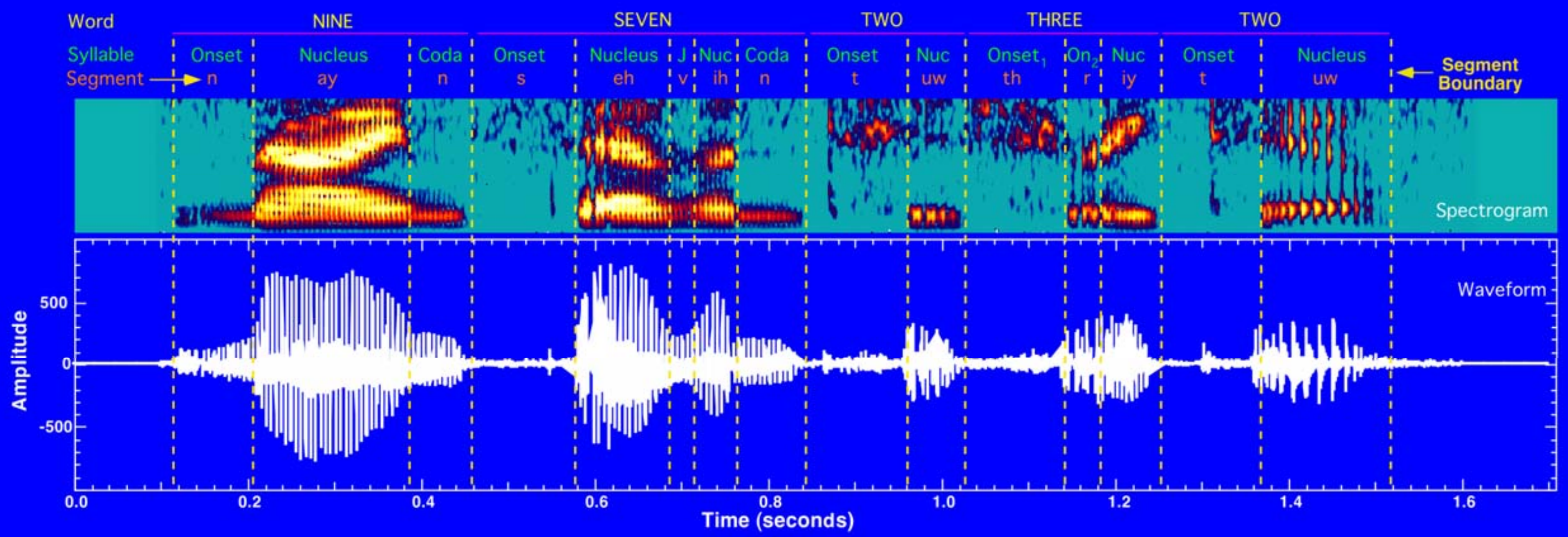
Beyond Landmark-Based Speech Recognition

Steven Greenberg
Johns Hopkins University
August 16, 2004

Or …

How Automatic Speech Recognition Might Work in 2020



“The purpose of computing is insight, not numbers”

Richard Hamming

What's Wrong with the Landmark Approach?

The "Landmark" approach to automatic speech recognition makes some simplifying assumptions that are not necessarily correct

Among the most important are:

(1) The speech signal can be adequately characterized by a sequence of acoustic signatures (termed "landmarks")

(2) The acoustic signal can be mapped back to articulatory configurations that can be modeled with precision (this is the basis of the "Dynamic Bayesian Network" of articulatory dynamics, and represents in effect a "Motor Theory" of automatic speech recognition)

(3) Lexical models bear some systematic relation to sequences of acoustic landmarks

(4) Such landmark-based clusters are sufficiently discriminative at the lexical level to significantly improve recognition performance (relative to the conventional phoneme-based models)

Speech is NOT a Sequence of Landmarks

The Landmark approach is essentially a passive acoustic detector with a variable (and potentially uncertain) relation to speech production

The acoustic properties of the speech signal are viewed as the inevitable consequence of articulatory movements associated with words

Within this perspective, words are STILL viewed essentially as sequences of segments, with the caveat that certain articulatory properties associated with the segments can "desynchronize;" and under certain conditions landmarks and segments may delete or reduce

Entropy – The Missing Dimension

What's missing from the current landmark approach?

–ENTROPY (Information) (–ENTROPY) (Information) (–ENTROPY)

With respect to the recognition task at hand,– entropy translates into acoustic patterns capable of reliably distinguishing among words

One approach to applying this principle is through "Confusion Networks" (Katrin Kirchhoff's project for this workshop)

The problem with confusion networks is their reliance on large amounts of training material and the specific corpus-centric nature of the lattices and n-best lists used to generate them

Moreover, confusion networks do not provide a clear path towards the development of future-generation speech recognition systems

An Entropic Approach to ASR Development

What is required to make a non-phonemic approach a viable alternative to current-generation systems?

