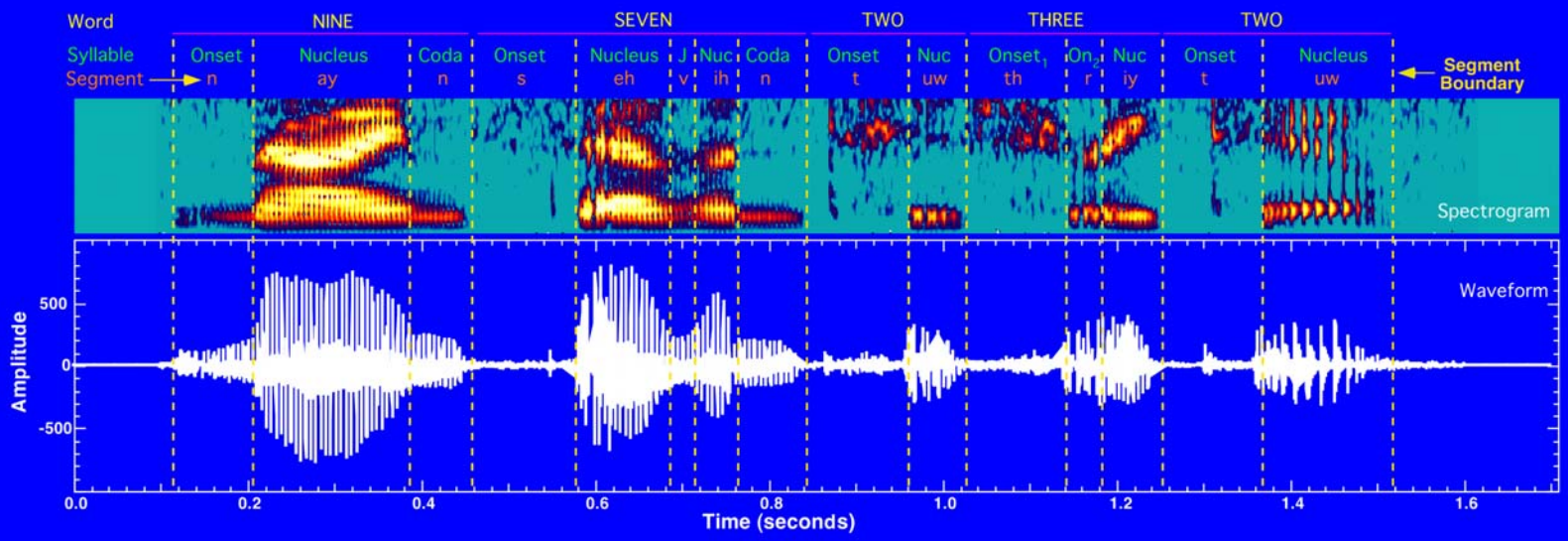
The Analysis of Spoken Language:

Past, Present, Future

Steven Greenberg

The Speech Institute



Acknowledgements and Thanks

Research Funding

U.S. Department of Defense

U.S. National Science Foundation

Scientific Collaborators

Takayuki Arai

Hannah Carvey

Shuangyu Chang

Ken Grant

Leah Hitchcock

Structure of the Presentation

A Vision of the Future

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

Key Questions –

What are the relevant units of analysis?

What are their physical signatures (in time, frequency and space)?

How can they be automatically and reliably extracted from the speech signal?

How can these units be combined for an accurate, detailed characterization of spoken language?

How can such information be exploited to enhance speech recognition and synthesis performance?

The Path to Utopia

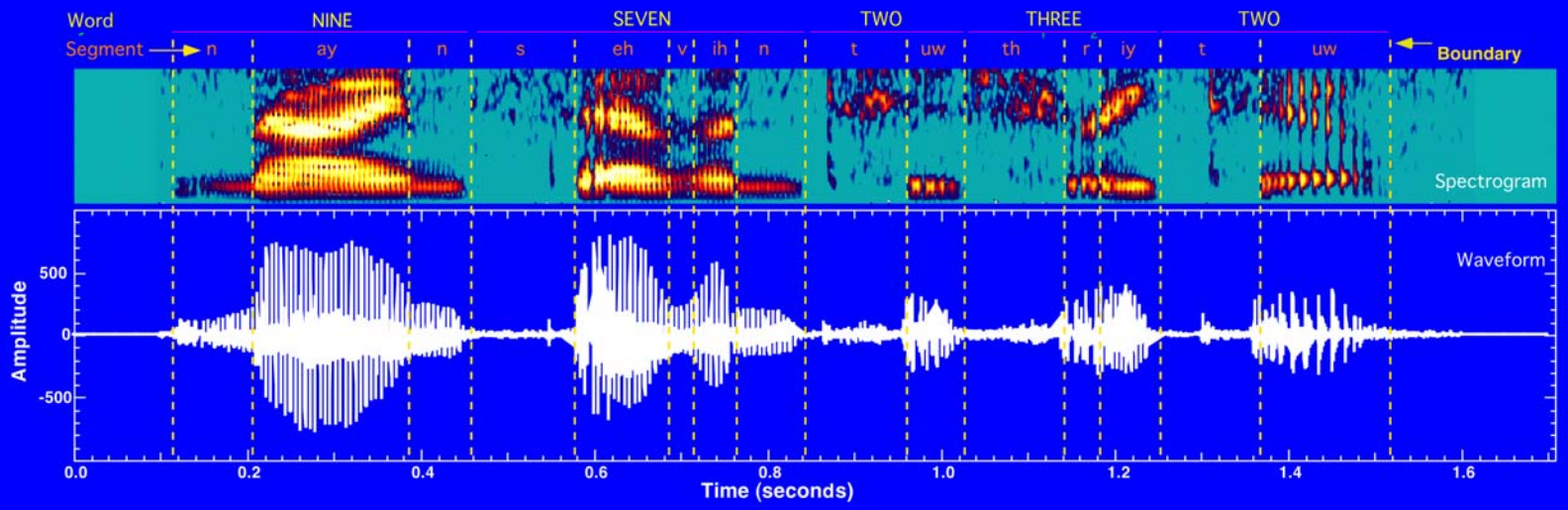
Where are we now, and how did we get here?

Where should we head?

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 transparent means with which to analyze and characterize the speech signal, as shown below

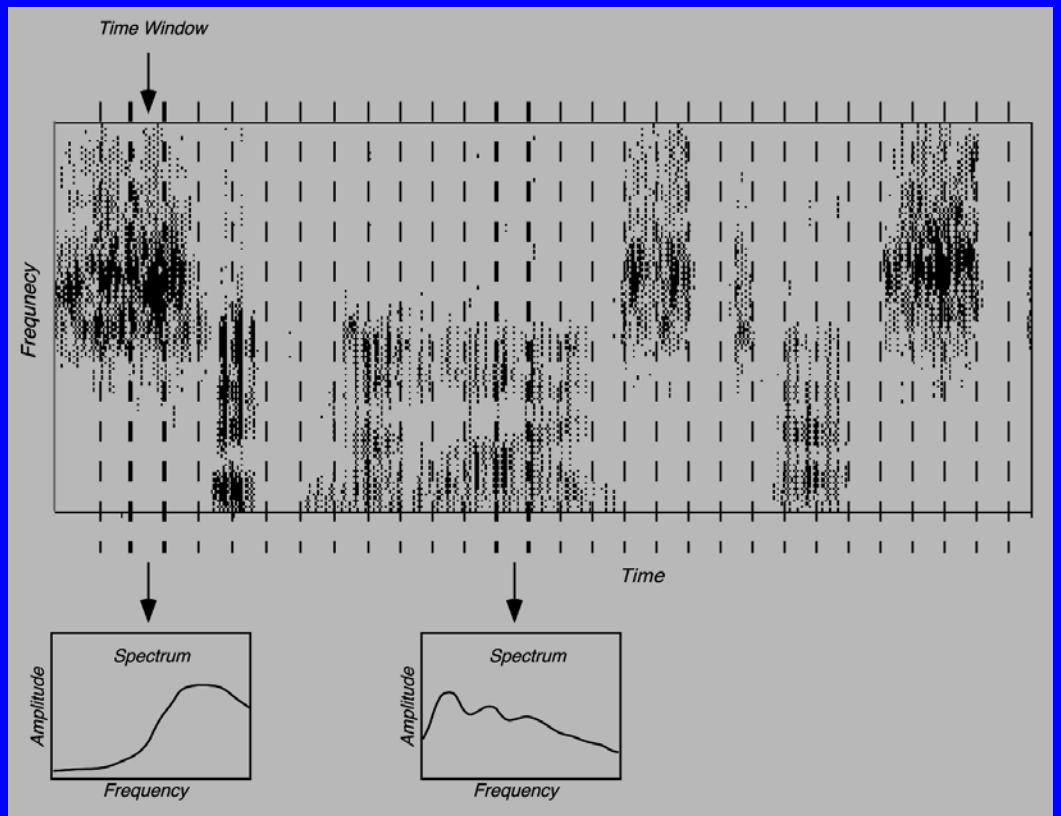


The Serial Frame Analysis Perspective

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)

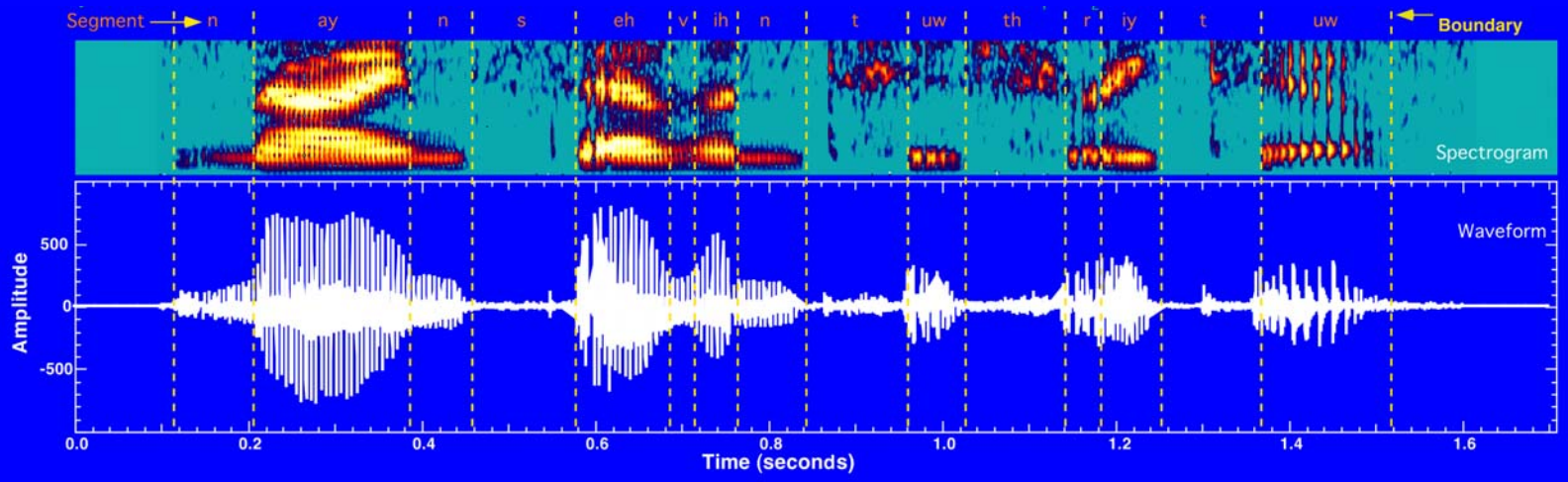
This method of analysis is relatively transparent to perform, as it requires no <u>a priori</u> knowledge of the signal



The Serial Frame Analysis Perspective

Within the framework of automatic speech recognition, each frame is associated with a symbol, representing either a phone or non-speech class (e.g., "silence," cough, filled pause, etc.)

Within the HMM framework, each interval of speech MUST be labeled (as something)

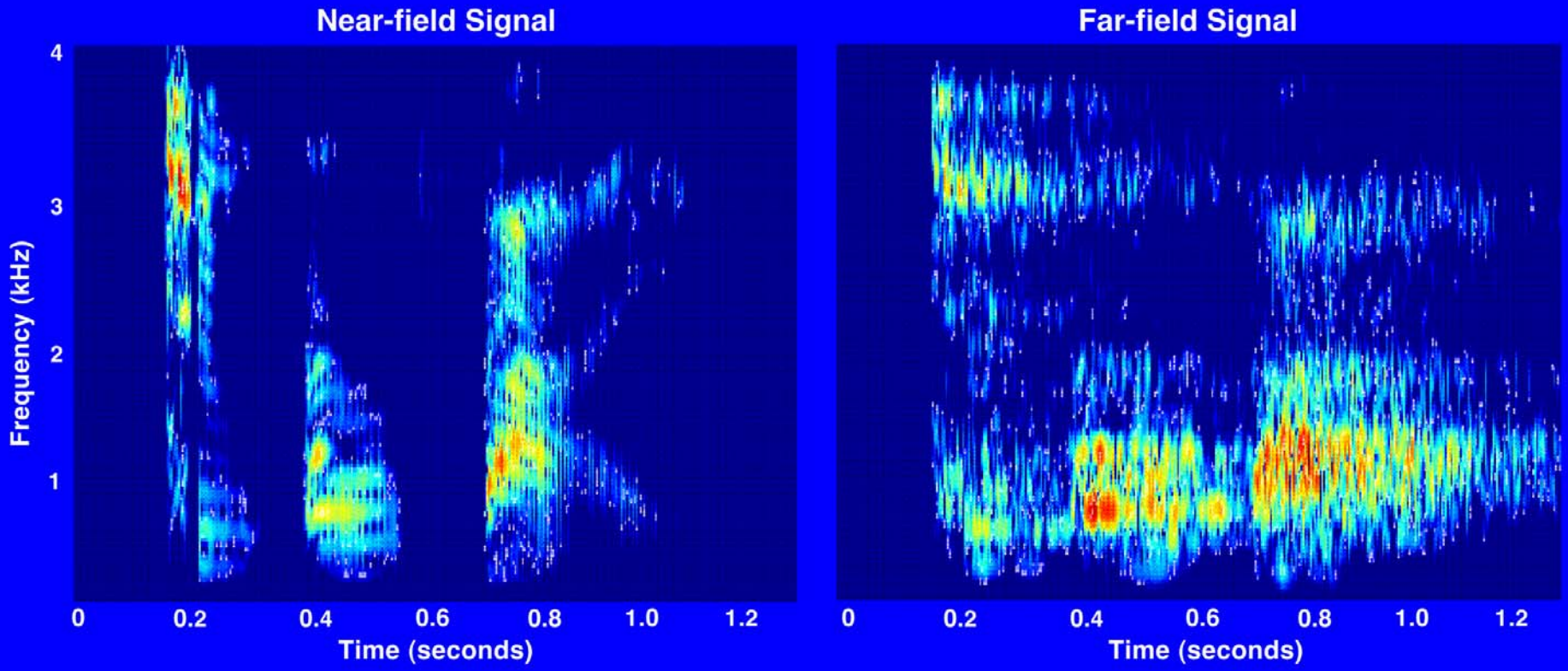


Challenge # 1 – Environmental Variability

As seductive as the serial-frame, egalitarian framework may be, there are three 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 "that" (Switchboard)

N	Pronunciation			N	Pronunciation		
53	dh	ae	t	2	t	aw	
31	dh	ae		2	n	ah	t
13	dh	ae	dx	2	d	ae	
10	dh	eh	t	2	dh	ih	
9	dh	ax		2	dh	ah	dx
9	dh	aw		2	ah	dx	
8	n	ae	t	1	z	d	ae
7	n	ae		1	z	ah	p
7	dh	ax	t	1	t	dh	ae
5	dh	eh		1	t	b	ae
5	dh	ah	t	1	t	ax	
4	dh	lh	t	1	t	ae	
3	th	ae	t	1	th	eh	t
3	d	ae	t	1	th	eh	
3	dh	ax	dx	1	th	ax	t
3	ae			1	th	ax	

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

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

Challenge #3 – Variation in Time and Spectrum

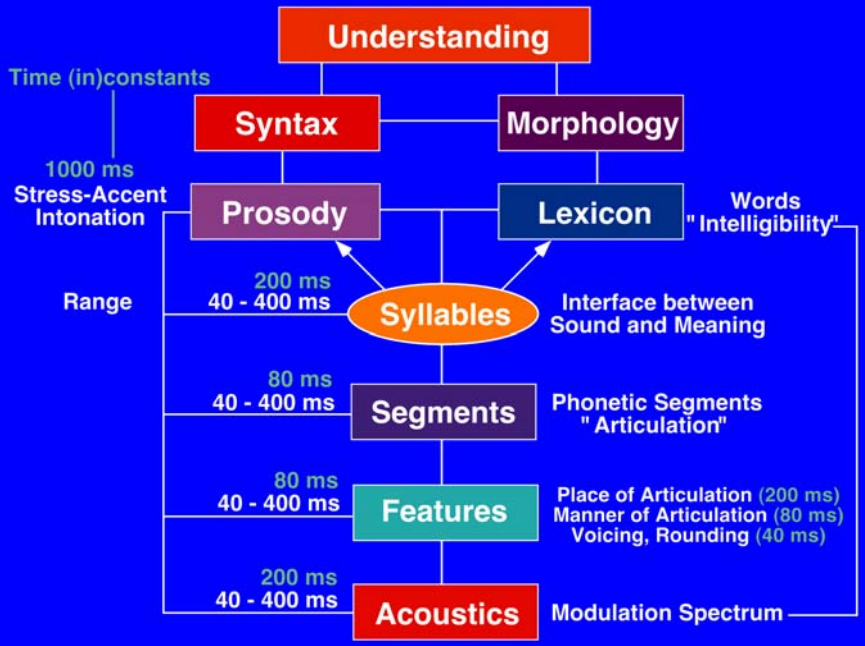
Third, the "units" of spoken language vary with respect to duration, frequency and space, thus

