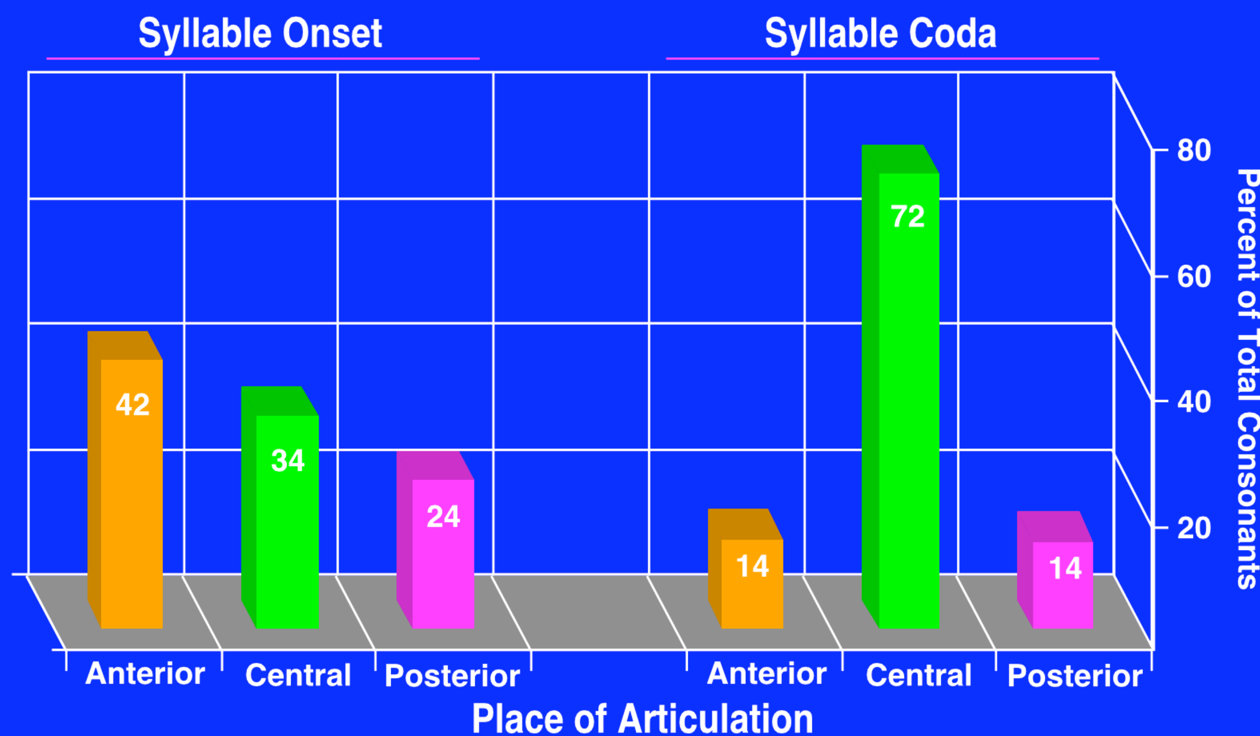
The disparity in place distribution in coda position implies that coronals are a "default" category, able to sustain deletion without undue impact on the information contained within the syllable



All accent levels combined (canonical elements)

Preponderance of Coda Coronals

- *Nearly three-quarters of the CODA consonants are CORONALS*
- *In contrast is a far more equitable distribution across place among onsets*
- *The disparity in place distribution in coda position implies that coronals are a "default" category, able to sustain deletion without undue impact on the information contained within the syllable*
- *In this sense, codas carry far less information than onsets (at least wrt place)*