This forms the focus of the remainder of this presentation

The Importance of Segmentation

Currently, ASR systems do not attempt to explicitly segment the speech signal

There is no attempt to estimate the number of words, syllables or other constituents before recognition

Nor is there a concerted effort to delineate a linguistic structure prior to the final stage of recognition

If it WERE possible to accurately estimate the number of syllables, as well as their temporal demarcation, many aspects of the speech recognition would be much simpler

For one thing, it would be possible to estimate the number of phonetic constituents within each syllable and also ascertain whether syllables are likely to be relatively independent entities ("wallflowers") or bound together to form units such as words or word phrases ("withs"), as well as estimate the amount of entropy associated with any given span of speech

Segmentation can be performed using a variety of methods (as described on subsequent slides)

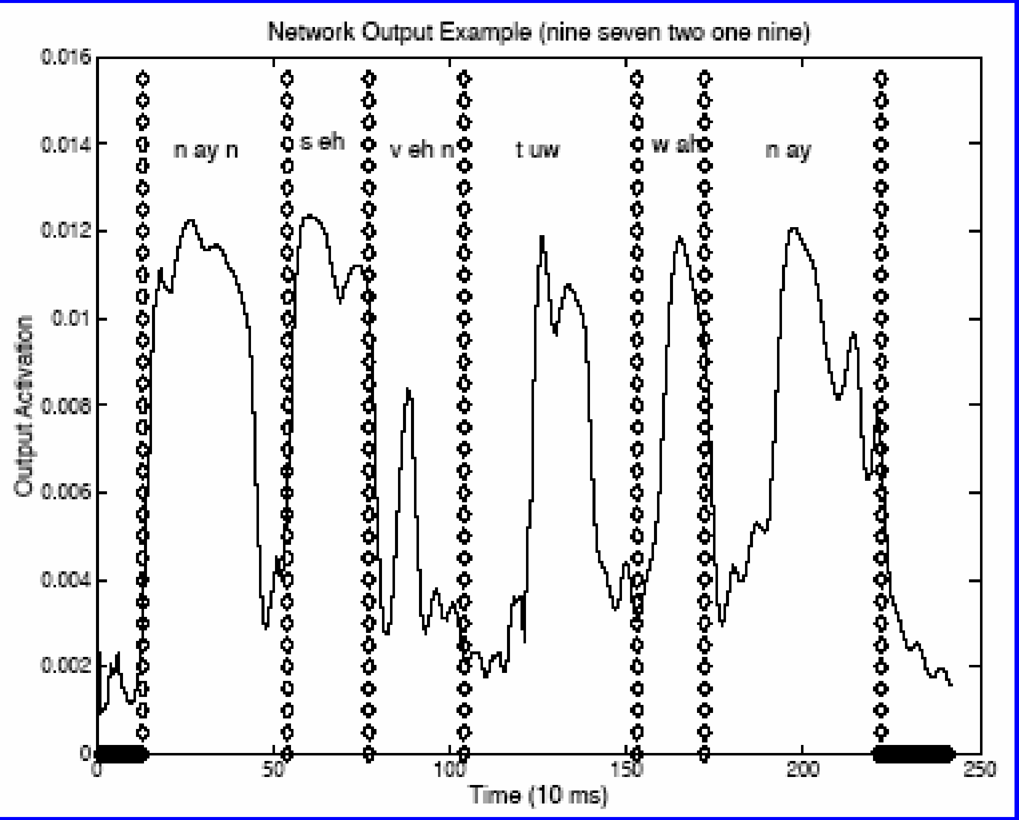
Syllabic Segmentation of the Speech Signal

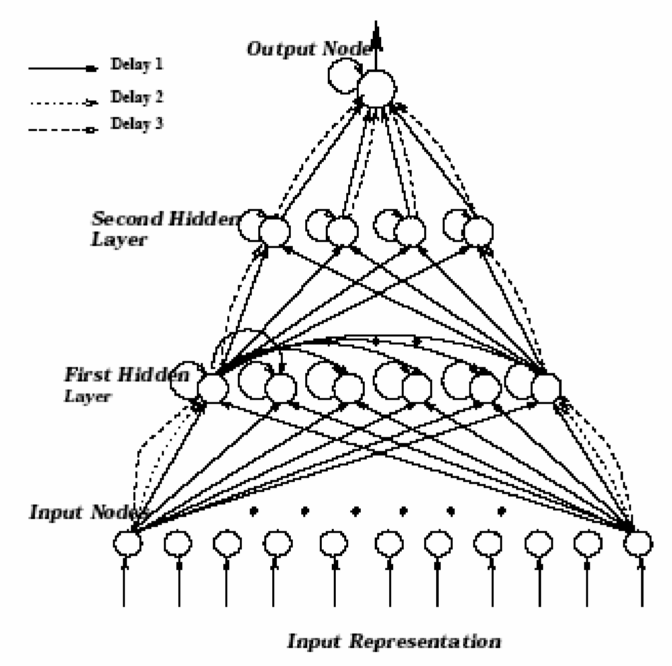
Signal-processing-based approaches

Group delay (phase) of the spectrum (Murthy and colleagues)

Neural networks using training data (e.g., Shastri, Chang & Greenberg, 1999)