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

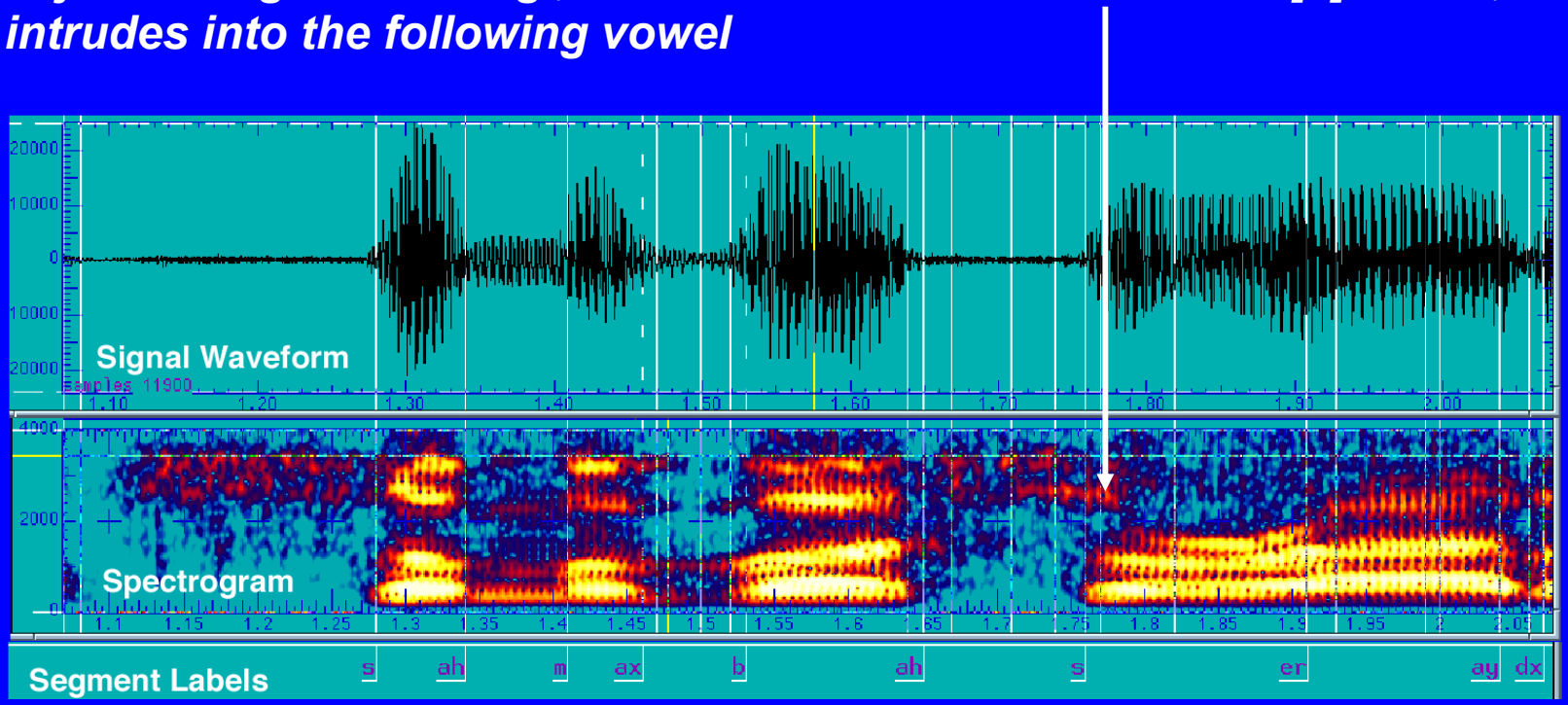
Hence, THERE IS NO SINGLE TIME INTERVAL that adequately captures all of the important acoustic and linguistic properties of spoken language



Challenge #3 – Variation in Time and Spectrum

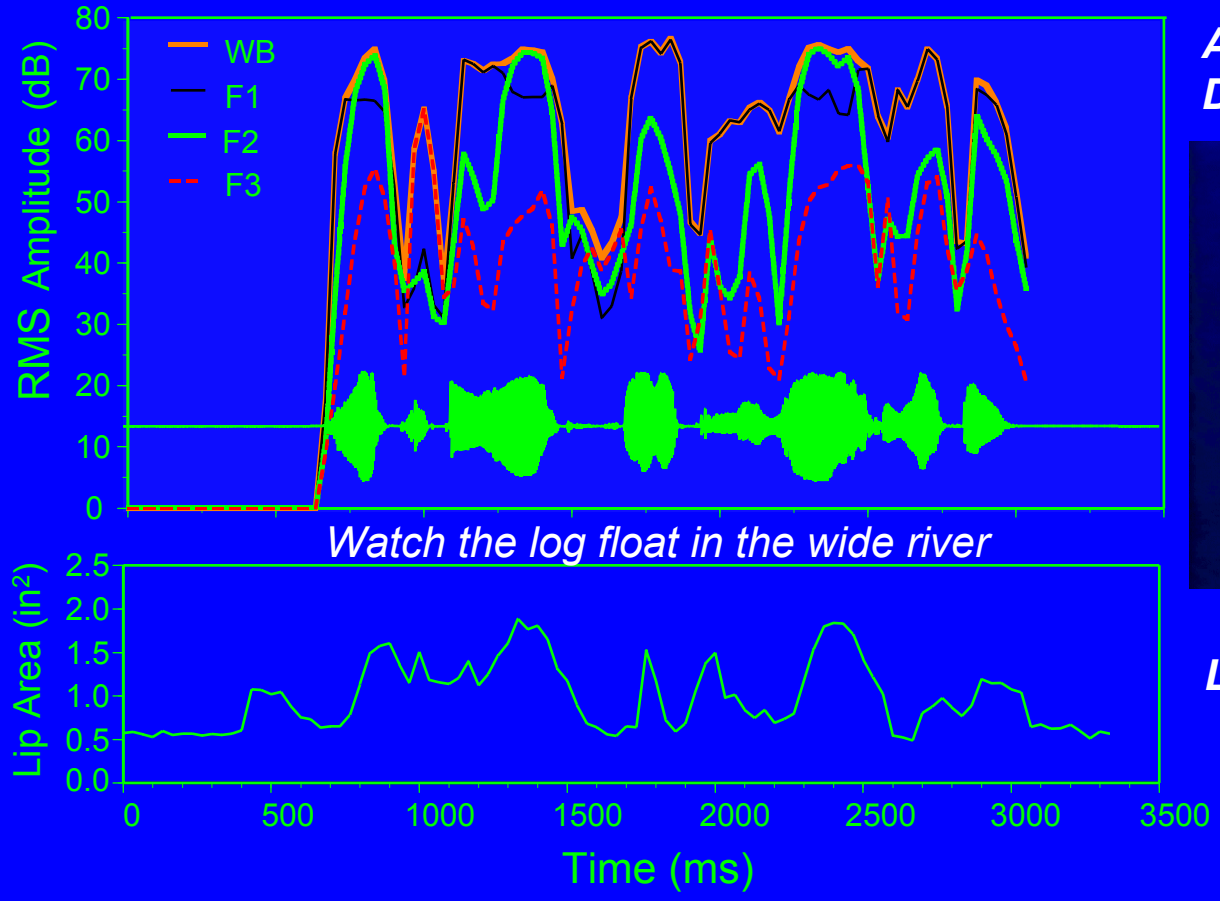
Moreover, the manner in which linguistic information is distributed across (spectral) frequency and time is non-uniform

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 #3 – Variation in Time and Spectrum

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?

In the remainder of this talk, I shall focus on three specific topics germane to the issues described in the introduction

First, I will discuss the feasibility of automatically segmenting the acoustic signal into time intervals useful for speech analysis and recognition

Then, I will discuss how such segmentation methods could be utilized within the framework of an articulatory-acoustic feature classifier that could be useful for automatic speech recognition

Finally, I will describe some recent work on automatically labeling prosodic prominence that directly ties in with articulatory-feature and phonetic analysis, and that also has implications for automatic segmentation

The Unstructured Acoustic Approach

Nick Campbell and Parham Mokhtari of ATR (Japan) are using an acoustically driven approach to annotate corpora for concantenative synthesis

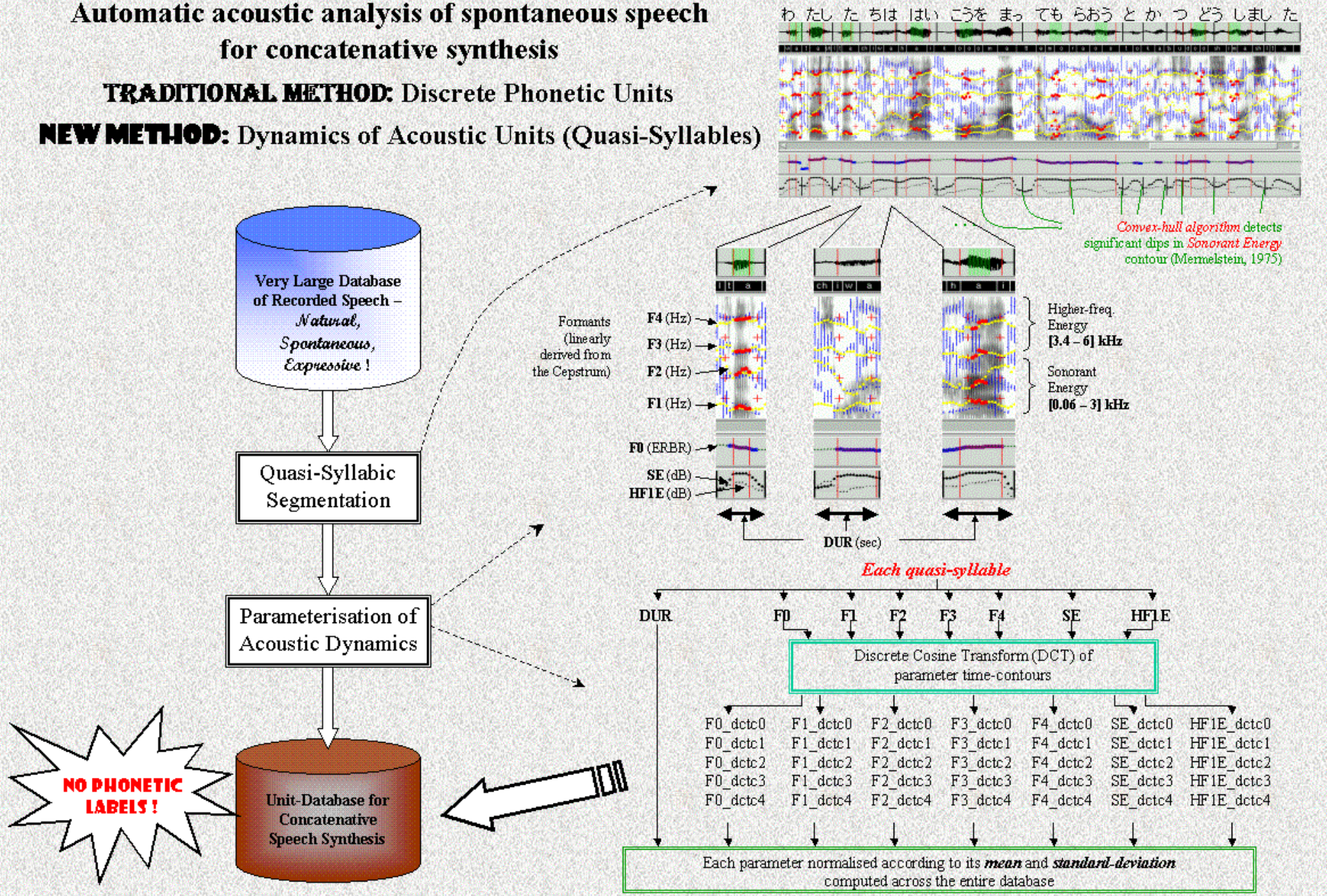
Juergen Schroeter will discuss synthesis in detail in the next talk

Here, I briefly mention the ATR approach as one alternative for analyzing the speech signal

Automatic acoustic analysis of spontaneous speech for concatenative synthesis

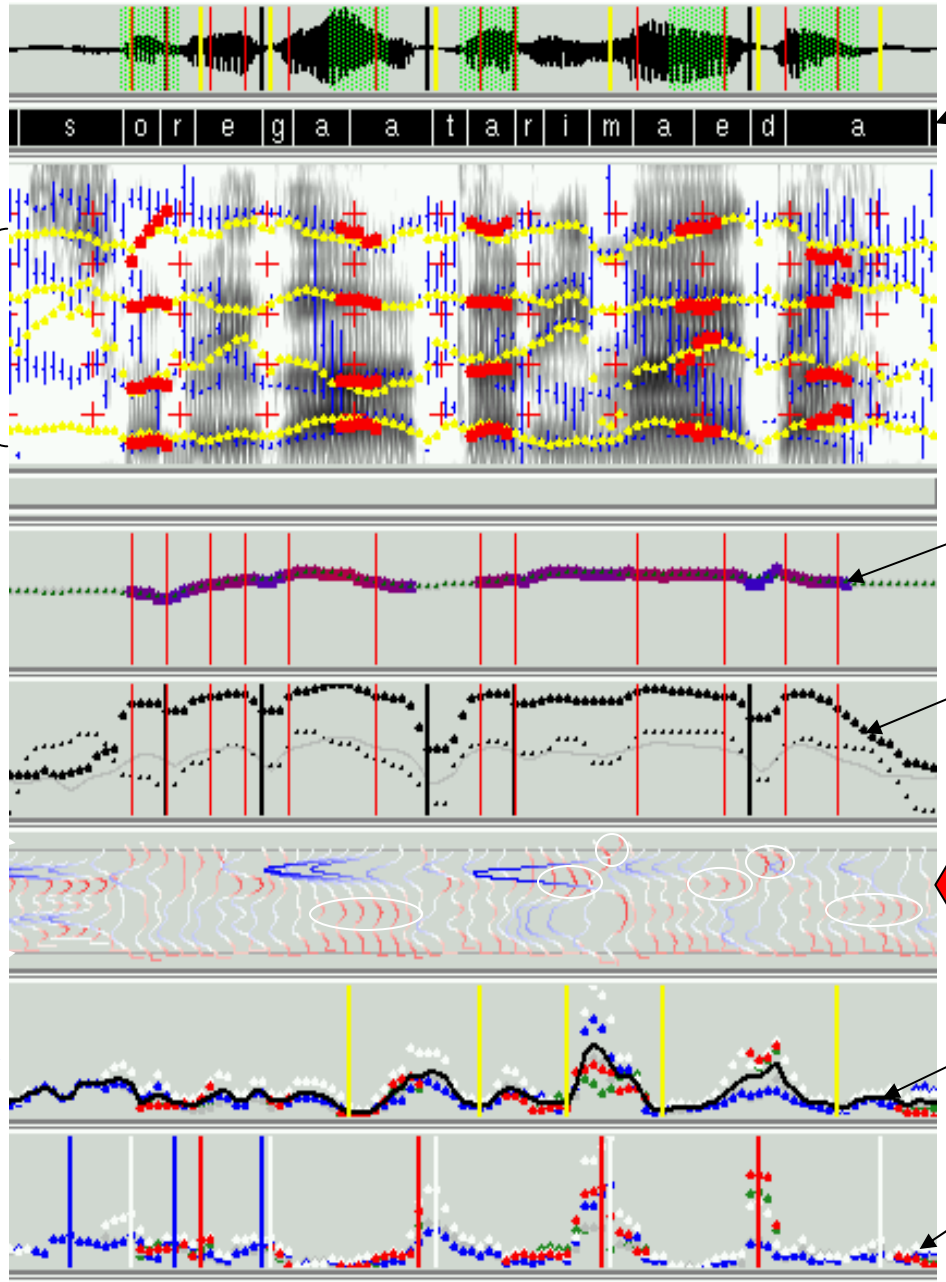
TRADITIONAL METHOD: Discrete Phonetic Units

NEW METHOD: Dynamics of Acoustic Units (Quasi-Syllables)



Material provided courtesy of Parham Mokhtari and Nick Campbell (ATR, Japan)

Quasi-Articulatory-Gestural Units from Continuous Speech: Acoustics → VT Area-Functions → SSs & Transitions



Contours of **formants** F1, F2, F3 and F4 estimated by linear transformation of the cepstrum as proposed by **Broad & Clermont (1989)**.

lips (13.1cm)

glottis (0cm)