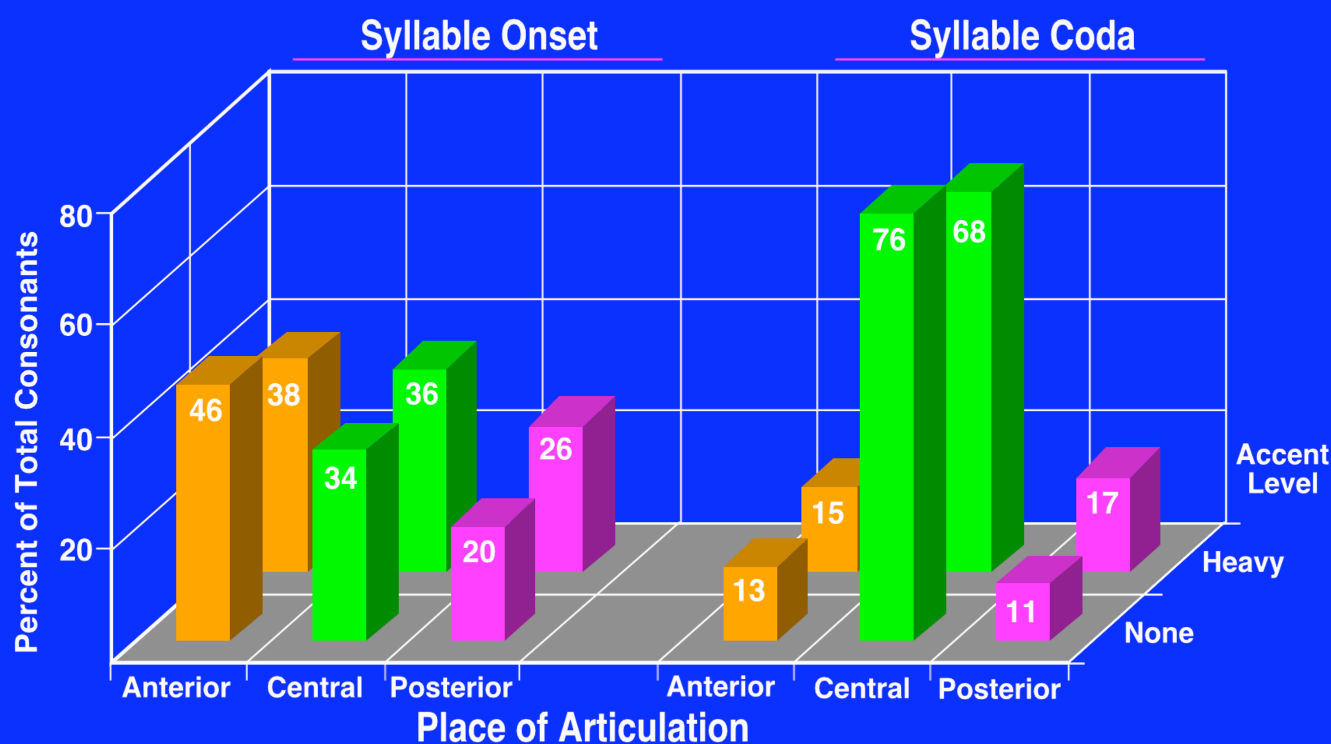
All accent levels combined (canonical elements)

Accent and Preponderance of Coda Coronals

Stress accent has relatively little impact on the distribution of place in either onset or coda segments

Particularly with respect to the preponderance of coronal segments in codas

And is consistent with the hypothesis that codas are inherently less informative than onsets regardless of accent level (and that place cues are less sensitive to prosodic factors than manner and voicing)



Unaccented and heavily accented levels combined (canonical elements)

Multi-Tier Theory – Summary

The SYLLABLE, rather than the PHONE, is the most basic organizational unit of spoken language – the patterns of pronunciation variation observed are incompatible with segment-based models

The syllable carries prosodic weight (a.k.a. "accent" or "prominence") that affects the manner in which its constituents are phonetically realized

The behavior of these syllabic constituents (a.k.a. "ONSET," "NUCLEUS" and "CODA") differ dramatically from each other, and influence the phonetic character of the syllable

Syllable position is probably as important as segmental identity for characterizing pronunciation

The MICROSTRUCTURE of the syllable can be delineated in terms of articulatory-acoustic features (e.g., voicing, articulatory manner and place)

MANNER of articulation most closely parallels (in time and behavior) the classical concept of the phonetic segment and sets the basic intensity mode for the sequence of syllabic constituents (a.k.a. the "ENERGY ARC")

The ENERGY ARC reflects cortical processing constraints on the acoustic (and visual) signal associated with the MODULATION SPECTRUM

Multi-Tier Theory – Summary

PLACE of articulation is an inherently TRANS-SEGMENTAL feature that binds vocalic nuclei with preceeding and following consonants

Formant transitions are unlikely to serve as the primary basis for articulatory place information (except, perhaps, under pristine listening conditions)

Rather, the visual,speechreading cues play an important role in decoding place of articulation information under many conditions

VOICING emanates from the nucleic core of the syllable and spreads both forward (towards the coda) and backward (towards the onset), the degree of temporal spreading reflecting the magnitude of prosodic prominence – in this sense, VOICING is a SYLLABIC rather than a phonetic-segment feature, in that it is sensitive to the prominence of the syllable

It is the PATTERN of INTERACTION among articulatory-feature dimensions across time that imparts to the syllable its specific phonetic identity

The specific REALIZATION of ARTICULATORY FEATURES is governed by prosodic PROMINENCE as well as their POSITION within the SYLLABLE

The PROSODIC pattern reflects the INFORMATION contained within the utterance

Therefore, it is ultimately INFORMATION (and lexical discrimination) that governs the detailed phonetic properties of spoken language

Implications

for

Speech Technology

How to Exploit the Patterns Observed

How can the insights described in this presentation be exploited for developing future-generation speech technology?