Performance is ca. 85-95% accurate within ± 10 ms tolerance limit



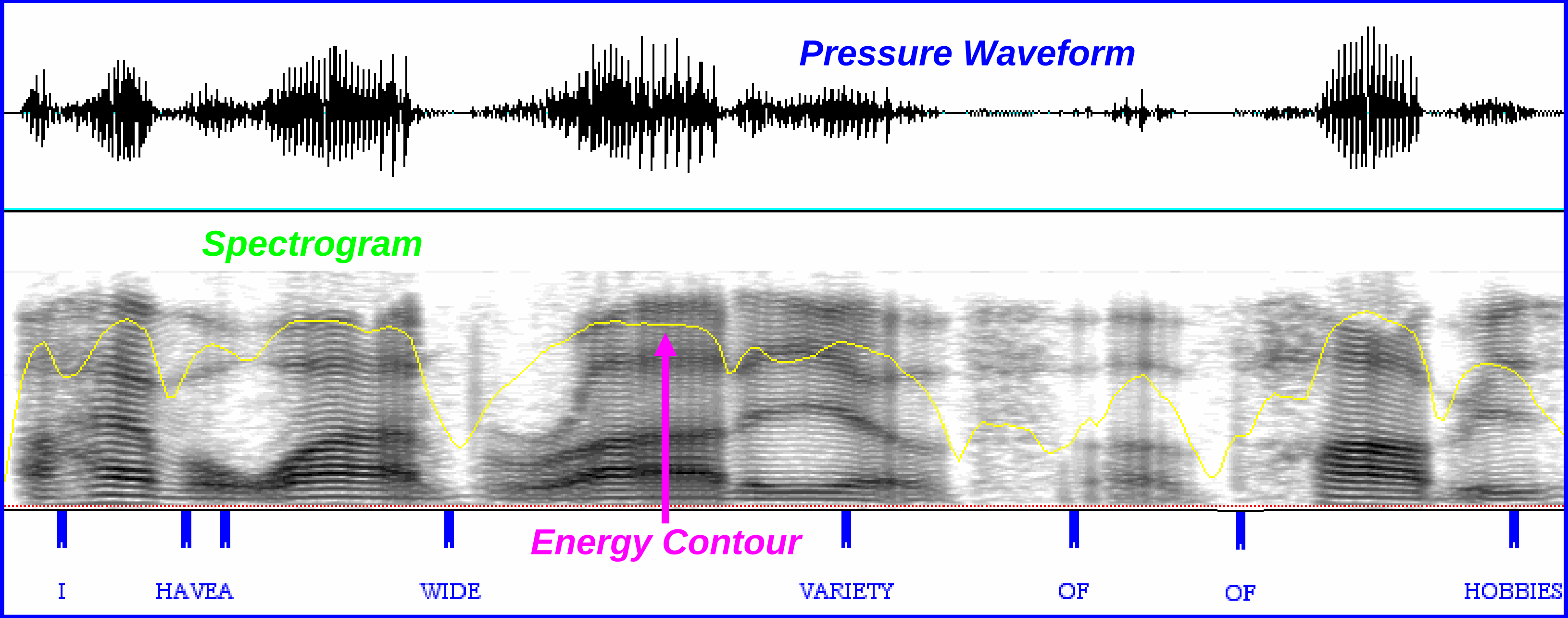


Shastri, Chang and Greenberg (1999)

Syllabic Segmentation of the Speech Signal

One possible segmentation of the acoustic waveform might look like….

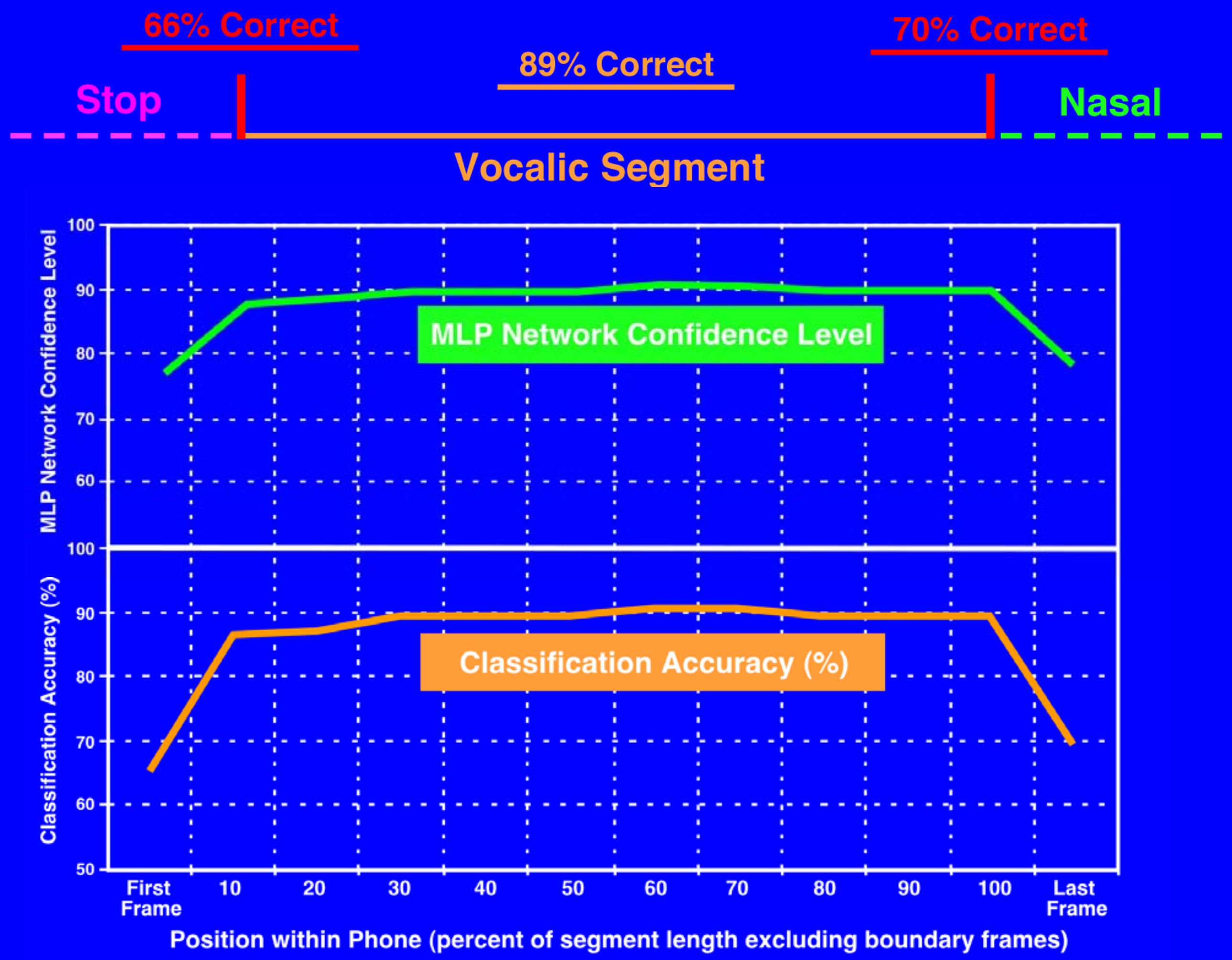
(where the syllabic (energy) contour is marked in yellow)