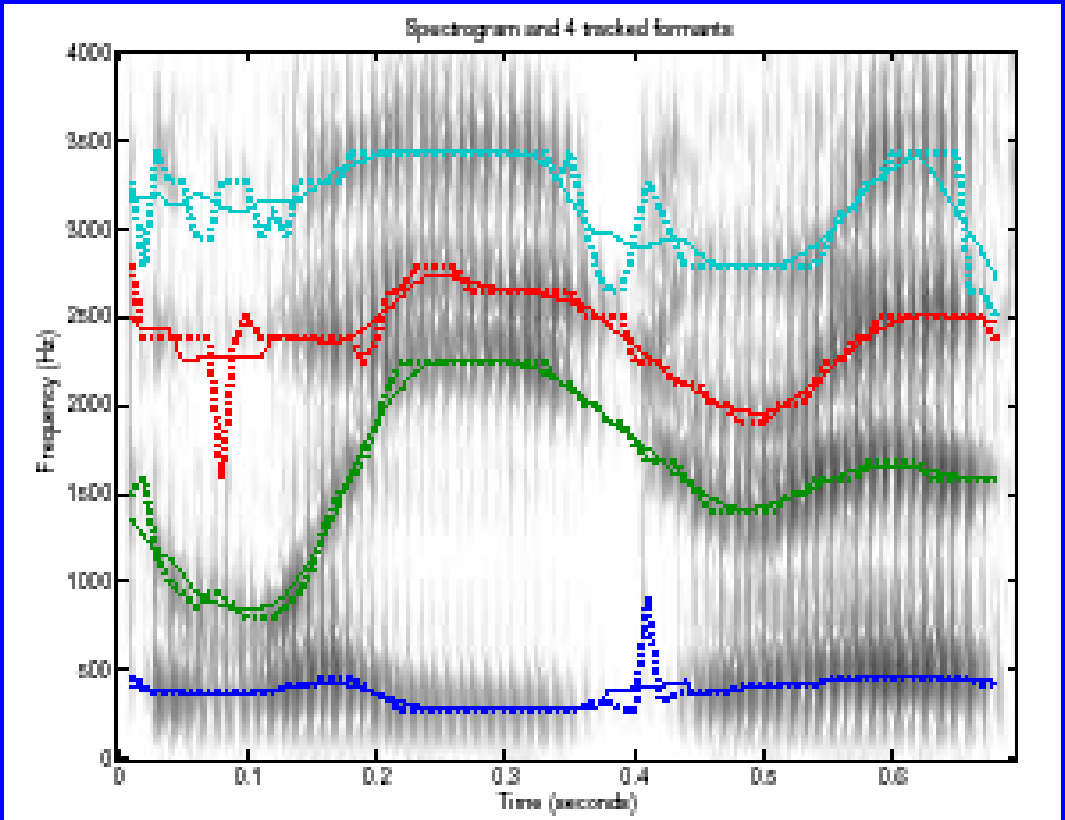
Material provided courtesy of Parham Mokhtari and Nick Campbell (ATR, Japan)

Formant-Tracking Approach

A related, but somewhat different approach is taken by Deng and colleagues

Formant patterns in the speech signal are automatically tracked as a means of restricting the acoustic model space, as shown below

In this first example, no explicit targets are provided, which results in certain errors in tracking, particularly during intervals of spectral transition and low amplitude

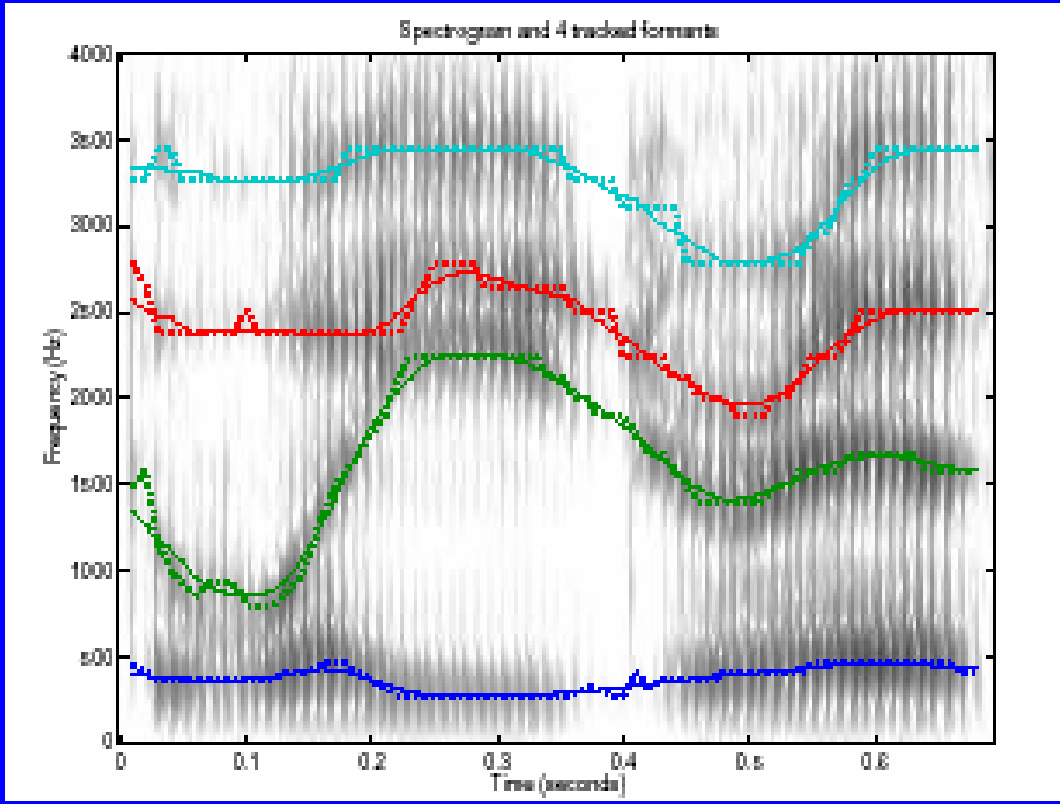


Deng et al. (2003)

Formant Tracking Approach

In the second example, explicit targets are provided

These targets improve the performance of the tracking algorithm considerably



Deng et al. (2003)

Drawbacks of the Acoustic Approach for ASR

The acoustics-driven approach does not attempt to relate the details of the acoustic signal to a more abstract representation

For this reason, it may be of limited utility for recognition (by itself)

Some form of linguistic structure would be useful for integrating the lower acoustic-phonetic tiers and the higher levels associated with words (and meaning)

Lack of explicit structure makes detailed error analysis difficult (see final section of this presentation) and therefore is not amenable to a scientific study of recognition systems

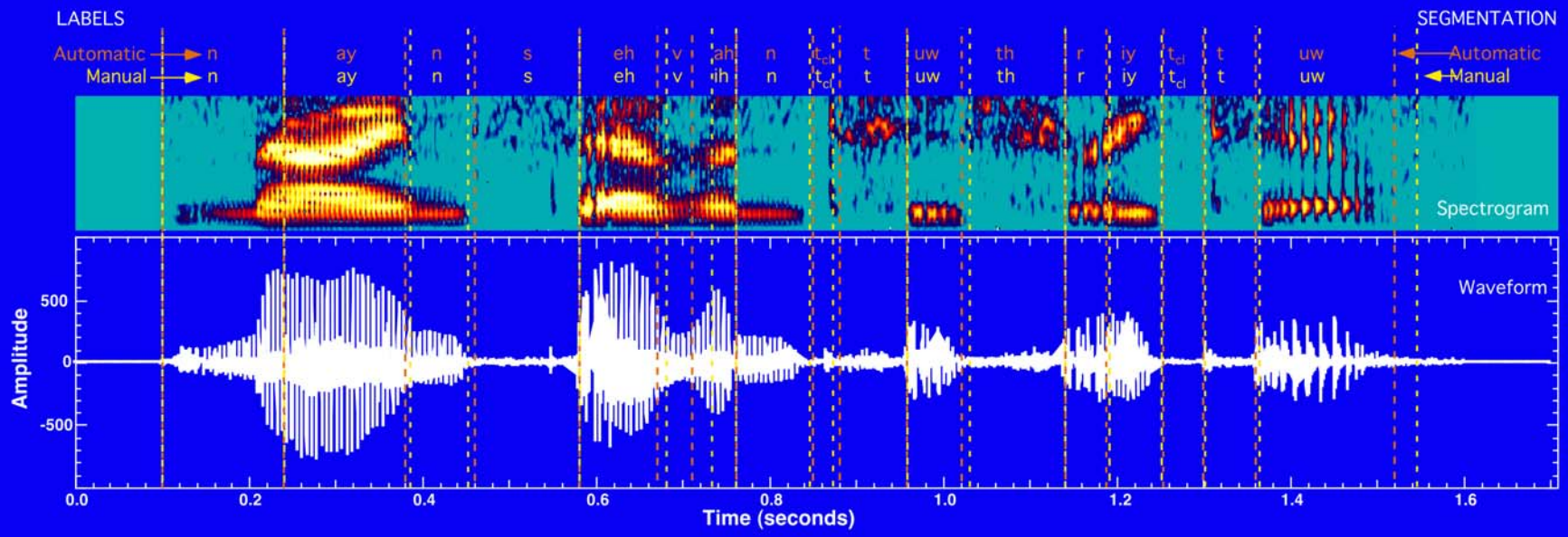
Automatic Segmentation – Phones

Currently, the lexical units in ASR systems are typically composed of phonemic elements

It would be useful to partition the speech signal automatically at this level WITHOUT RECOURSE TO A WORD TRANSCRIPT

Is this possible? Yes, for limited vocabulary tasks, such as OGI Numbers, as shown below for the ALPS system (using neural networks)

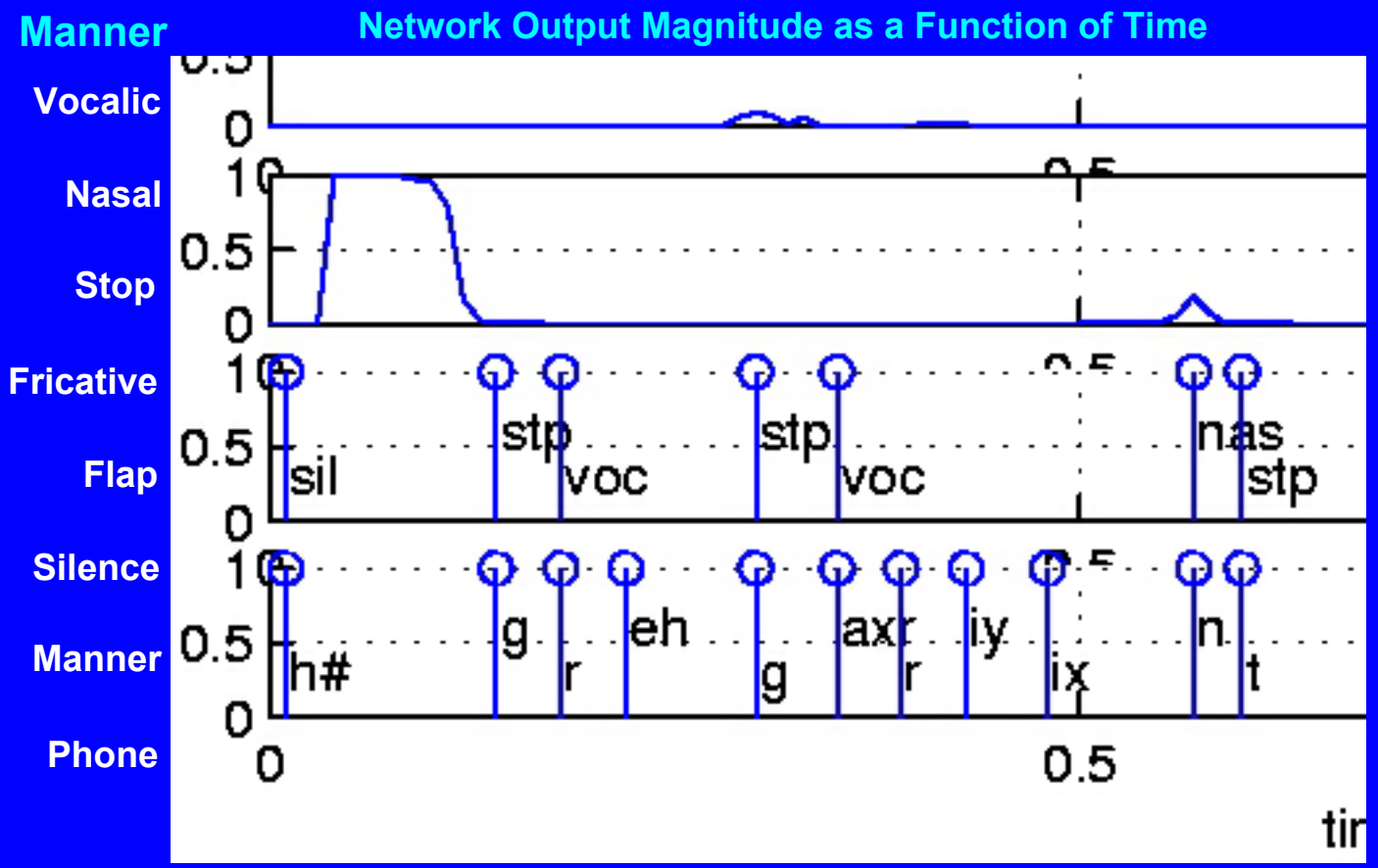
Average disparity between manual & automatic phone boundaries is 11 ms and the phonetic segment concordance is 83%



Automatic Segmentation – Manner

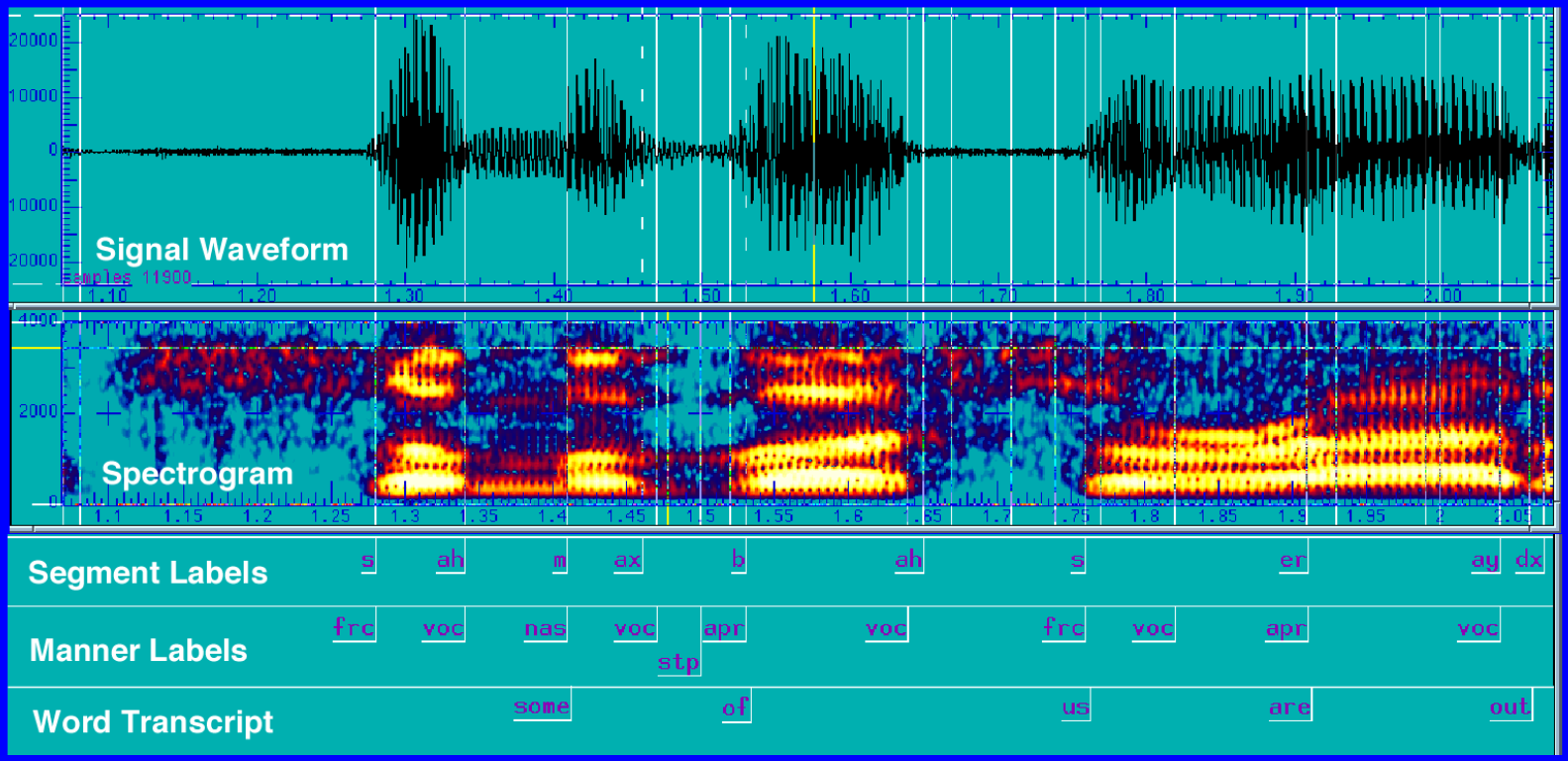
For more casual speech and larger-vocabulary tasks(e.g., Switchboard) a somewhat different approach may be warranted

In such instances, we may wish to perform an initial manner-of-articulation classification on the raw speech signal, using neural networks, as shown below