This multi-tier framework could be useful in many different ways

It could be used to improve the quality of pronunciation models for both recognition and synthesis (a topic in and of itself ….)

It could also be used to synthesize far more realistic sounding speech than is currently possible without the use of sophisticated unit-selection methods (and thus be able to simulate a broad range of emotions and speaking styles without the need to record representative materials for each new condition)

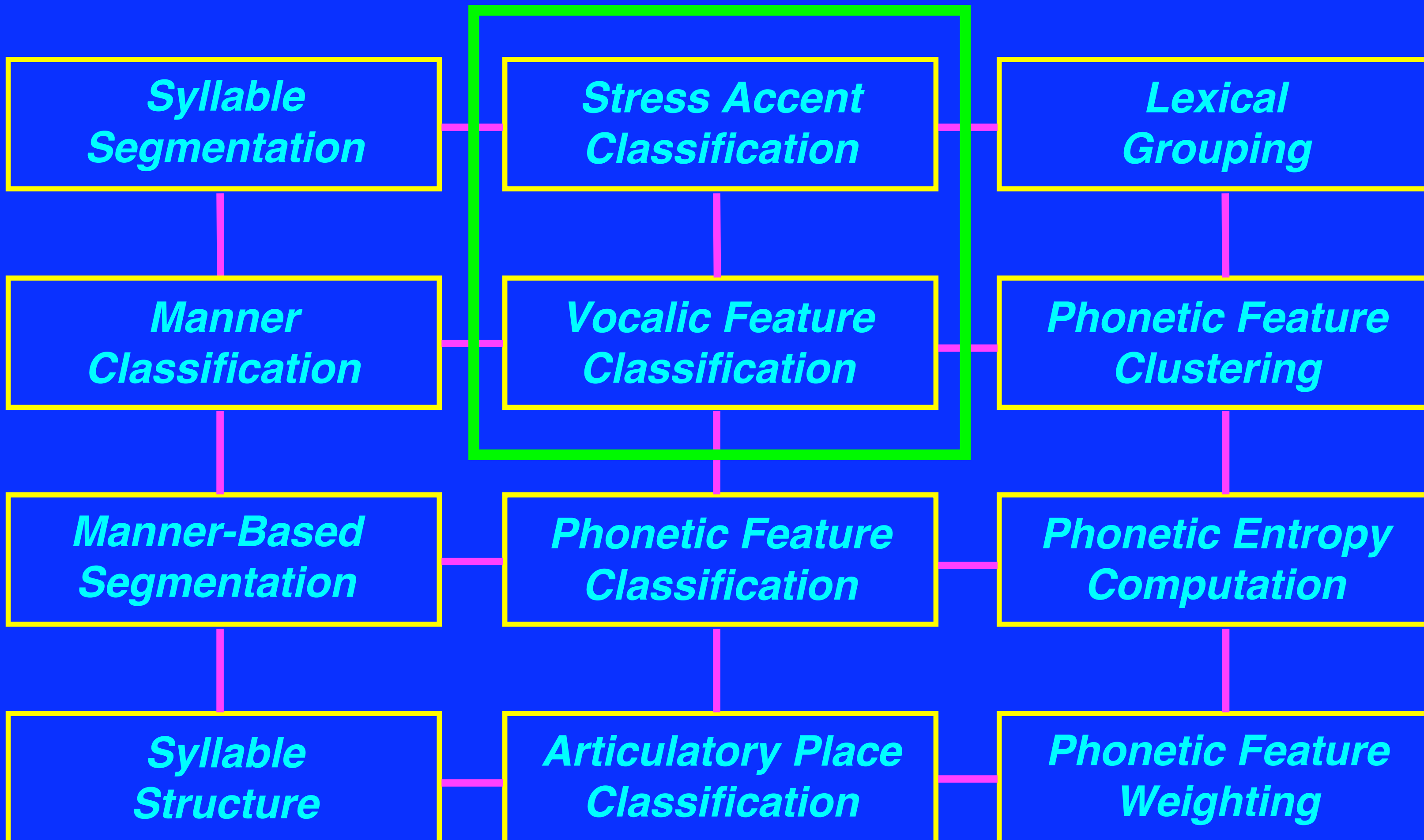
It could enable recognition systems to be freed from the bondage of extensive training material for each new speaking style and task

It could also be used to guide the signal enhancement algorithms for hearing aids and speech separation systems