Switchboard WS96 Development Data

Phonetic Segmentation

The confidence estimates of MANNER classifiers can be used to delineate temporal boundaries associated with the segment

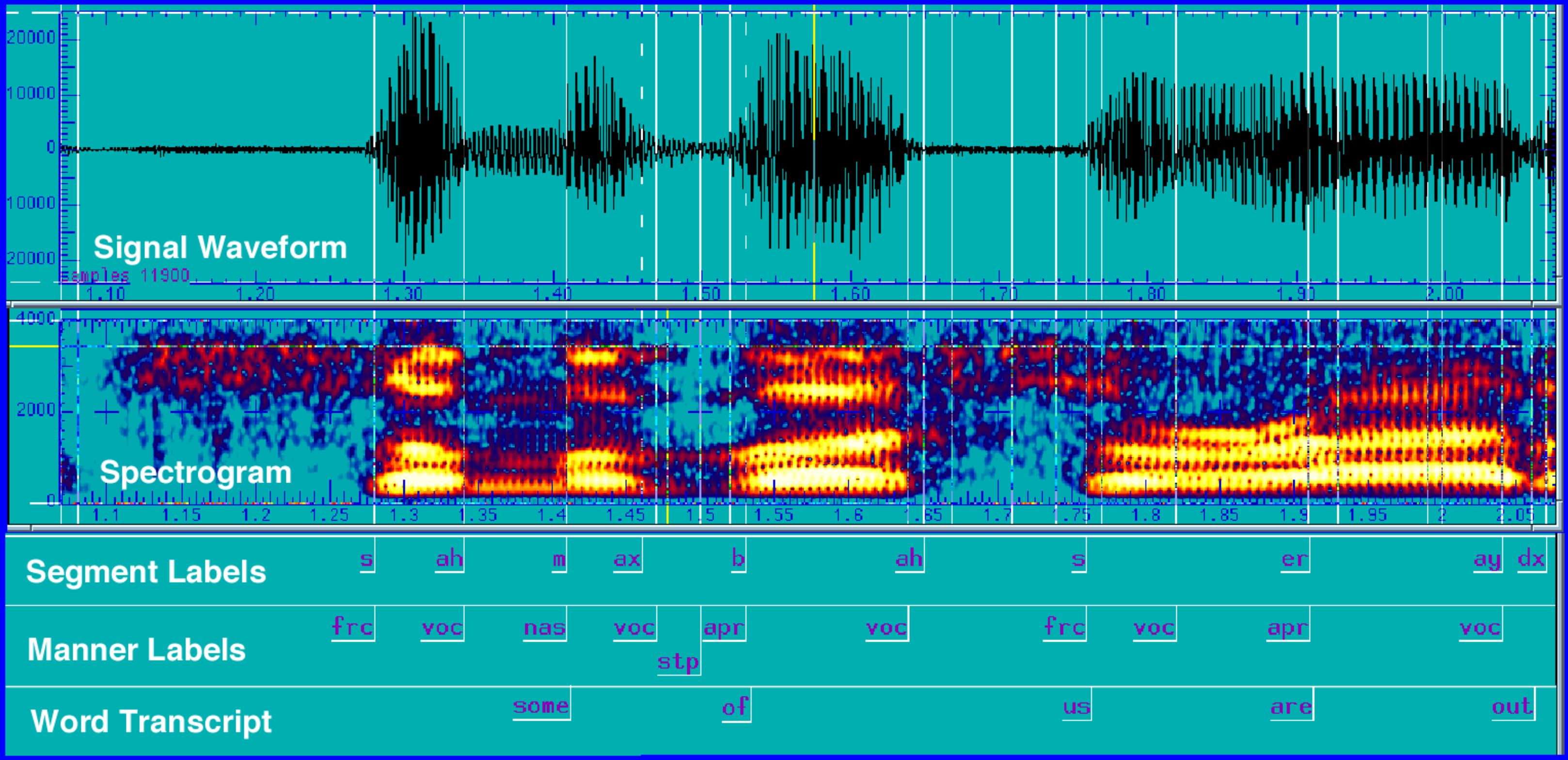


(Chang, Wester & Greenberg)

Phonetic Segmentation

Phonetic segmentation can be largely achieved through manner-of-articulation classification (as shown for the Switchboard corpus below)