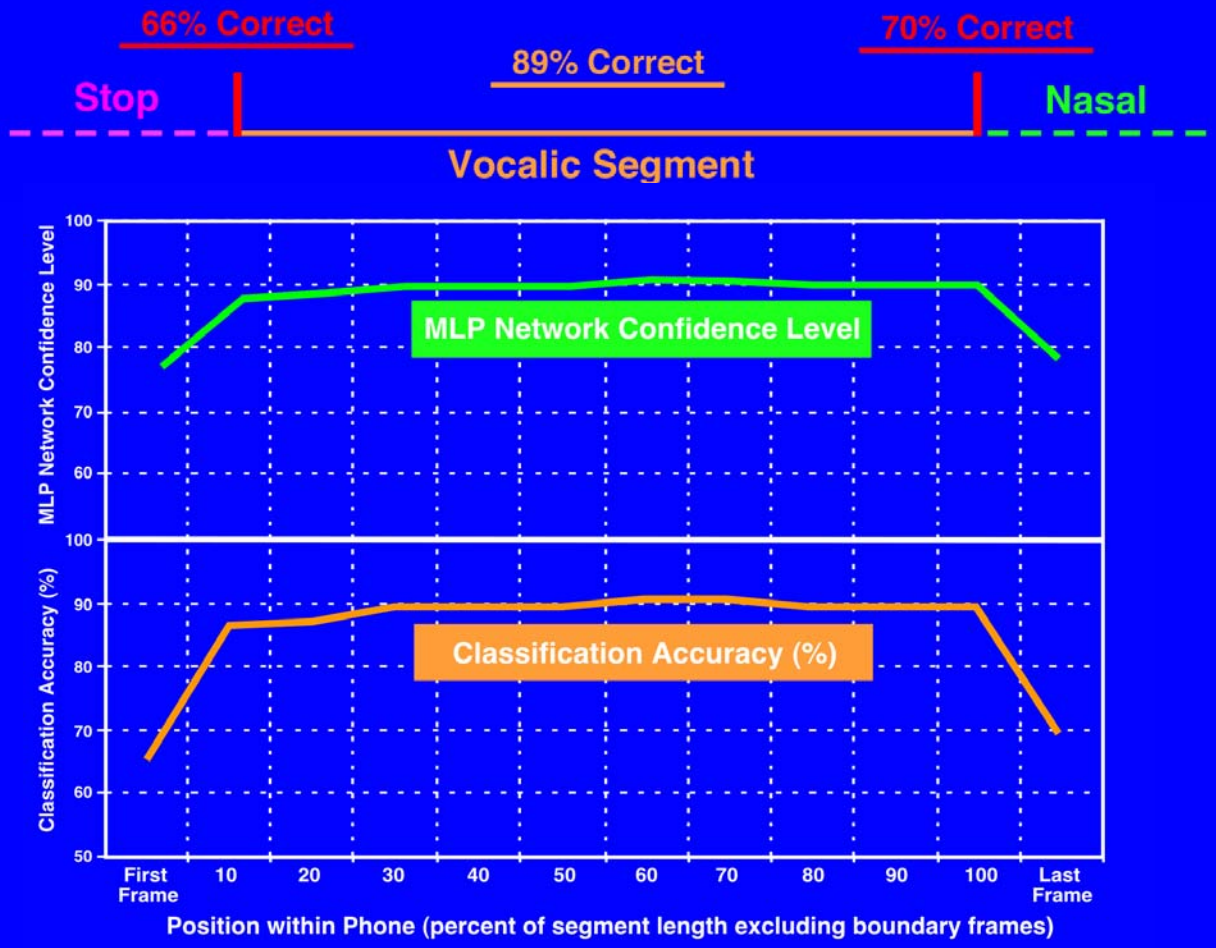
Manner of Articulation Isomorphic with Phones

The manner segmentation (fricative, vocalic, nasal, etc.) is temporally isomorphic with phonetic segments, as shown below



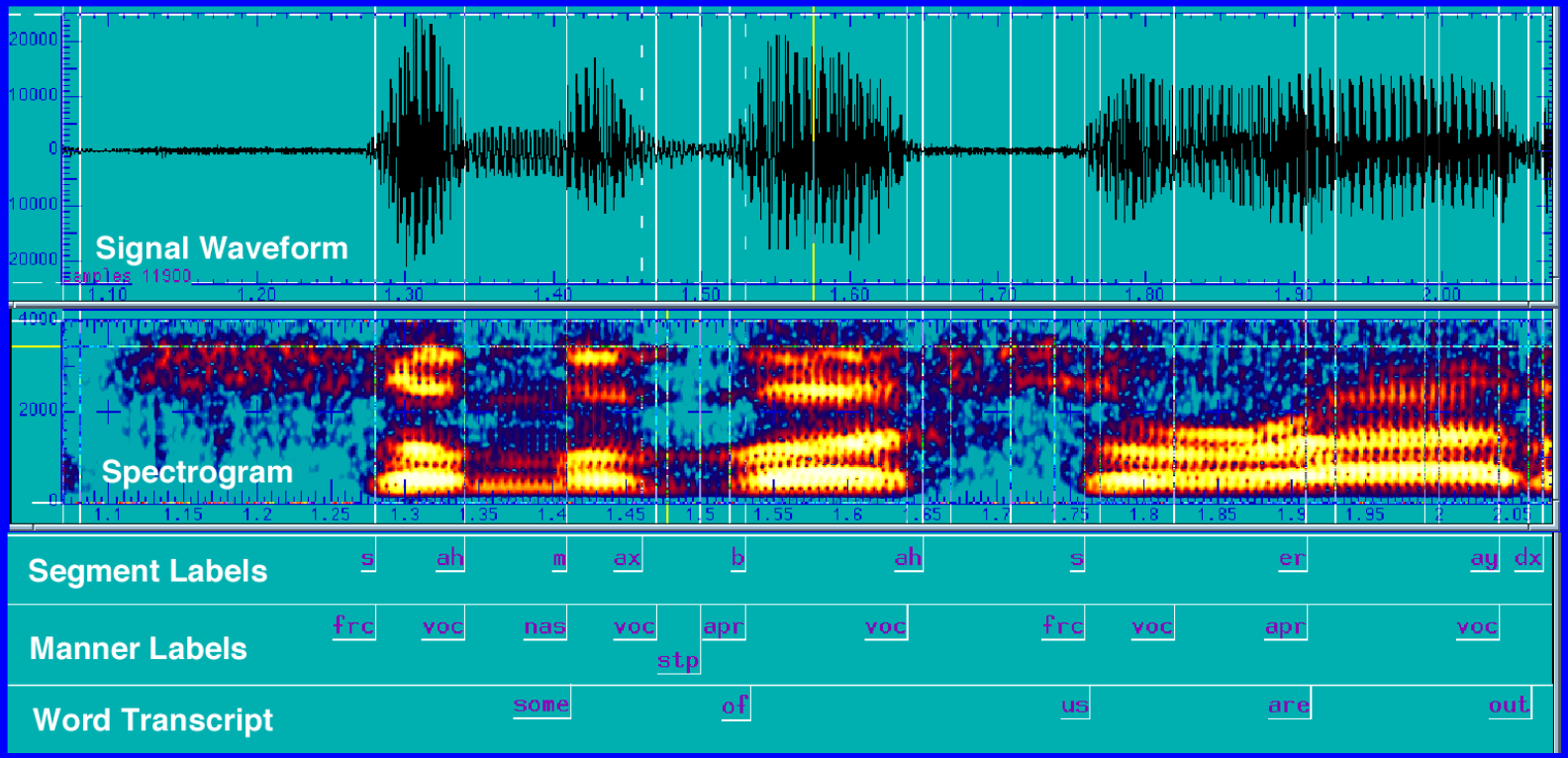
Using Manner to Spot Syllable Nuclei

One can exploit this observation using neural networks (MLPs) to implicitly segment the speech signal at the phone level by utilizing the network confidence level as an indirect indicator of the phone boundaries



Automatic Phone Segmentation with Manner

To obtain the implicit phone segmentation shown below using manner of articulation boundaries as exemplified on the previous slide



We shall return to the issue of articulatory feature classification later in this presentation

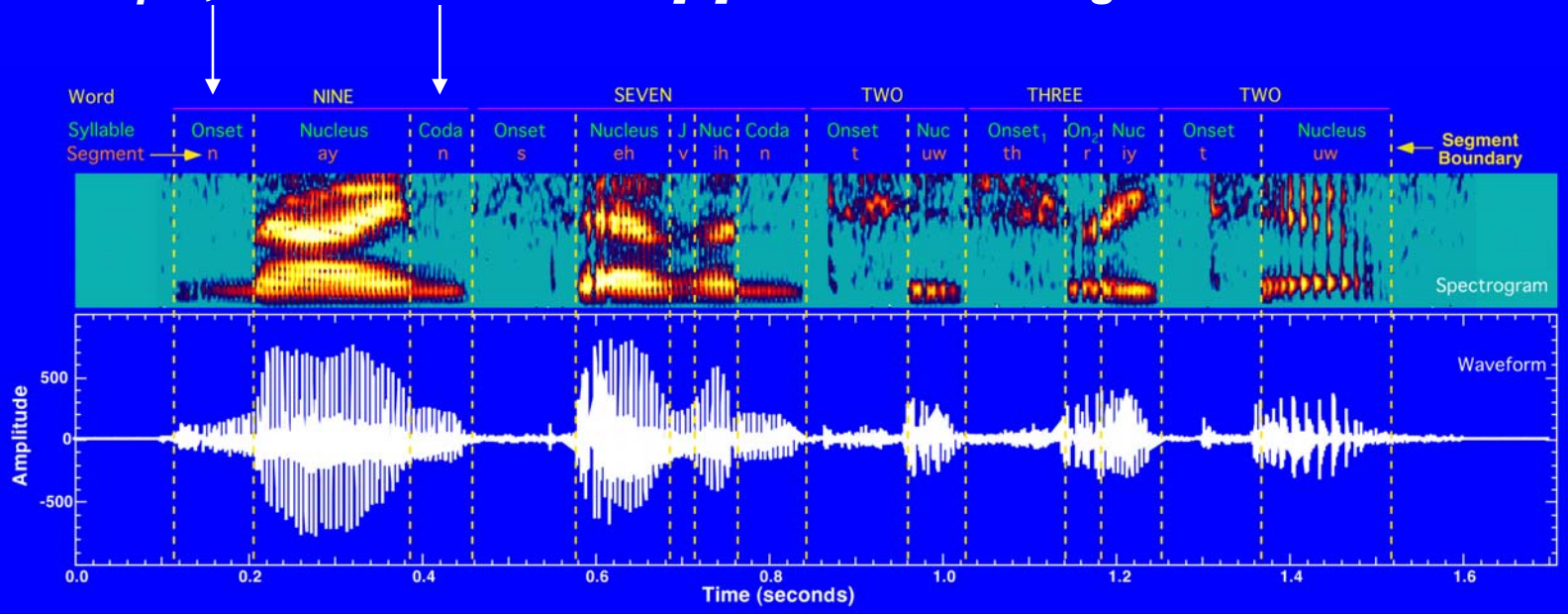
Syllable Segmentation

Words are, of course, composed of more than phones

Syllables are an important linguistic unit, and can be characterized in terms of three basic constituents – ONSET, NUCLEUS and CODA, as shown below

Syllables are more closely tied to the production process than phones and reflect articulatory gestures in a more transparent manner

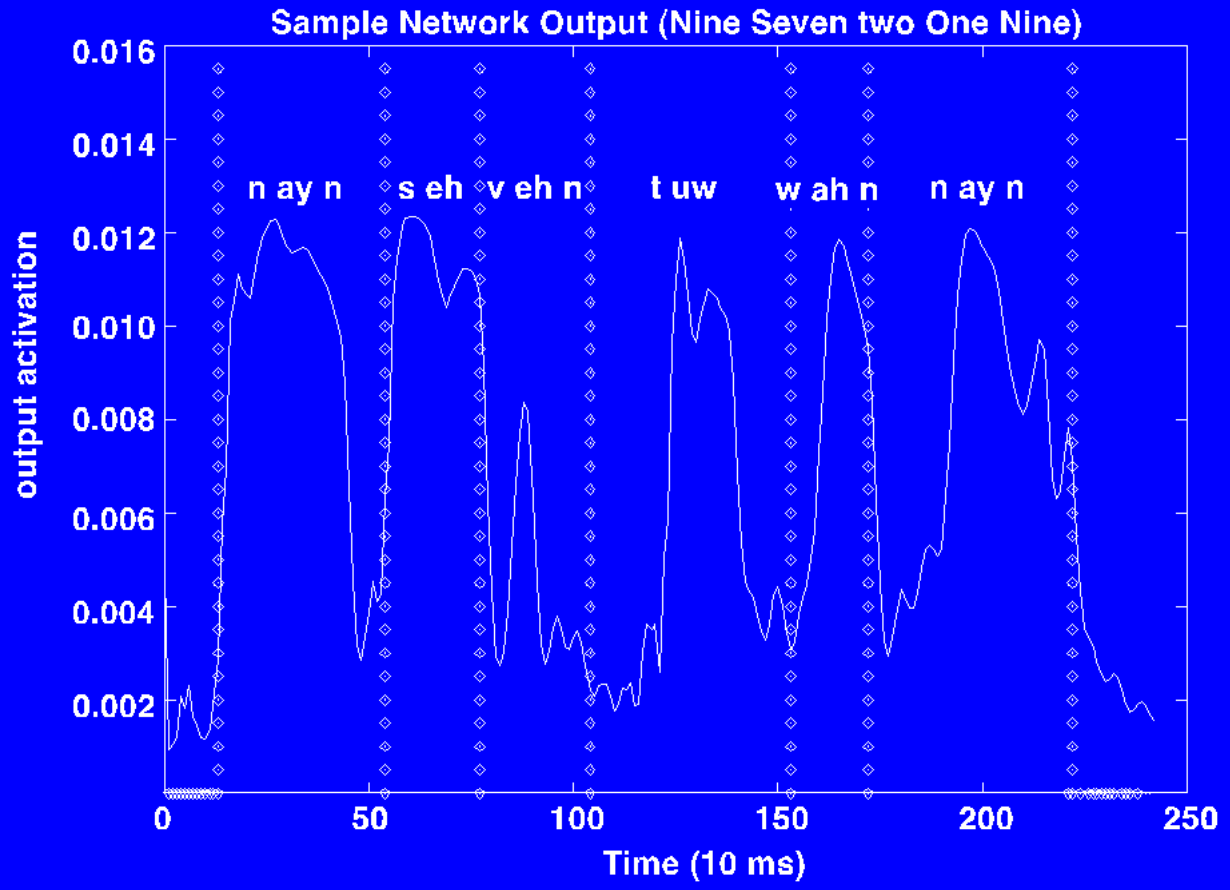
For example, notice that the first [n] in "Nine" is longer than the second



OGI Numbers (telephone) corpus

Syllable Segmentation – Network Based

It is possible to automatically segment the speech signal into syllable units using neural networks (in this instance, Temporal Flow Model networks), as shown below (the diamond lines indicate the estimated boundaries)



OGI Numbers (telephone) corpus

Syllable Segmentation – Signal Processing Based

It is also possible to automatically segment the speech signal into syllables using signal-processing methods. This has been done both by Michael Shire (while at ICSI) and by Hema Murthy and colleagues at ITT Madras, an example of which is shown below

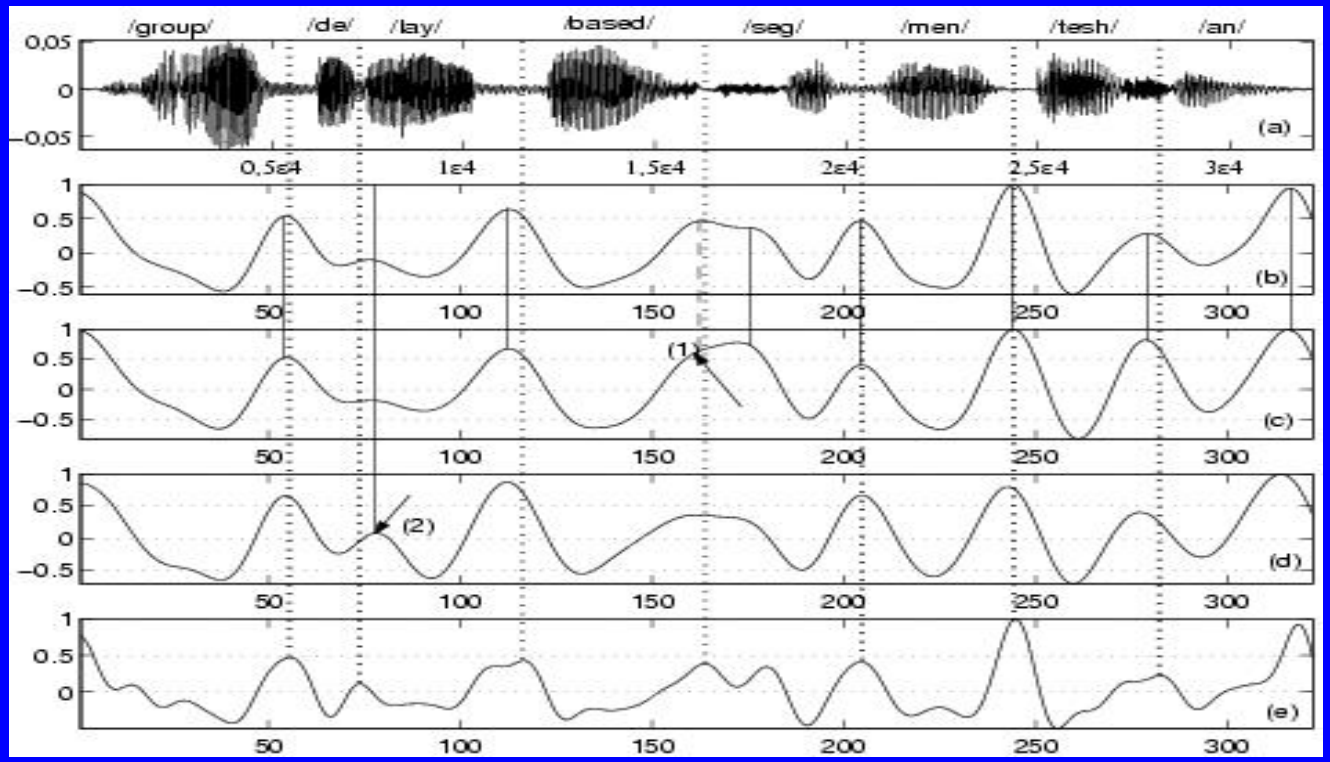
Group Delay Derived from

All Pass

LPF < 0.5 kHz

BPF 0.5 – 1.5 kHz

Hi Res, All Pass



Acoustic Waveform

From Syllables to Phonetic Classification

The syllable serves as a potentially useful unit with which to perform phonetic-segment and articulatory-acoustic feature classification