Speech Analysis – The Full Monty

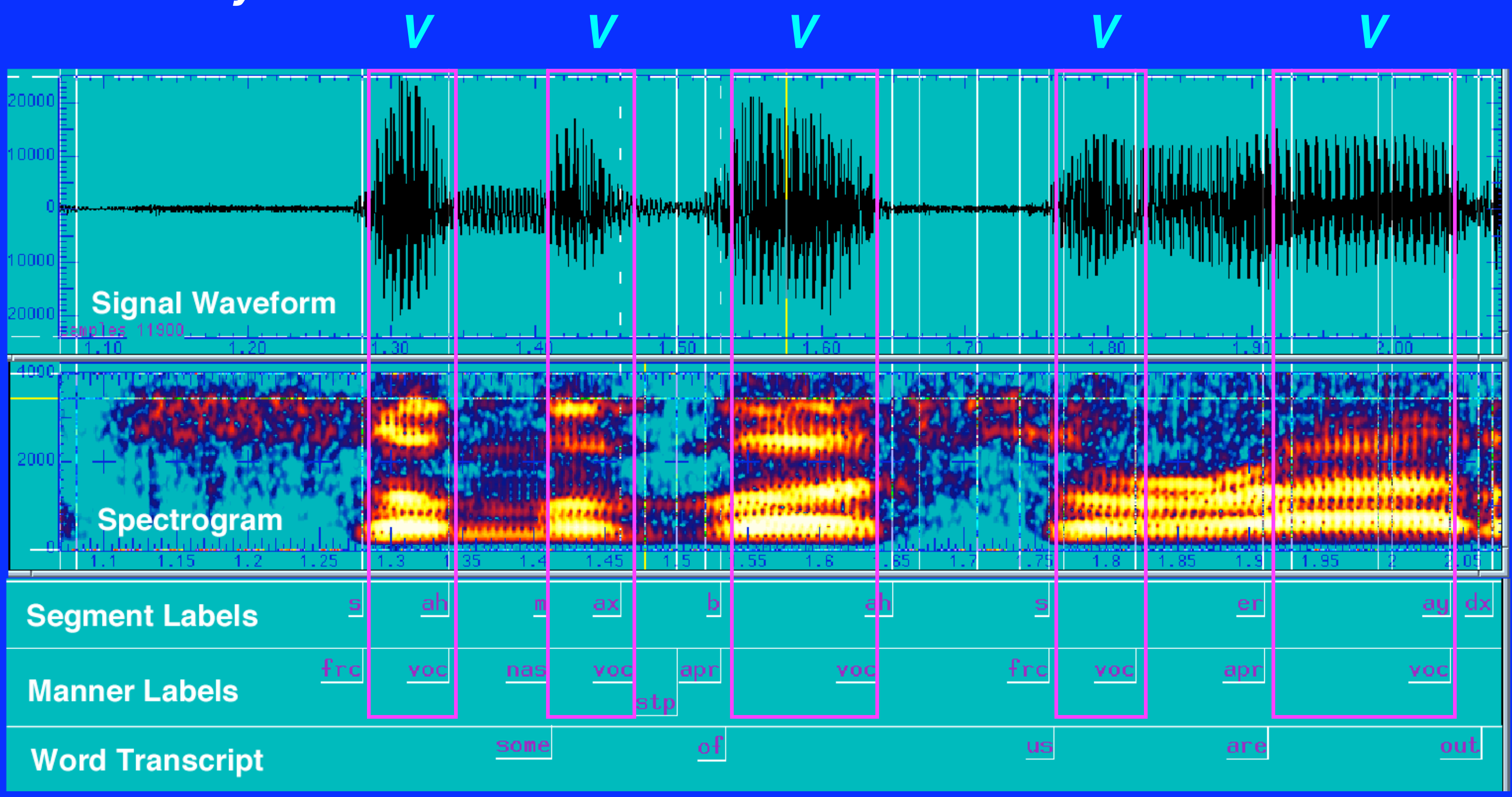
Time does not permit an exhaustive discussion, so I'll focus on a single prospective application

Namely, extracting the syllable nucleus and computing the prosodic weight of the associated vocalic constituent (on the following slides)



Using Manner to Spot Syllable Nuclei

- ***As mentioned earlier, manner of articulation is temporally isomorphic with phonetic segments***
- ***Manner classifiers are particularly adept at spotting vocalic segments with high precision***
- ***For this reason, it is possible to delineate syllable nuclei with a high degree of accuracy***