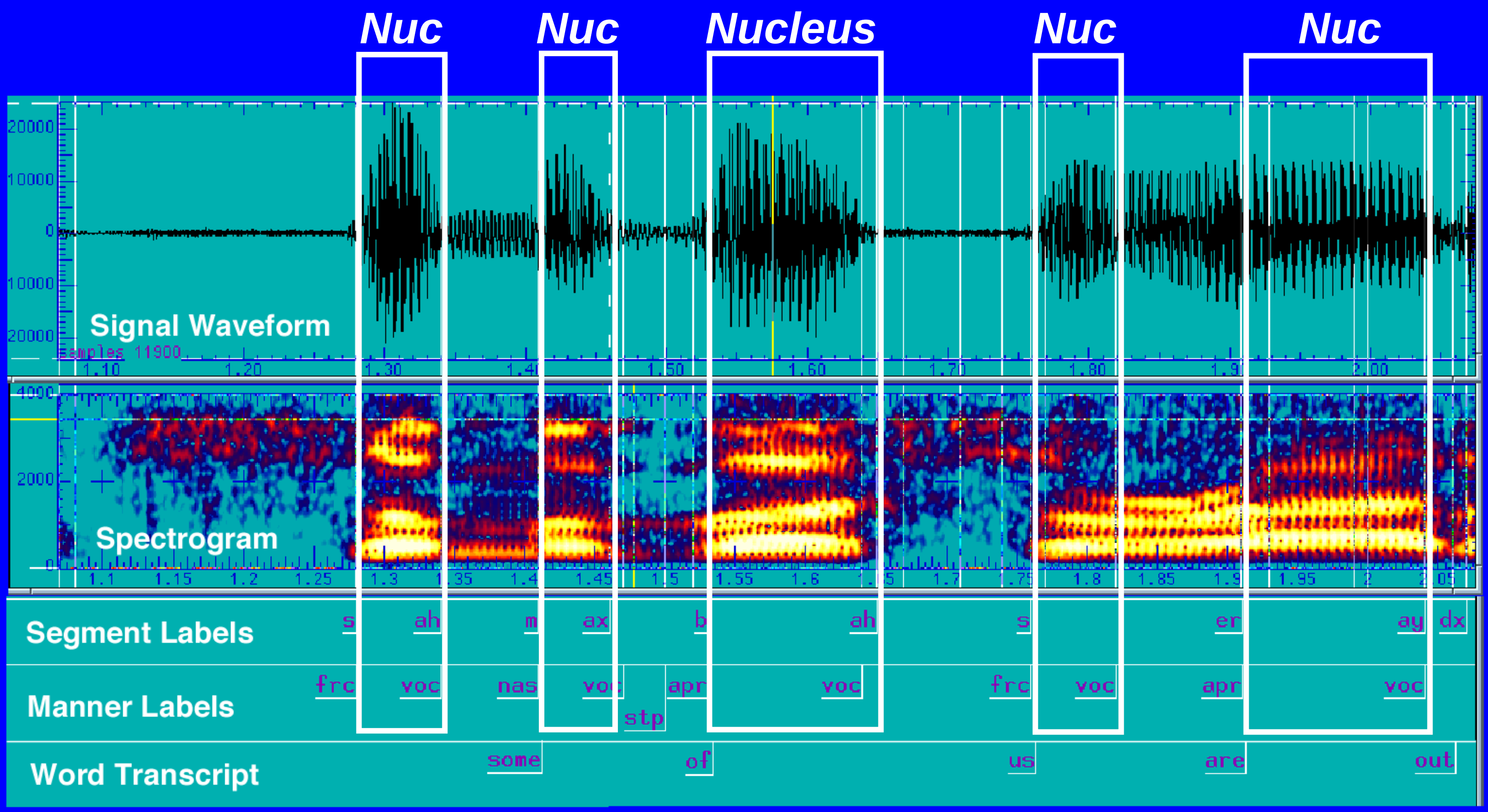
Manner is temporally isomorphic with the concept of the phonetic segment



Vowel Spotting – Implicit Syllabification

Virtually all syllables have vowels at their core (i.e., nucleus)

Vocalic-manner classifiers can be used to perform implicit syllabification



Phonetic Classification

***Once segment boundaries have been delineated,* it should be much easier (in principle) to classify the relevant portion of the signal with respect to:**

(1) Place of articulation

(2) Specific manner of articulation

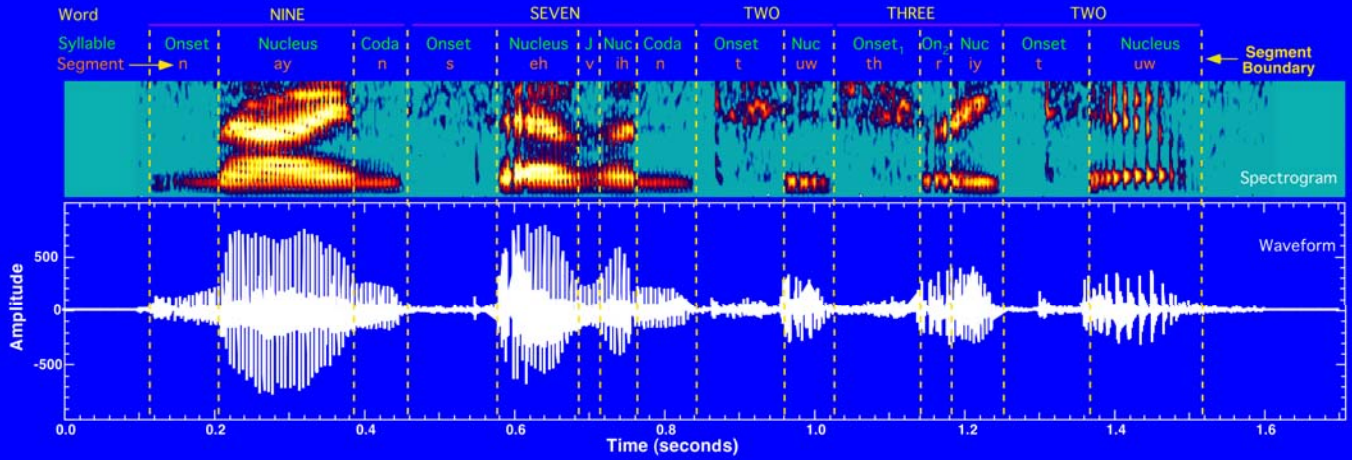
(3) Voicing

(4) Lip rounding, and so on

(5) Associating segments with syllable constituents (i.e., onset, nucleus, coda)