This is because many phonetic properties are organized at the level of the syllable

This is particularly the case for articulatory-acoustic features (AFs), which form the essential building blocks of the syllable's micro-structure

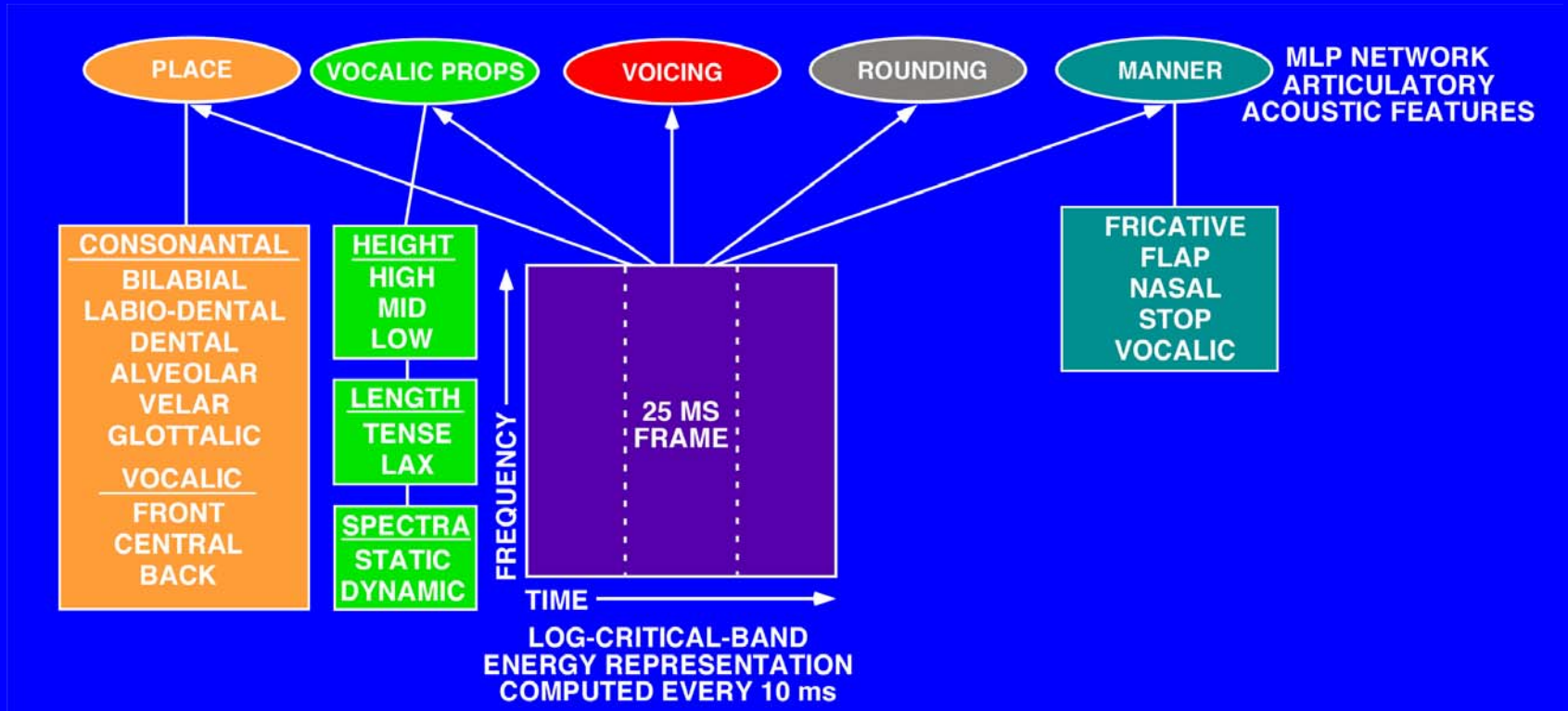
AFs can be used to delineate various constituents of the syllable, as well as characterize the nature of their interaction within a syllabic framework

Articulatory-Acoustic Feature Extraction

A singular advantage of AFs relative to phones is the small number of classes per dimension – instead of 40-60 phones, there are four to six articulatory dimensions, most with between 2 and 4 classes

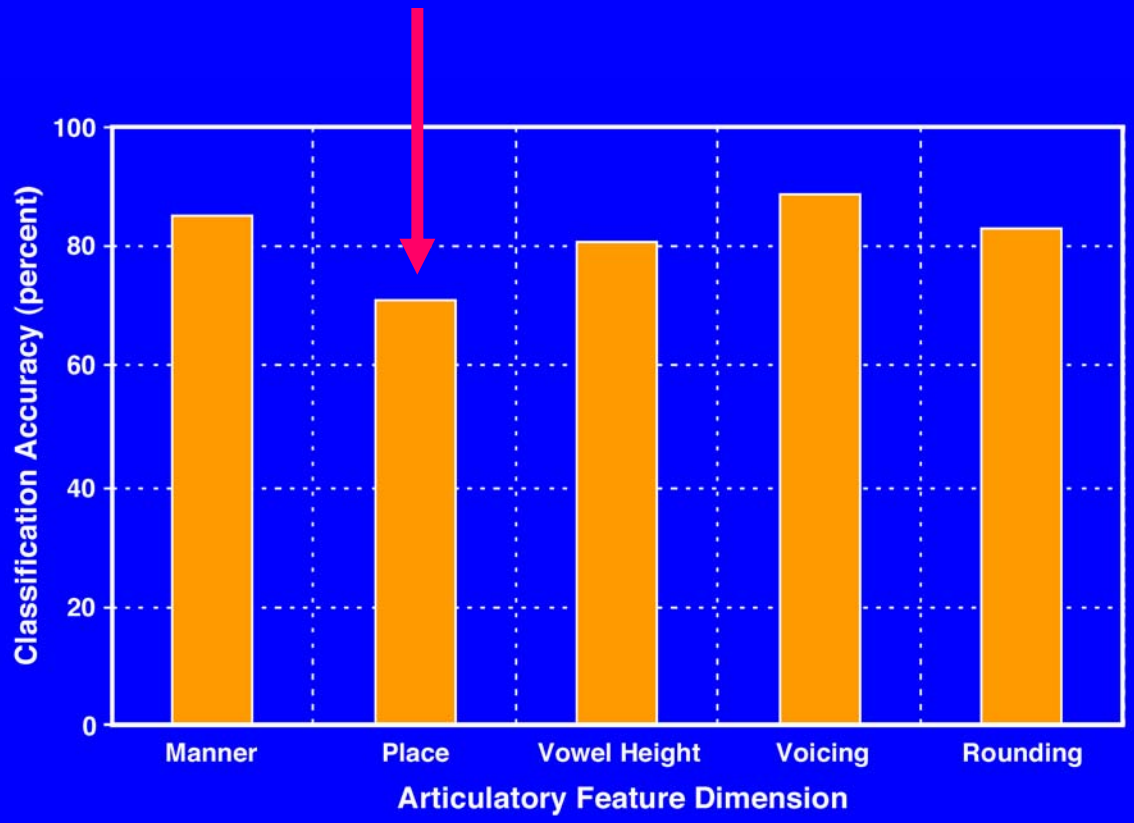
Multilayer Perceptron (MLP) Neural Network Classifiers

- ***Critical-band, log-compressed spectral representation***
- ***Single hidden layer of 200-400 units, trained with back-propagation***
- ***Nine frames of context used in the input***



Annotation of Articulatory Primitives

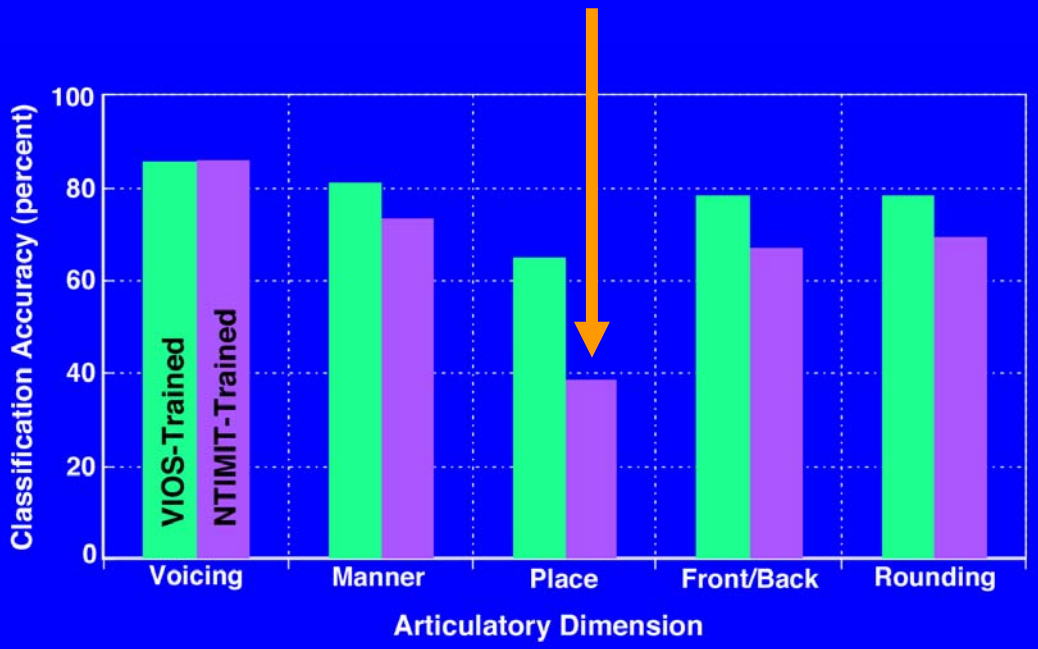
Using such classifiers it is possible to obtain good to excellent performance for all articulatory feature dimensions except place of articulation



Cross-Linguistic Transfer of Articulatory Features

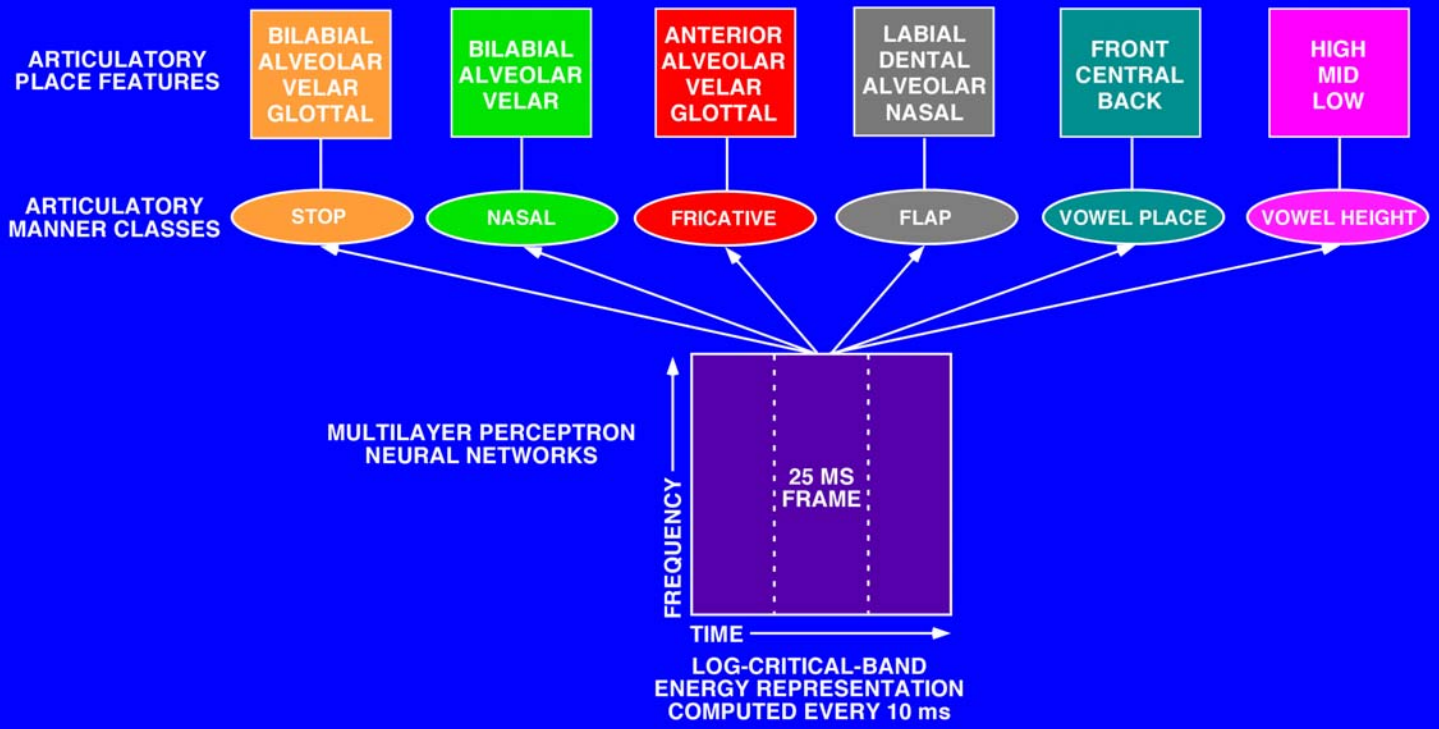
Articulatory feature classifiers trained on one language (or speaking style) may be transferable to other languages and speaking styles, as has been demonstrated for English and Dutch by Wester et al. (2001)

In that study, voicing and manner of articulation features trained on English transferred well to an independent Dutch corpus, as did other AF dimensions, such as Front/Back (vowels) and Rounding, – only the place-of-articulation dimension did not transfer well



Manner-Dependent Place of Articulation

Although there are 10 distinct places of articulation in English, there are typically only three per manner class

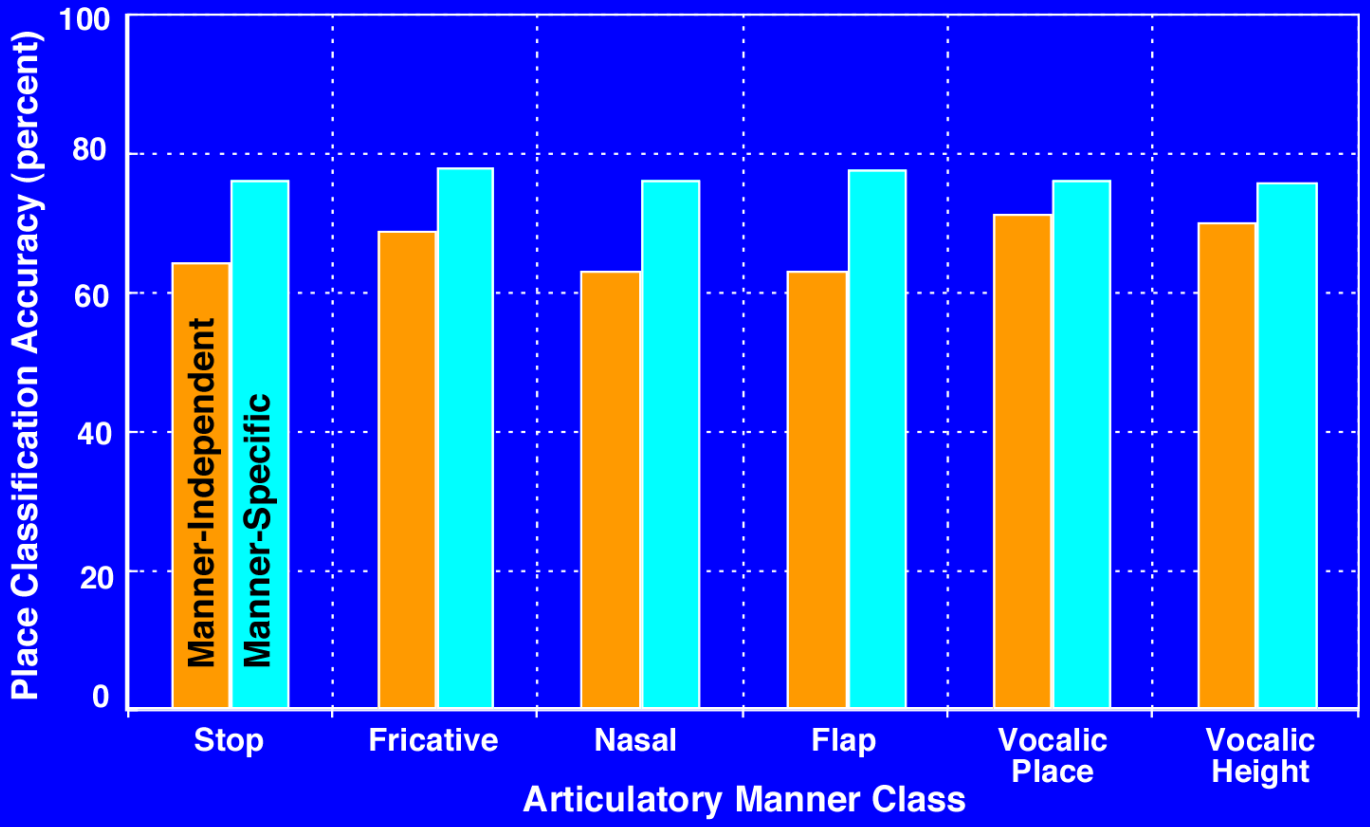


Chang et al., 2001

Manner-Specific Place of Articulation

This observation can be used to improve place classification

Using manner-specific place classifiers improves performance for all manner classes



Chang et al., 2001

Articulatory Features and Syllables

The SYLLABLE, rather than the PHONE, is the basic organizational unit of spoken language (certainly with respect to pronunciation variation)

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

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 – thus, syllable position may be 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 (more about this shortly)