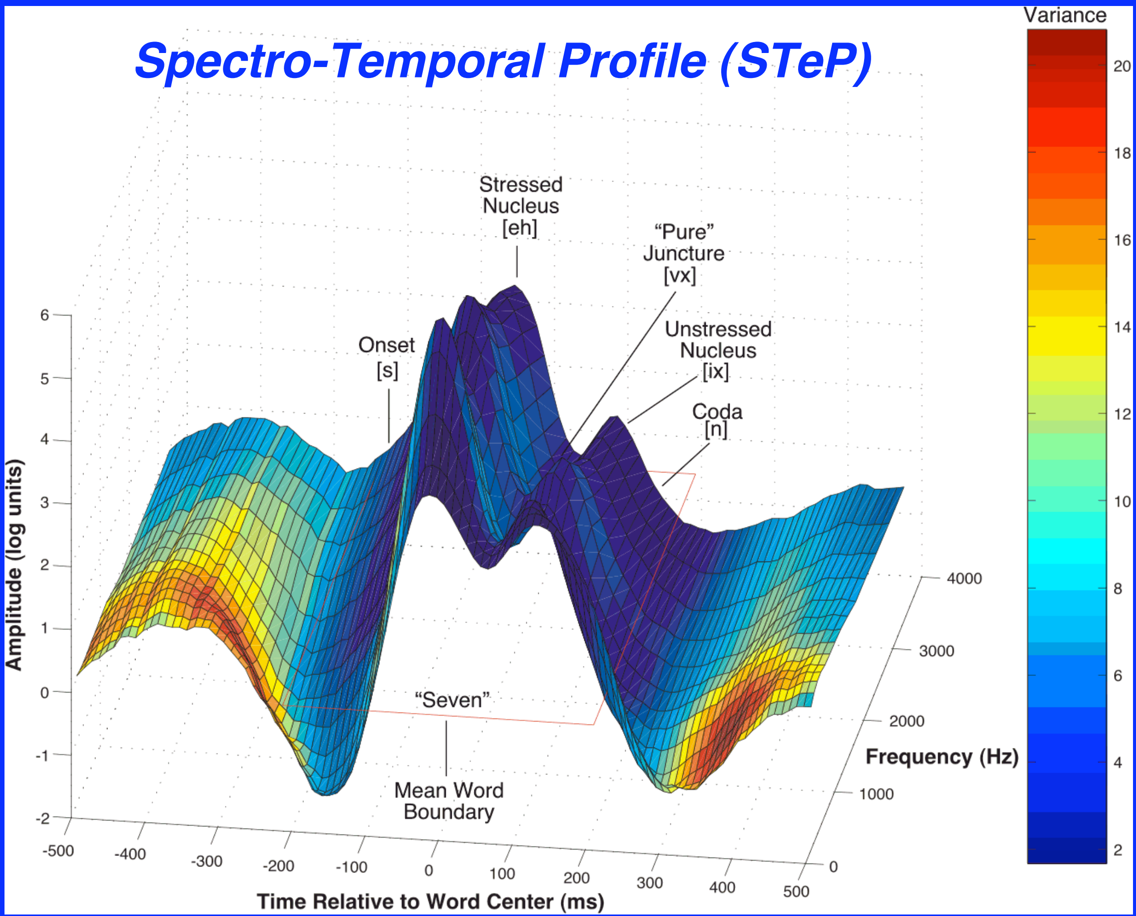
From Syllable Nucleus to Prosody

The nucleus contains much of a syllable's energy

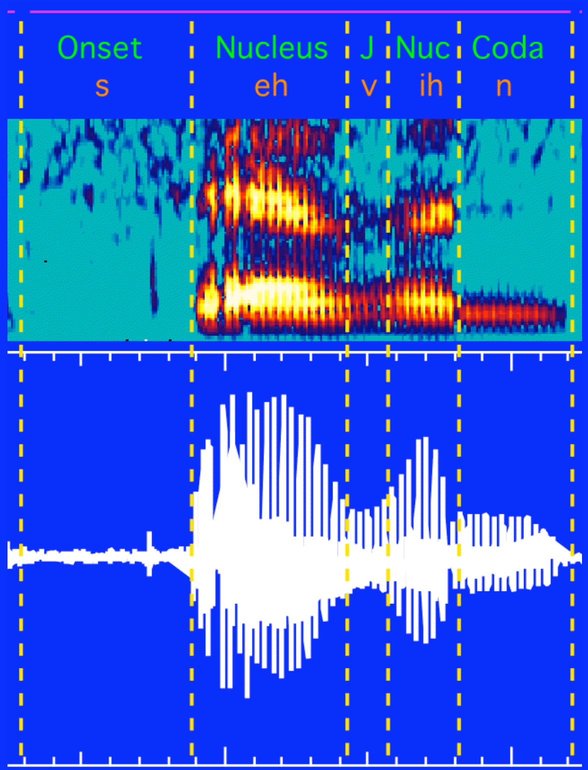
And also conveys important information about the syllable's prominence or "accent" (for languages such as English, a.k.a. "stress")