To a certain degree, this was done as part of the Landmark Speech Recognition project this summer

And has also been performed by others in the past, including Chang, Wester and Greenberg, Juneja and Espy-Wilson

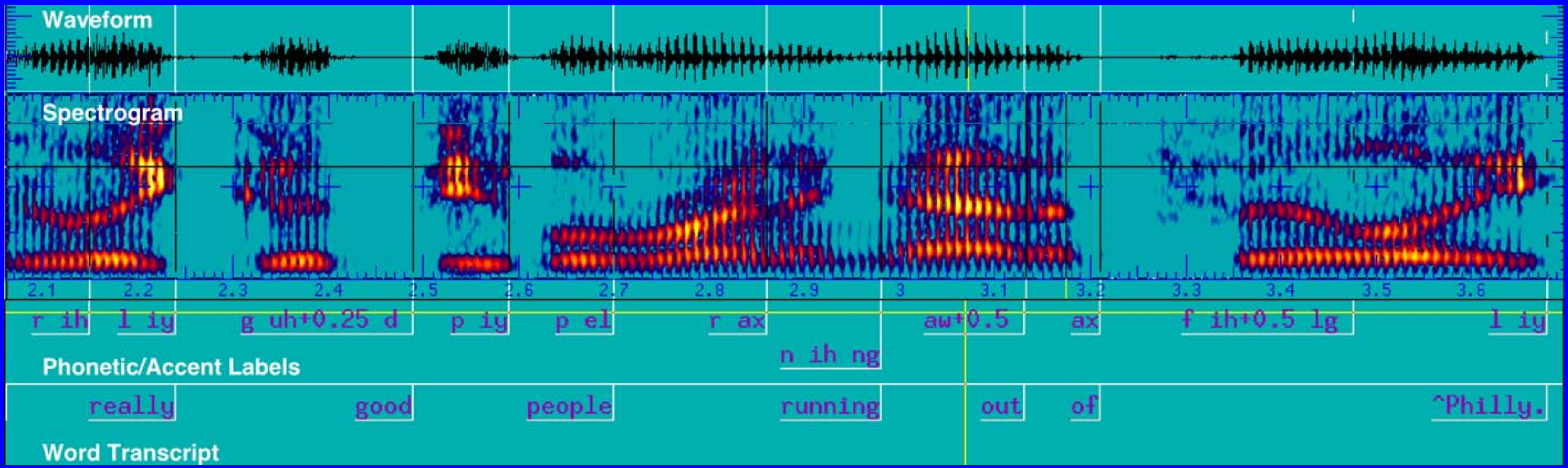


Importance of Stress Accent & Segmentation

There's an enormous amount of variation in the pronunciation (and hence articulation and acoustic properties) of words in conversational speech

Detailed statistical analyses of the Switchboard corpus demonstrate that much of this variation is structured and systematic at the level of the syllable, particularly when the accent weight of the syllable is known (Greenberg, 1999; Greenberg et al., 2002, 2003; Hitchcock & Greenberg, 2001)

Fortunately, the accent can be reliably computed directly from the syllable nucleus (usually vocalic)



Importance of Stress Accent & Segmentation

Within the context of this summer's workshop, support vector machines (SVMs) were designed that reliably label the stress accent of syllables based on the features:

(1) Duration of the syllable nucleus

(2) Normalized energy of the nucleus

(3) Vocalic identity of the nucleus (in terms of vowel height, frontness, and tenseness)

The SVM classifier (developed by Vidya Mohan and Amit Juneja) is able to simulate manual labeling of Switchboard data extremely well using these features (which were used in an MLP implementation developed by Greenberg and Chang a few years ago for the Switchboard corpus)

Accent Affects Phonetic Properties

Many aspects of pronunciation variation are related to accent weight of the syllable

The probability of segmental deletion, vocalic identity (particularly height), and voicing are all related to the syllable's accent weight

Why should this be so?

Because accent reflects INFORMATION associated with the syllable (and more)

And information is what determines the specific properties of pronunciation variation (not the "laziness" of articulators or talkers)

Accent Reflects Information

Syllables that are lexically and semantically discriminative are far more likely to be accented than their unaccented counterparts

Accent can thus be used to compute the amount of information associated with a syllable…

Along with other phonetic properties of the syllable

Unaccented syllables tend to be shorter and contain fewer segments than their (heavily) accented counterparts

In fact, the INTRINSIC information of a syllable (and a word) can be computed from its phonetic properties alone, without recourse to lexical and phonetic context (as shown on the following slide)

Computation of Information within a Syllable

The syllable can be decomposed into the following phonetic dimensions:

(1) PLACE of articulation (the most important dimension entropically)

(2) MANNER of articulation (also quite important in terms of information)

(3) VOICING – potentially discriminative, but often not so effective

(4) LIP ROUNDING – potentially discriminative, but often not particularly so

(5) ACCENT – functions as a meta-feature affecting the interpretation of other syllabic and phonetic properties

PLACE of articulation is affiliated with SYLLABIC constituents and MORPHEMES (NOT segments)

There's generally a single place ENTROPE associated with each ONSET, CODA and NUCLEUS constituent

EXCEPT when associated with a bound morpheme (e.g., past-tense marker / – t / "kept" where the /p/ and /t/ are separate place entropes and morphemes

Articulatory PLACE is an entropic dimension par excellence

The more entropes contained in a syllable, the more intrinsic information – this is consistent with basic information theory (particularly Mandelbrot's information-theoretic extension of Zipf's law)

Computation of Phonetic Entropy – "Strings"