The Energy Arc Illustrated

Syllables rise and fall in energy over the course of their duration

Vocalic nuclei are highest in amplitude

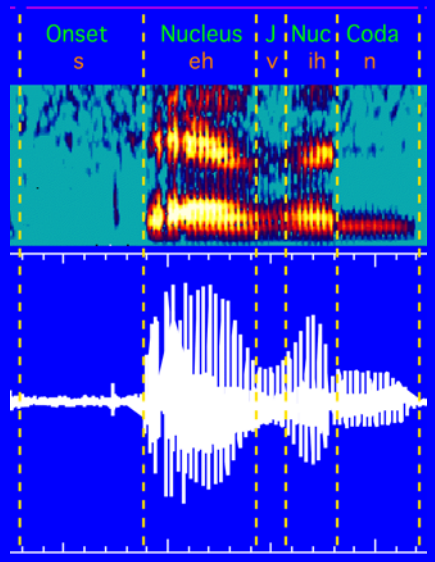
Onset consonants gradually rise in energy arching towards the peak

Coda consonants decline in amplitude more abruptly than onsets

The energy arc can account for the sequential order of segments within a syllable (organized by manner of articulation)

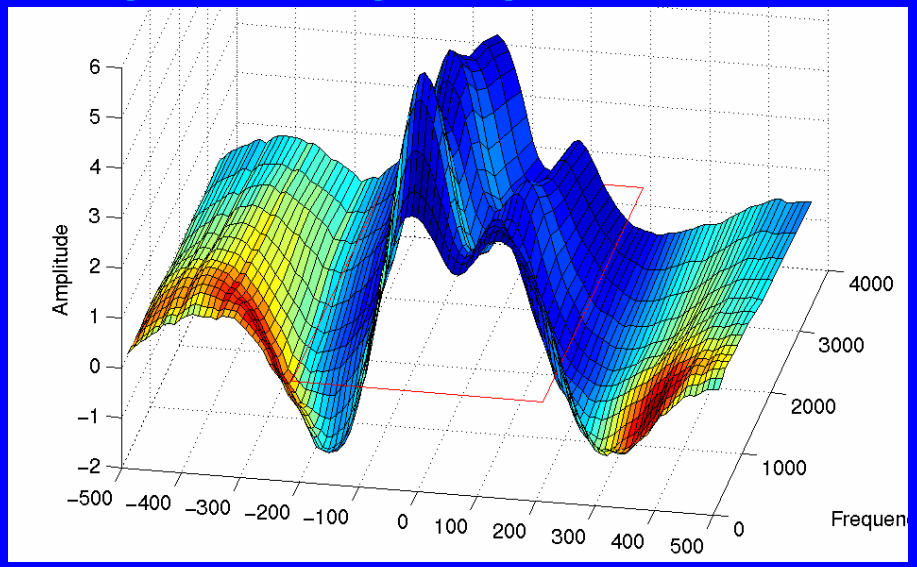
And also pertains to the low-frequency (2-20 Hz) modulation spectrum

Spectrogram + Waveform



"seven"

Spectro-temporal profile (STeP)



Articulatory Features and Syllables

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

VOICING spreads 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 prosodic prominence magnitude – 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 prosodic pattern of an utterance reflects the information contained within the utterance

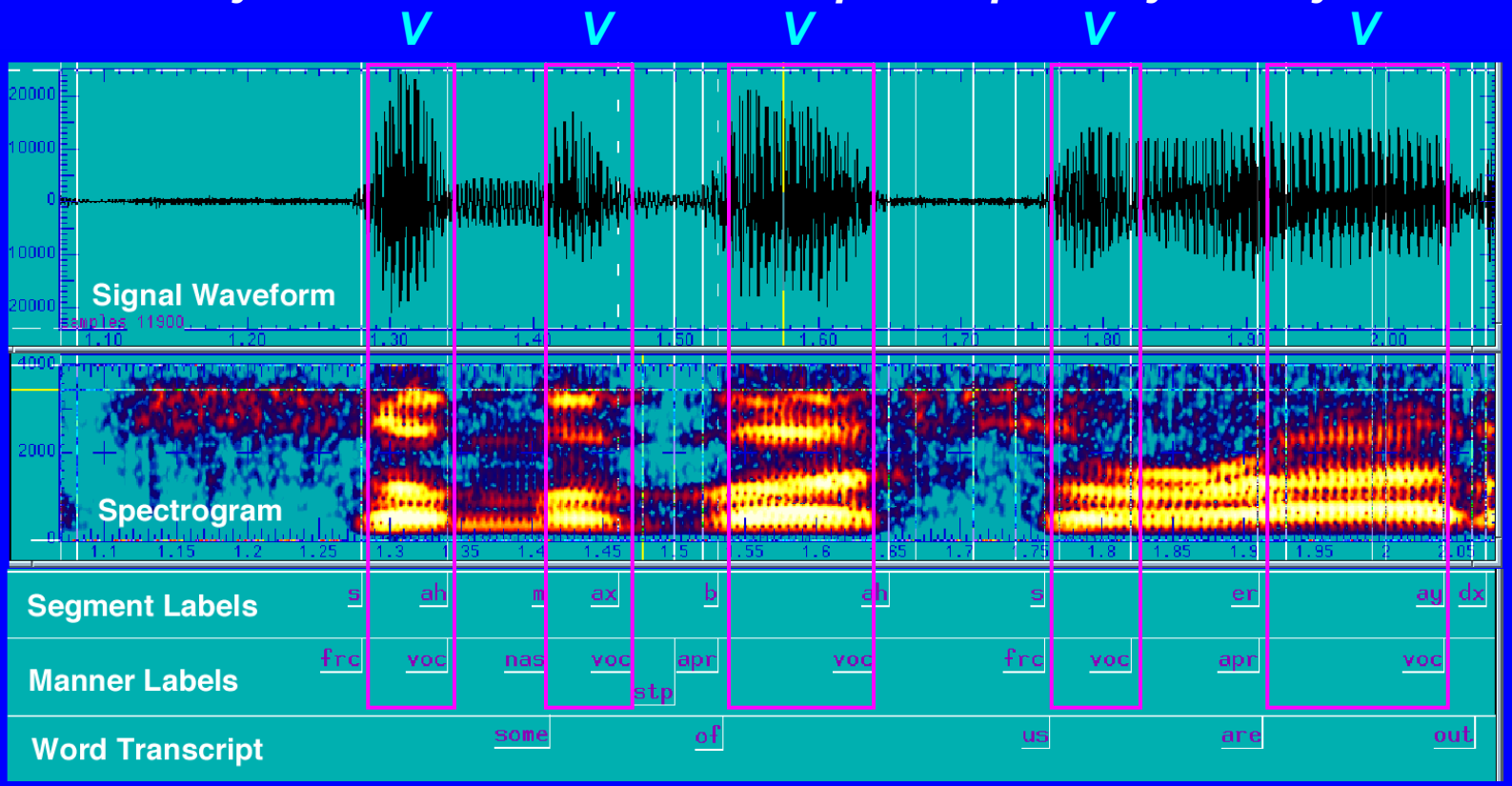
Therefore, it is ultimately INFORMATION (and lexical discriminability) that governs the detailed phonetic properties observed in an utterance

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 – we shall return to this topic via prosody shortly

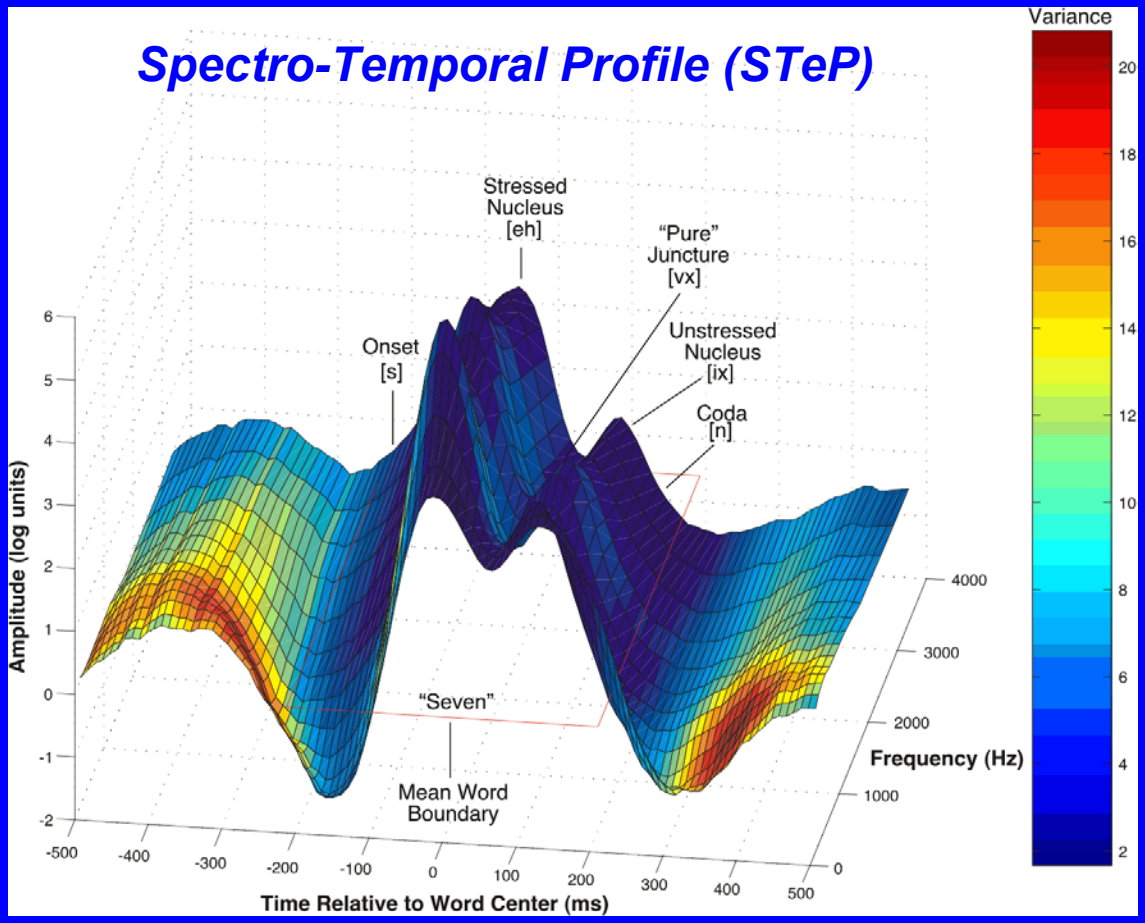


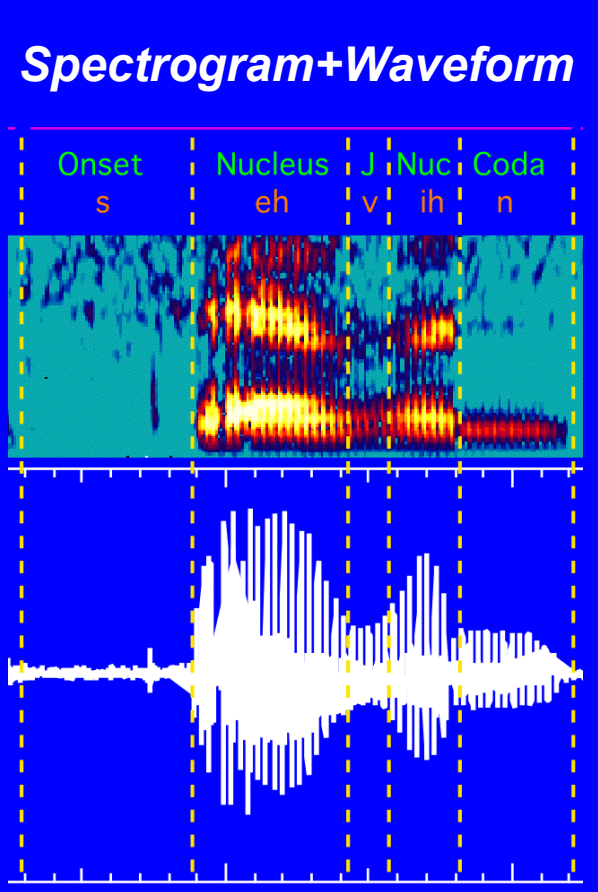
From Syllable Nucleus to Prosody

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"

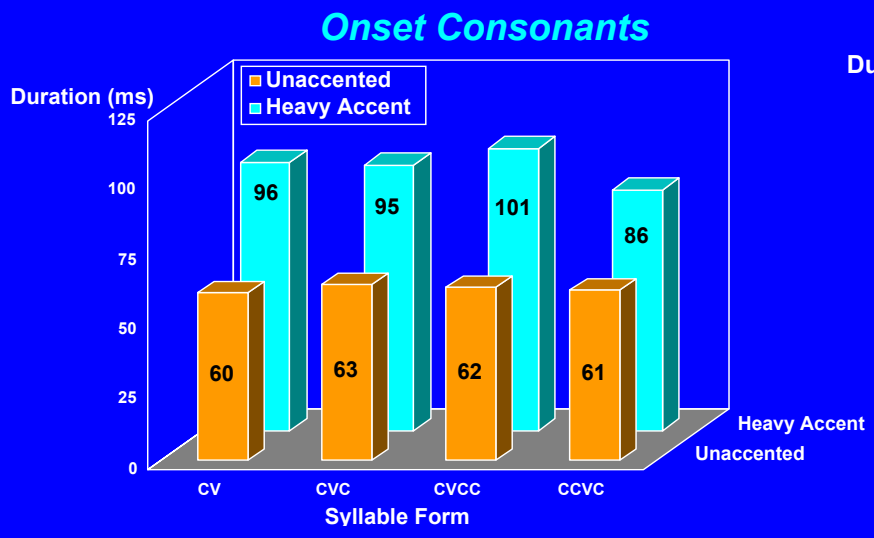


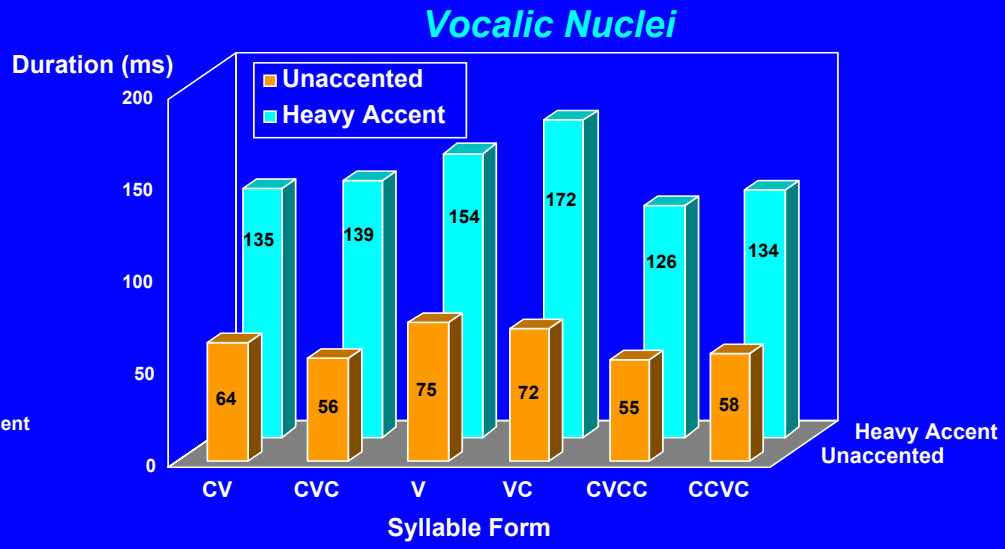


Prosody's Importance – Duration

The prominence of the syllabic nucleus affects many phonetic properties of the syllable, including:

The duration of nucleic and onset segments:

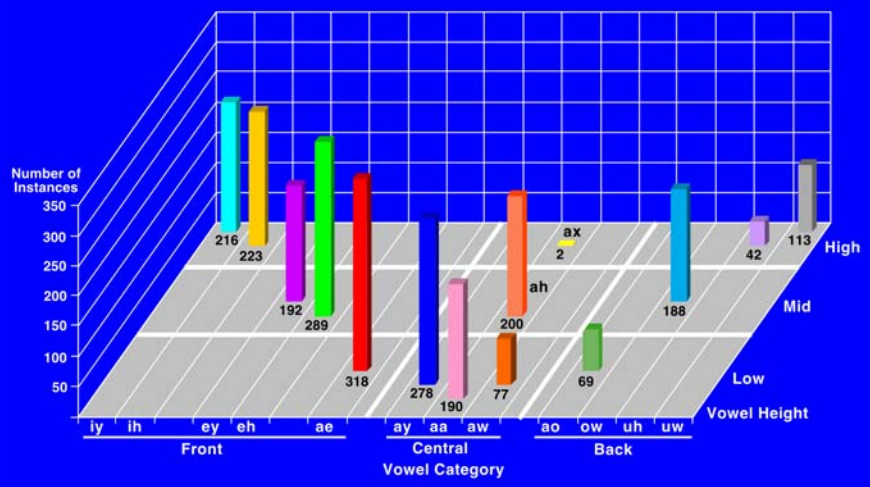




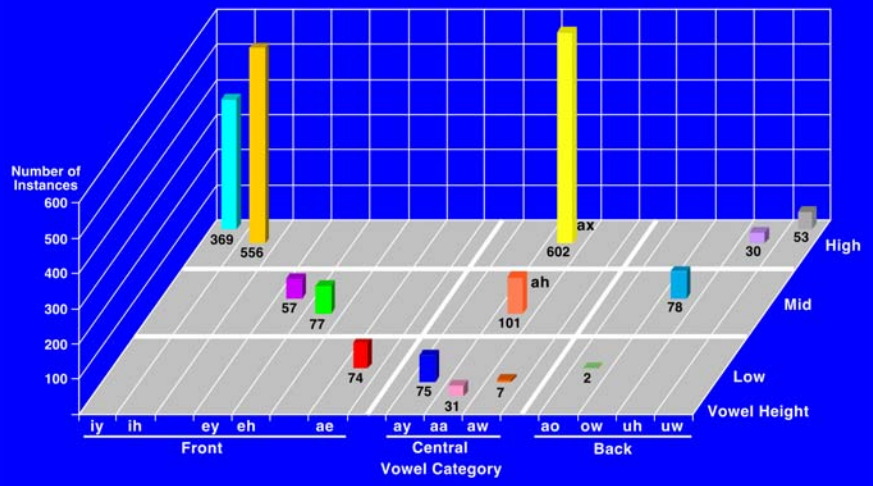
Prosody's Importance – Vocalic Identity

As well as the specific identity and articulatory configuration of vocalic segments