As shown below for the word "seven"

Spectro-Temporal Profile (STeP)



Spectrogram+Waveform



Greenberg et al. (2003)

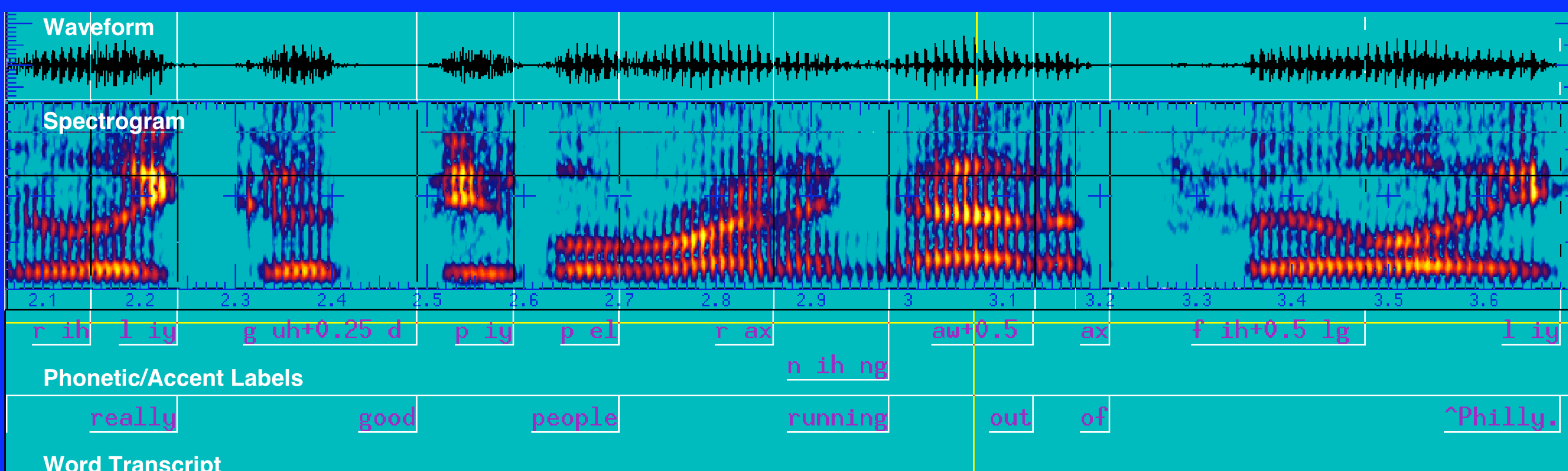
Automatic Annotation of Stress Accent

Given the importance of stress accent for characterizing the phonetic properties of the speech, is it feasible to automatically label a corpus in this way?

An automatic stress accent labeling system (AutoSAL) is capable of labeling the Switchboard corpus using 5 levels of stress

Heavy (1)   Moderate (0.75)   Light (0.5)   Very Light (0.25)   None (0)

An example of the annotation (attached to the vocalic nucleus) is shown below. In this example most of the syllables are unaccented, with two labeled as lightly accented (0.5) (and one other labeled as very lightly accented (0.25))