Schematic - for illustrative purposes only

"STRINGS"

Segment	Place	Manner	Voicing	Entropes	Cum
s	ø*	Fricative	–	1	1
t	Central	Stop	±**	2	3
r	ø	Rhotic	+	1	4
I	Front	Vocalic	+	2	6
N	Back	Nasal	+	2	8
s	Central	Fricative	±**	2	10

* *In consonant clusters /s/ usually has no articulatory place apart from that of the dominant consonant*

** *Voicing is optional in these contexts (voicing is a syllabic feature reflecting accent)*

Computation of Phonetic Entropy – "And"

Schematic - for illustrative purposes only

"AND" – Canonical and Stressed … phonetically – [ae] [n] [d]

Segment	Place	Height	Manner	Entropes	Cum
ae	Front	ø	Vocalic	2	2
n	Central	ø	Nasal	2	4
d	ø	ø	Stop	1	5

"AND" – Conversational and unstressed … phonetically – [n]

Segment	Place	Height	Manner	Entropes	Cum
ae*	ø	ø	ø	ø	ø
n	ø	ø	Nasal	1	1
d*	ø	ø	ø	0	1

* *Segment is "deleted" from pronunciation*

Intrinsic Information and Pronunciation

The lower the INTRINSIC information associated with a word, the more highly variable is its pronunciation over a broad range of contexts (and the more the ACTUAL information will vary)

Thus, this entropic metric can be used to estimate the likely variability associated with any word (particularly if the unigram frequency is known), and indirectly the likelihood of lexical confusability in a speech recognition system

Accented syllables are likely to be canonically pronounced most of the time, and are also likely to have a high degree of intrinsic information

Unaccented syllables are more likely to contain relatively little information and be far more variably pronounced that their accented counterparts

Entropy-Based on Pronunciation Variation

From the foregoing it follows that the amount of variability observed in pronunciation is likely to be correlated with the intrinsic information of a word

Therefore, it is possible to estimate the amount of information associated with a word by measuring the amount of pronunciation variation

Words with a high degree of variability are likely to have low –entropy

While words with little variability are likely to have much higher –entropy

(one problem with this approach is that the number of instances of high-entropy words far fewer than their low-entropy counterparts, thus lowering the possibility for observed variability)

But the principle probably holds despite this complication

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

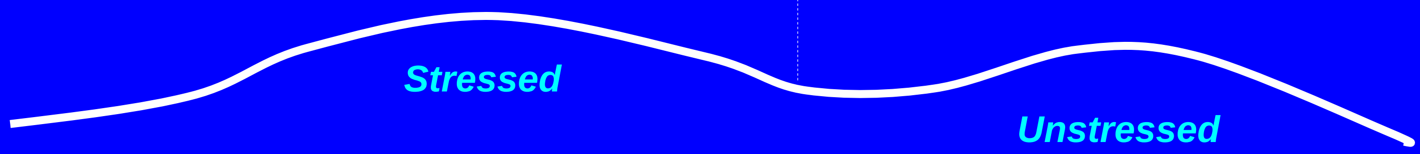
Energy contour

And so on ….

WORD – “Strengthen”

	SYLLABLE – “streng”					SYLLABLE – “then”		
	ONSET			NUCLEUS	CODA	ONSET	NUCLEUS	CODA
Segment	s	t	r	ɛ	N	T	I	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