Heavily Accented



Unaccented



There is a relatively even distribution of vowels across the articulatory space in heavily stressed syllables (in English)

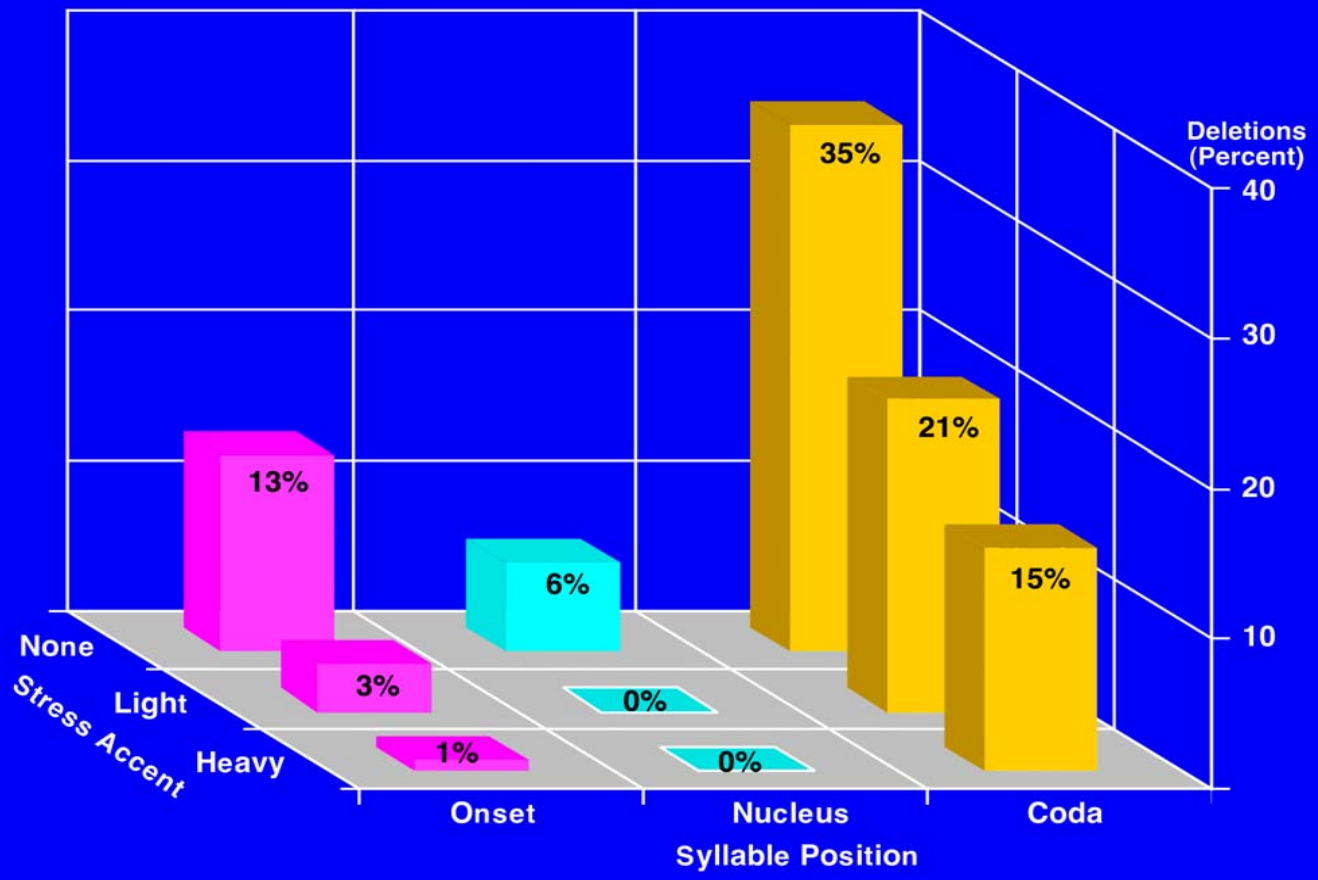
However, in unstressed syllables vowels consist mostly of [ih], [iy] and [ax]

In this sense, the vowel system appears inextricably linked to stress-accent

Vowels may not function in the same way as consonants, a potentially important observation for ASR systems

Prosody's Importance – Coda Deletions

The probability of coda consonant deletion is much higher in unstressed syllables relative to heavily stressed ones



Prosody's Importance – Coda Deletions

Most of the deletions are concentrated among three segments: [t], [d], [n] (where the disparity between white (Canonical) and orange (Transcribed) yields the number of deletions

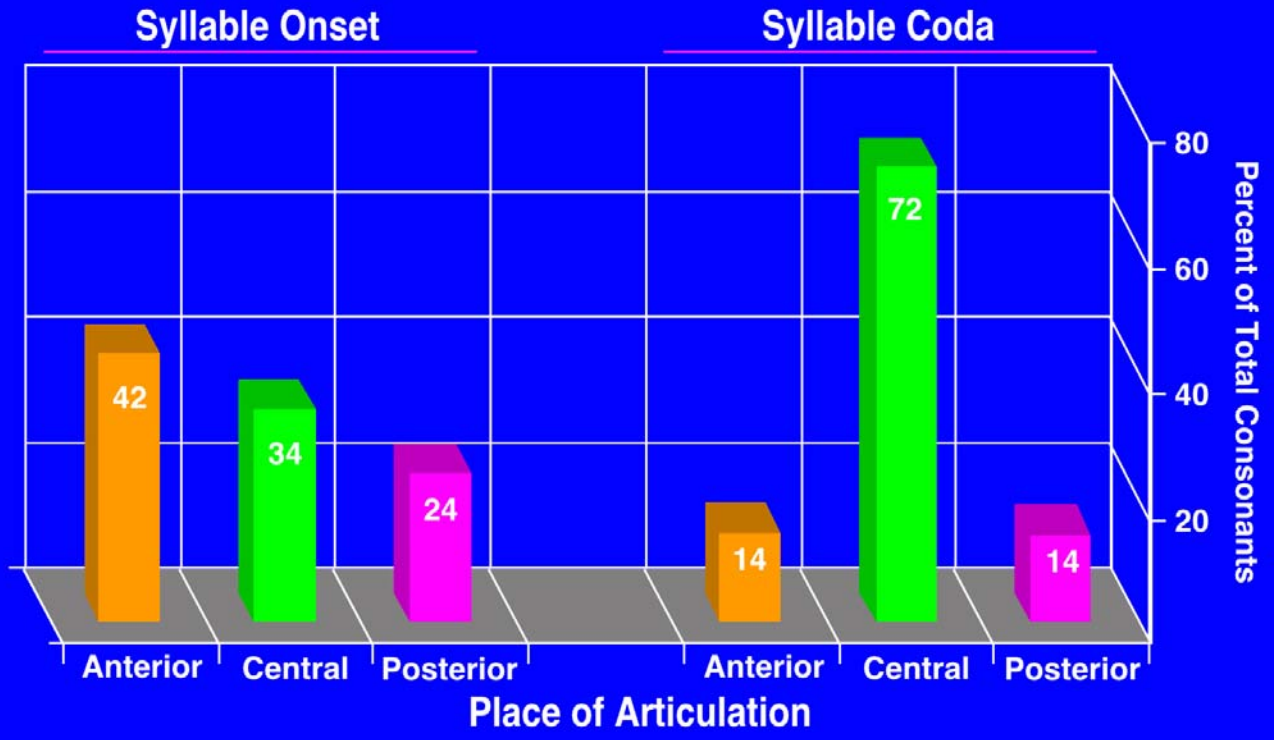
Accent	Heavy		Light		None		Total	
Segment	Can	Trans	Can	Trans	Can	Trans	Can	Trans
t	322	126	575	191	562	172	1459	489
d	200	119	295	127	370	96	865	342
n	311	237	498	381	773	542	1582	1160
s	142	135	202	214	151	155	495	504
z	179	149	258	208	271	221	708	578

Coda Deletions – Place of Articulation

And where three quarters of the consonant codas are coronals (central articulation)

With a place-of-articulation distribution quite different from the onsets

Hence, the entropy associated with syllable constituents appears to interact with the stress-accent system and exert a profound impact on the phonetic realization of the speech signal



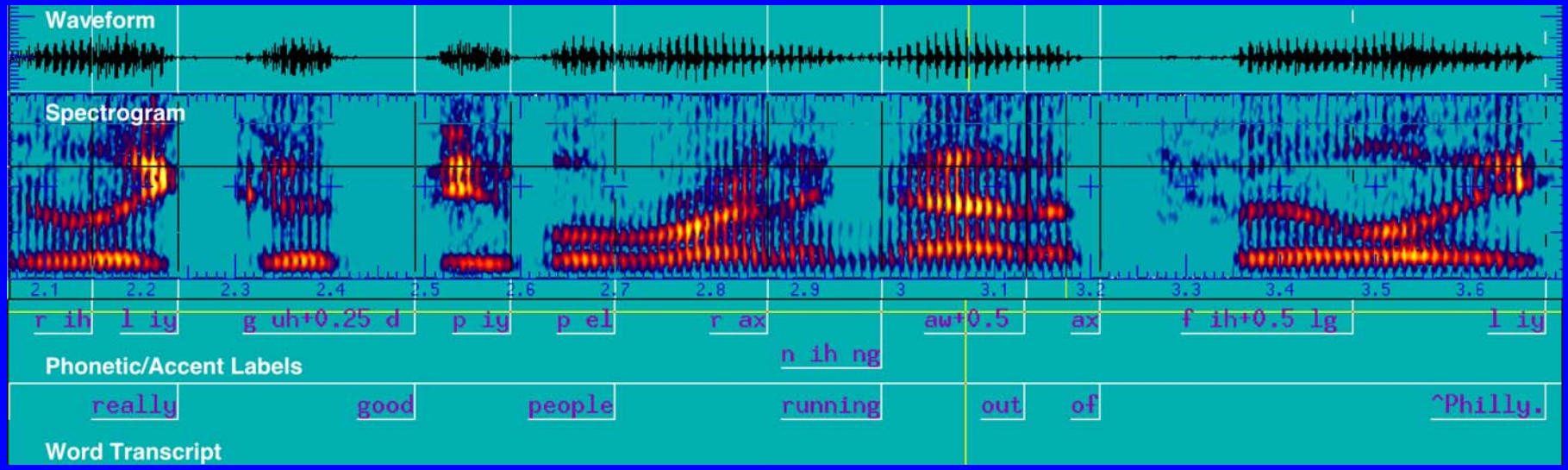
Automatic Annotation of Stress Accent

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))

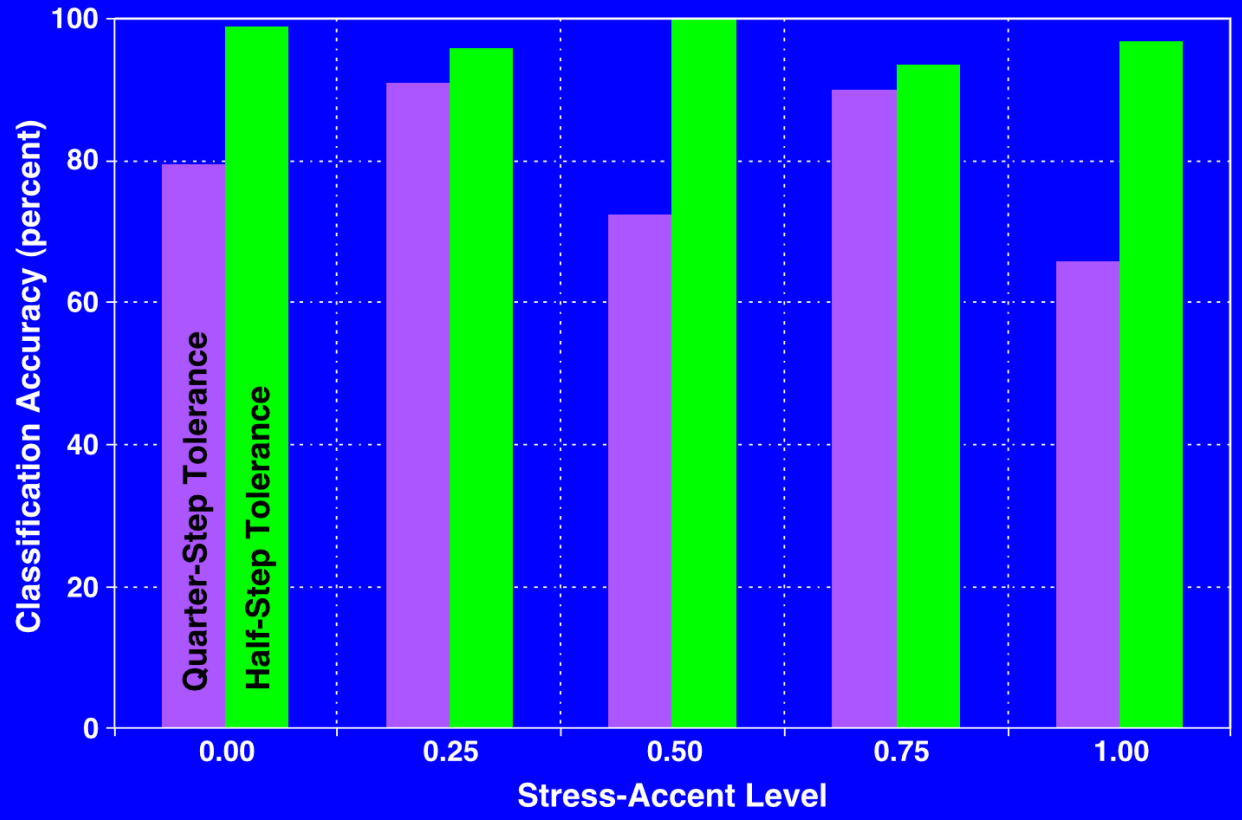


How Good is AutoSAL?