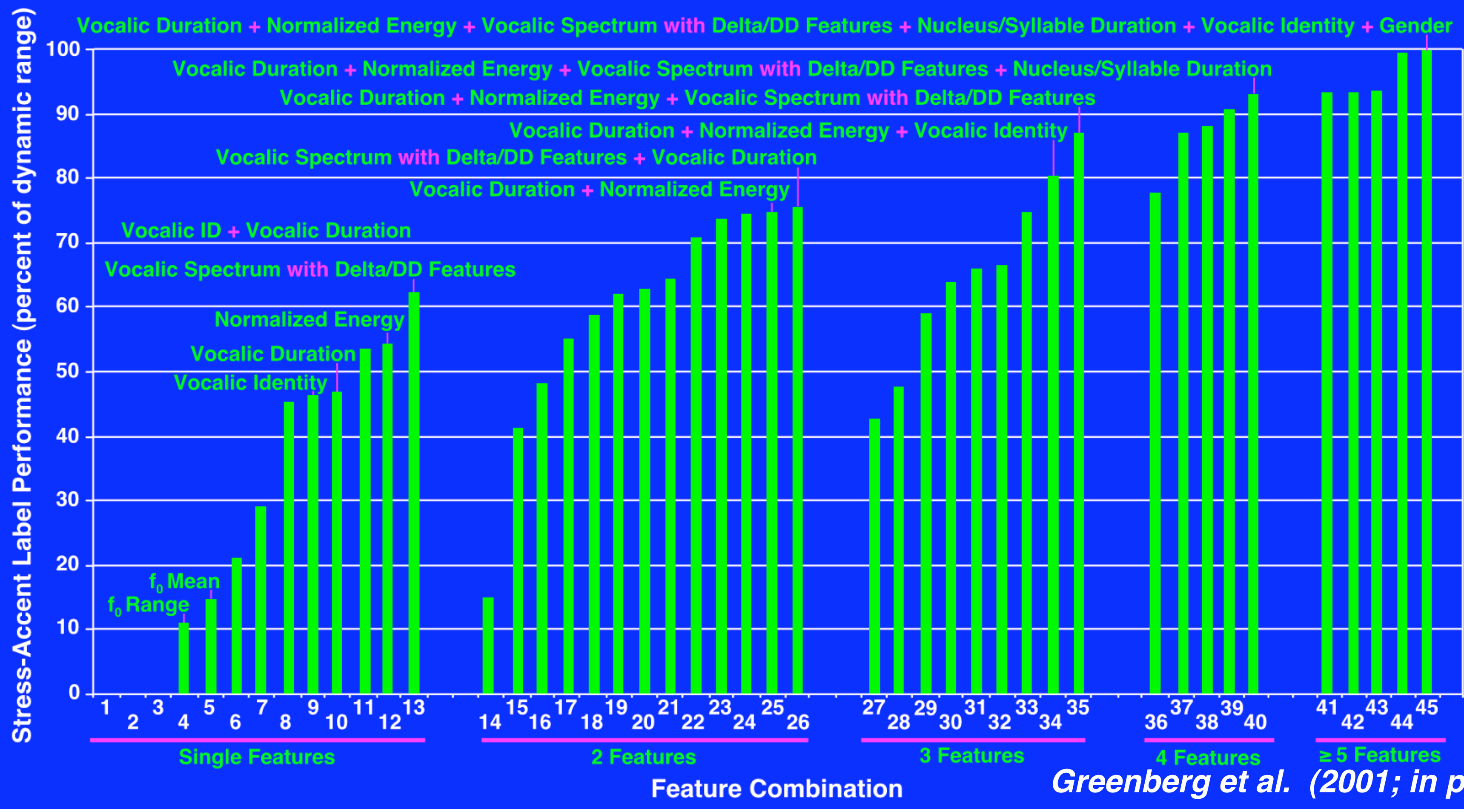
Greenberg et al. (2001)

The Acoustic Bases of AutoSAL

What are the most important features for simulating stress-accent labeling using AutoSAL?

Duration, (normalized) energy, vocalic identity (and its acoustic correlates)

Pitch-related features are (relatively) unimportant for stress-accent labeling



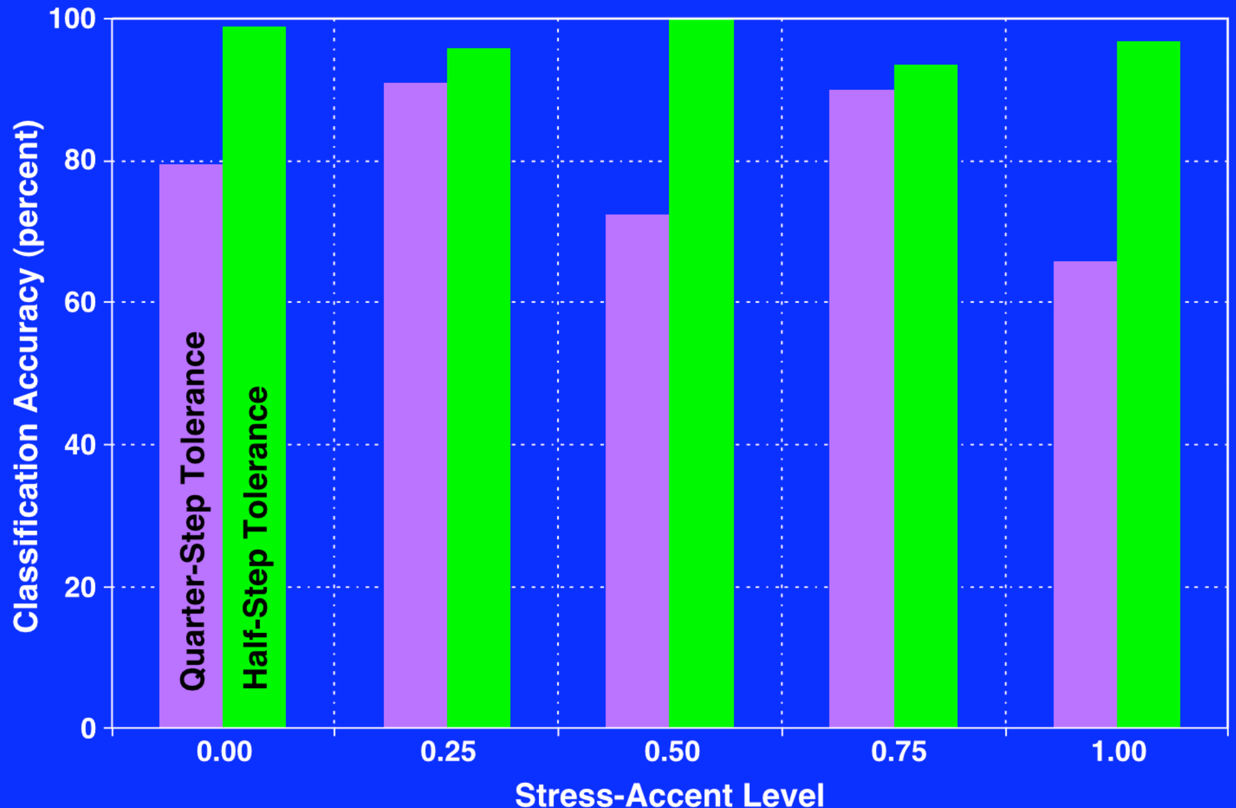
Greenberg et al. (2001; in press)