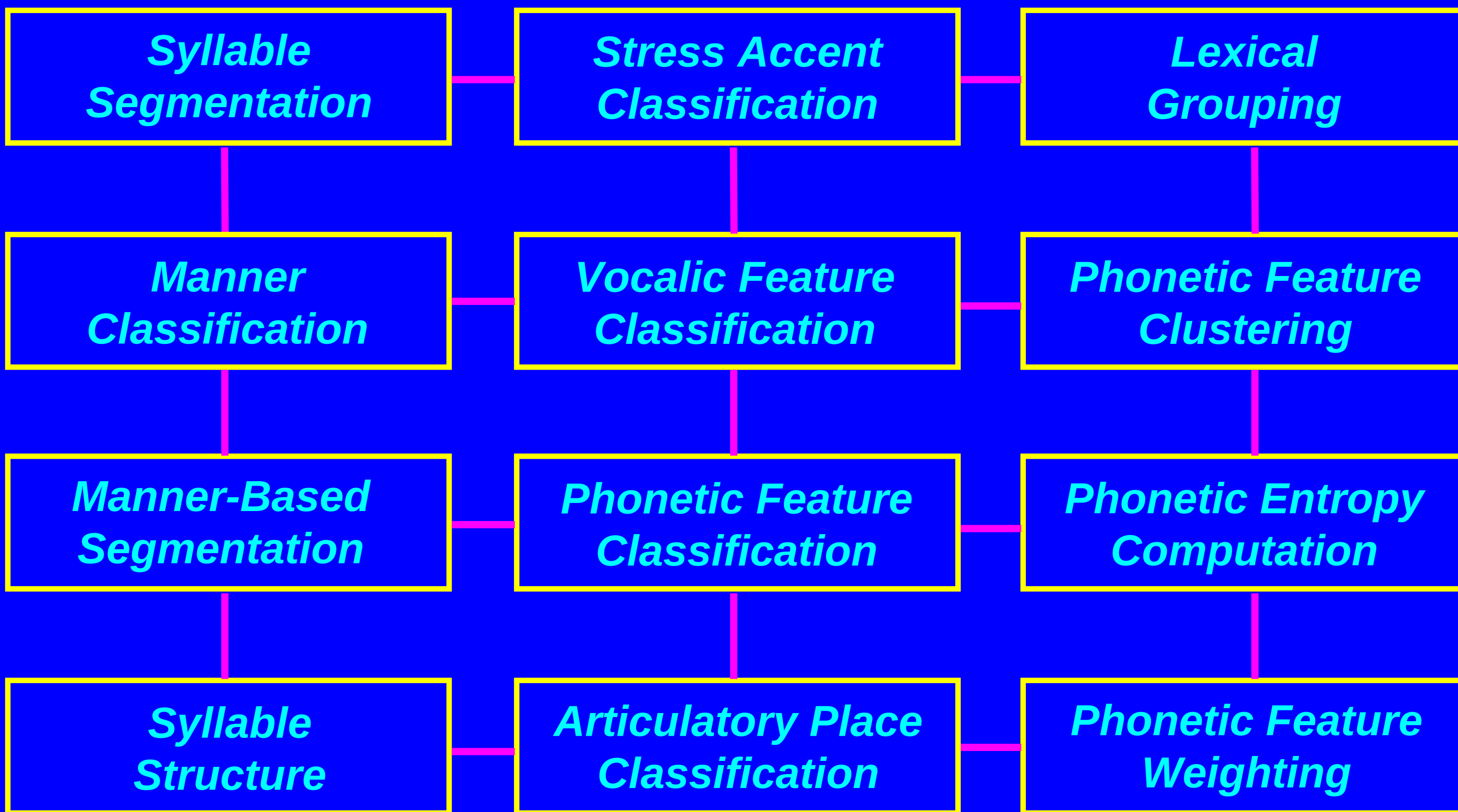
Energy Contour



An Alternative Architecture

Words are composed of phonetic-prosodic features, which can be derived in the following way…

(n.b. – this is NOT word recognition per se, but rather a specification list, where most of the steps are intertwined)



Conceptual Basis of the Lexicon

The lexical representations represent an attempt to encode information likely to be used by human listeners

Thus, duration and energy dynamics should be part of the lexical representation

The lexicon assumes that manner, place and syllable position are the key parameters underlying the specification of a word

Automatic Generation of Pronunciation Models

A pronunciation lexicon can be generated in the following manner …

Where each feature set can be an n-dimensional object with a statistical distribution (mean and variance, etc.)

Pronunciation models are distributions within a high-dimensional space

The key is matching the lexical pronunciations to the classifier output and vice versa (à la McAllaster et al., 1998)

Word X_i

Phonetic Features
Prosodic Features
Other Features

Word X_j

Phonetic Features
Prosodic Features
Other Features

Word $X_{k\ldots}$

Phonetic Features
Prosodic Features
Other Features

Word X
Ensemble Representation

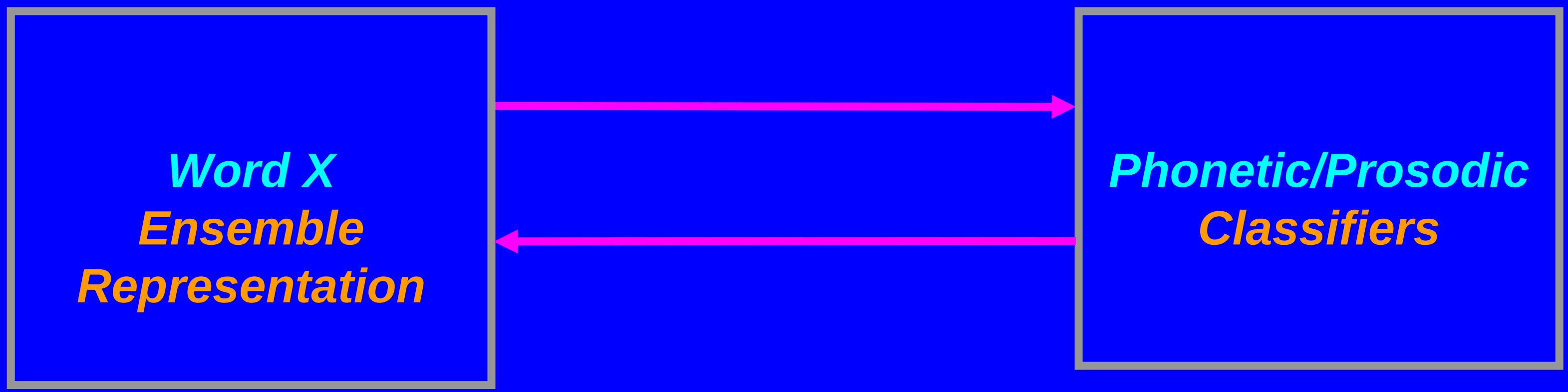
Phonetic Features
Prosodic Features
Other Features

Tuning the Lexicon to Recognition Features

- ***The actual lexical entries should be far more comprehensive, encompassing all of the major pronunciation variants AS RECOGNIZED by the classifiers***
- ***This can be performed by having the classifiers operate on training material comparable to the test data***
- ***Each word in the training corpus can be clustered with comparable words and the classification patterns associated with each word incorporated into the recognition lexicon***

Tuning Recognition Features to the Lexicon

- ***Once the lexical representations have been developed, some form of linear discriminant analysis could be performed in order to lower the dimensionality of the representation***
- ***And to leave only truly discriminative features in the lexicon***
- ***It is these LDA-based features that the phonetic classifiers need to use for word recognition***
- ***If performed, this would accomplish a data-model concordance, as suggested by McAllaster et al. (1998)***
- ***And thereby substantially reduce word error rate***



That's All

Many Thanks for Your Time and Attention

Additional information can be obtained from the Landmark Speech Recognition Web Site – www.clsp.jhu.edu/ws2004/ws041dmk