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

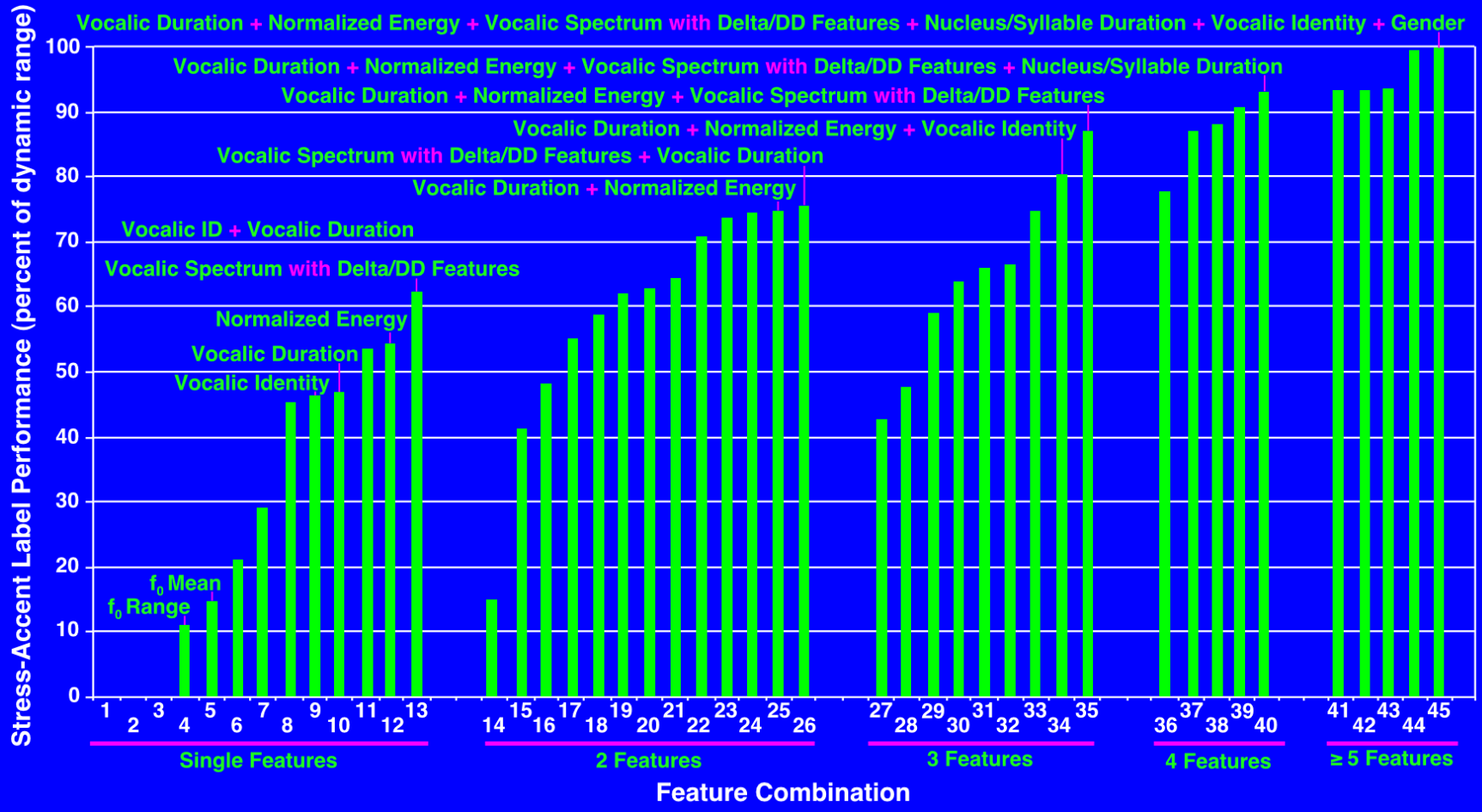


The Acoustic Bases of AutoSAL

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

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

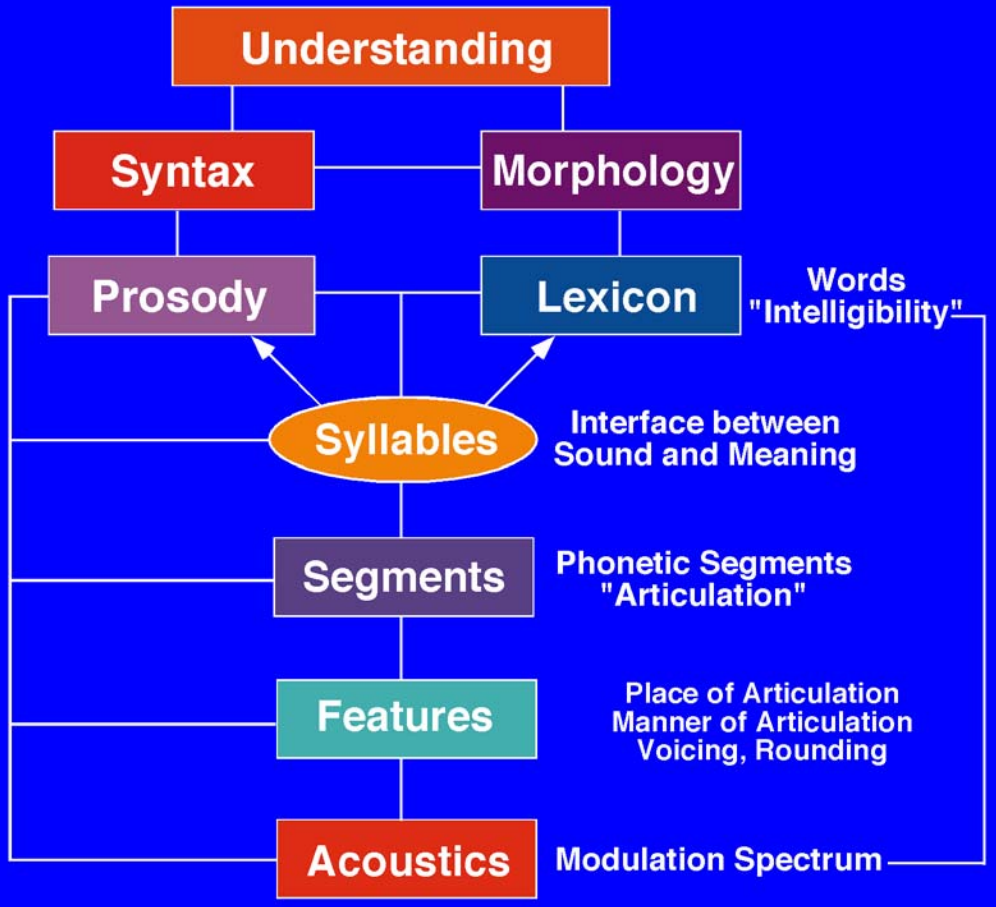
Pitch-related features are (relatively) unimportant for stress-accent labeling



Multi-Tier Representations

It is essential that we understand the relationship among linguistic levels in order to characterize the speech signal with parsimony and accuracy

One attempt at embedding such an integrated, multi-tier approach into ASR has been performed by Shawn Chang as part of his thesis at ICSI

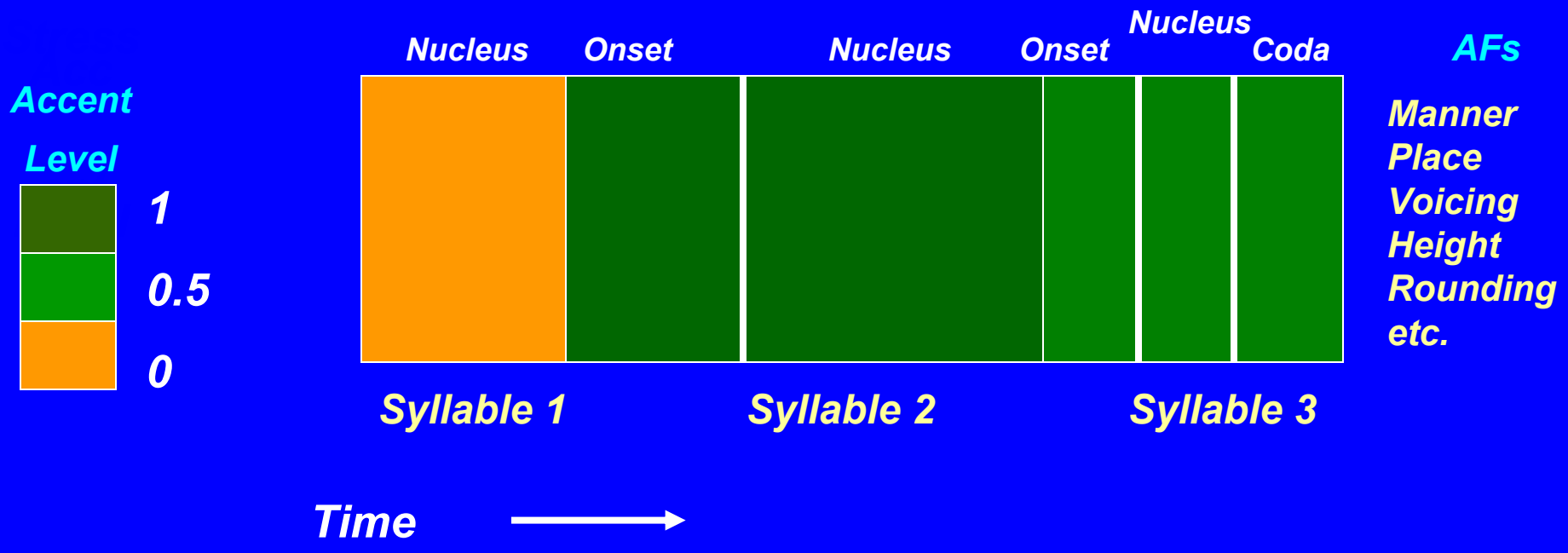


A Multi-Tier ASR System Illustrated

In this work, the speech signal is modeled as a sequence of syllables each with a variable amount of prominence

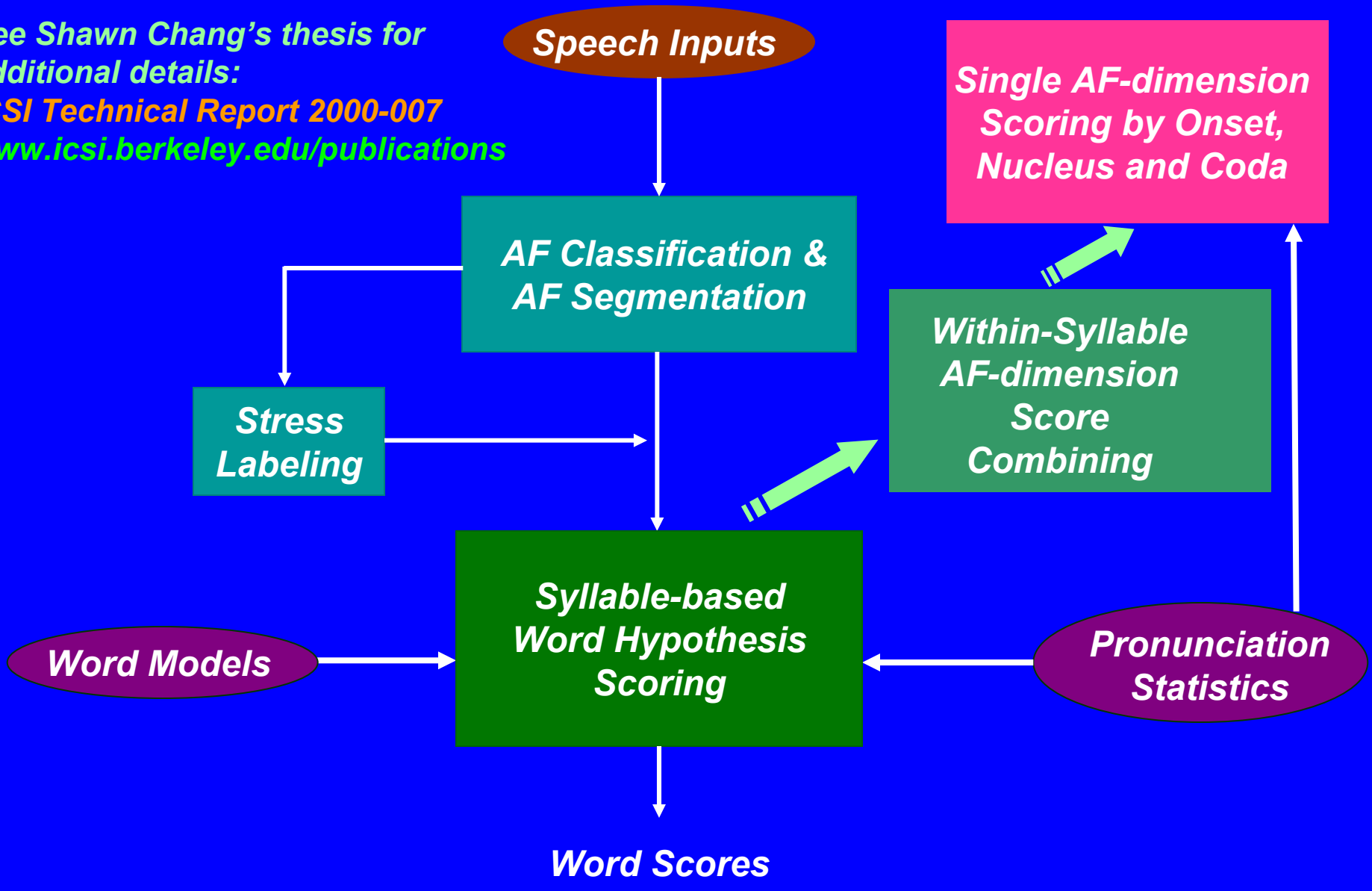
Each syllable consists of a vocalic nucleus, and optionally contains onset and coda elements

Each syllabic constituent is specified in terms of articulatory-acoustic features, most of which are inherently trans-segmental



Structure of the Multi-Tier System

See Shawn Chang's thesis for additional details:
ICSI Technical Report 2000-007
www.icsi.berkeley.edu/publications



Take Home Messages

The SYLLABLE, rather than the PHONE, is the basic organizational unit of spoken language – hence the difficulty for any phonetic orthography to accurately delineate pronunciation patterns in fine detail

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 – thus, syllable position may be 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 phone (and phonetic segment) and sets the basic intensity mode for the sequence of syllabic constituents ("ENERGY ARC")

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

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

Take Home Messages

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

VOICING spreads from the nuclei core of the syllable and spreads both forward (towards the coda) and backward (towards the onset), the degree of temporal spreading reflecting prosodic prominence magnitude – 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 their position within the syllable, as well as prosodic prominence

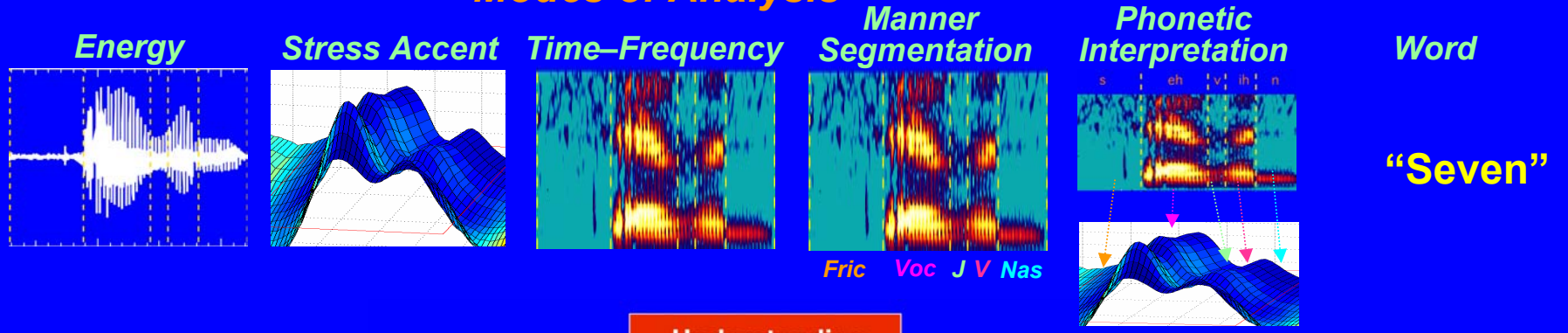
The prosodic pattern of an utterance reflects the information contained within the utterance

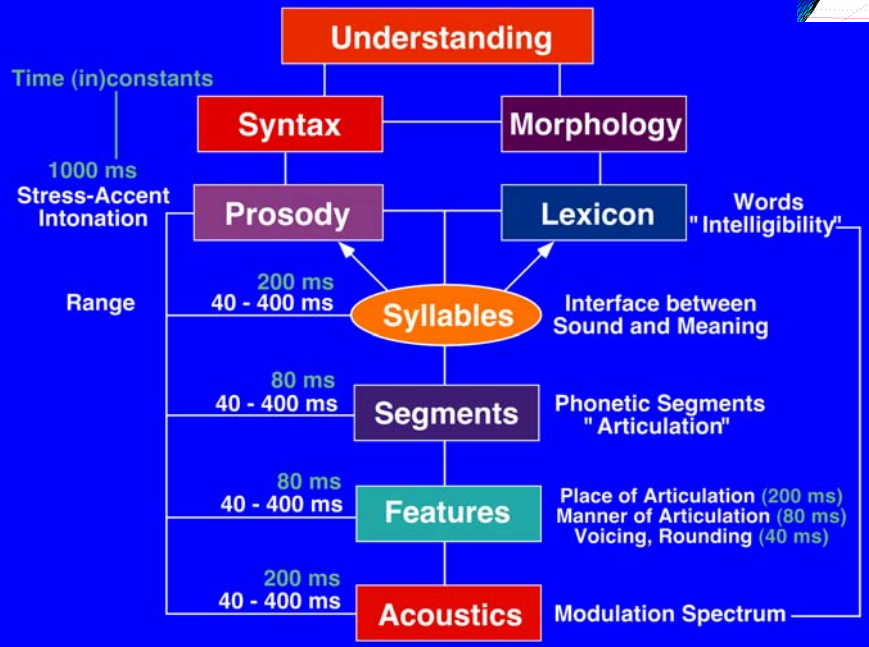
Therefore, it is ultimately INFORMATION (and lexical discriminability) that governs the detailed phonetic properties of spoken language, and hence pronunciation variation largely reflects information contained in spoken language

Language - A Syllable-Centric Perspective

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

Modes of Analysis



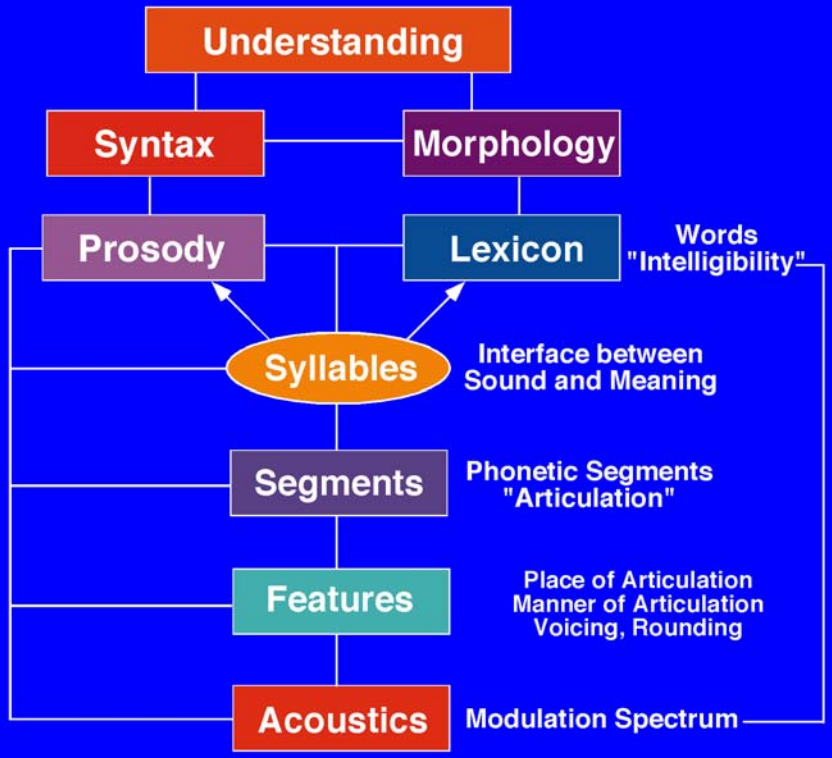


Linguistic Tiers

Conclusions

The speech signal needs to characterized on a variety of linguistic levels in order to be most useful for automatic recognition

How these representations interact with each other and can be combined for optimum recognition forms the frontier of speech research



Many Thanks

for

Your Time and Attention