How Good is AutoSAL?

There is an 79% concordance between human and machine accent labels when the tolerance level is a quarter-step

There is 97.5% concordance when the tolerance level is half a step

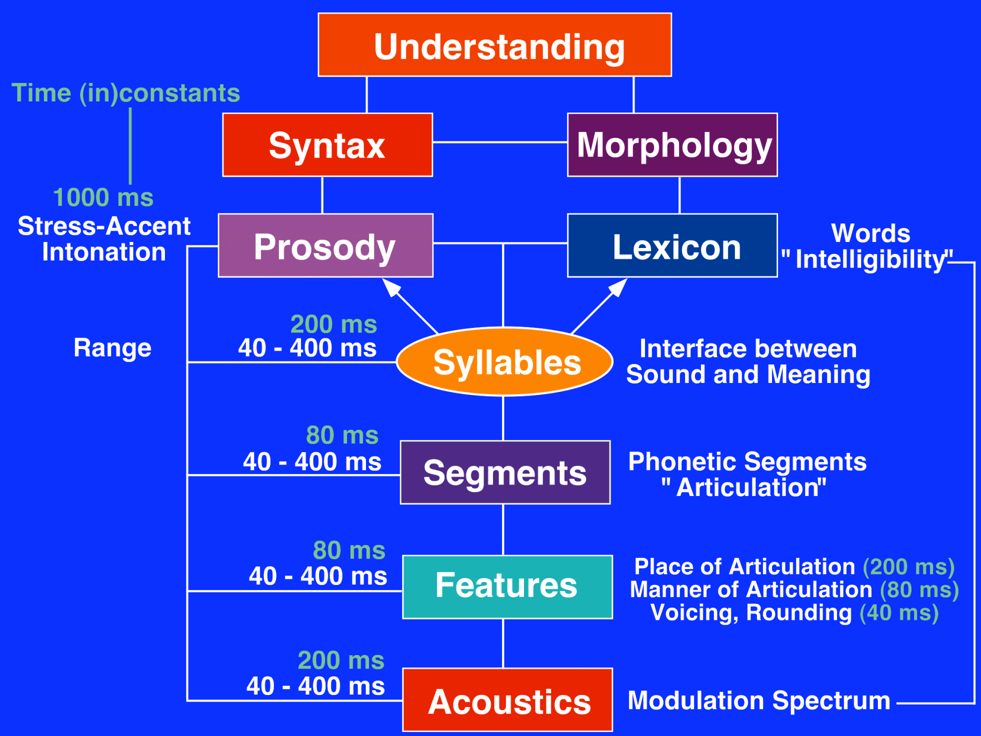
This degree of concordance is as high as that exhibited by two highly trained (human) transcribers



Greenberg et al. (2001)

Summary and Conclusions

- ***All great technology is based on a solid scientific foundation***
- ***A reliable means of establishing such a foundation is through melding sophisticated theoretical development and empirical research***
- ***A multi-tier perspective is a promising approach to developing the requisite scientific base***
- ***One that focuses on the interaction among the linguistic levels and relates this knowledge to speech spoken in the real world***



That's All

Many Thanks for Your Time and Attention