Automatic Identification and Classification of Words using Phonetic and Prosodic Features

Vidya Mohan
Center for Speech and Language Engineering
The Johns Hopkins University

Introduction

Motivation

- Substitution errors in Automatic Speech Recognition (ASR) tasks could be reduced if finer grained units than phonemes were used to capture changes in the waveform.
- A syllable structure, along with its constituent components - onset, nucleus, coda - could better represent acoustic/articulatory/prosodic features.
- By building better acoustic-phonetic models, where features are weighed according to their discriminative ability, word error rate (WER) might be decreased.
- Current speech recognizers generally accord the same level of importance to the onset and the coda, as well as to accented and unaccented parts of speech.
- It would therefore be useful to more fully understand how words are related to their acoustic/articulatory/prosodic features.

Introduction

Statement of Proposal

- Build *word models* by identifying which phonetic and prosodic features are critical in recognition, and thus being able to create a word templates defining them.
- Evaluation - Structured word identification and classification.
- *Note:*
  - Word identification would be used for proof of concept
  - Classification in confusion networks would allow for integration with the current ASR systems

Candidate Features

- We are interested in features that preserve as much ***information*** of the speech waveform as possible
- We already have feature detectors for most of these features
- Features of interest are:
  - Articulatory
  - Acoustic
  - Prosodic
- *Articulatory Features*
  - Manner - Fricative, spirant, stop, nasal, flap, lateral, rhotic, glide, vowel, diphthongs
  - Place - anterior, central, posterior, back, front, tense, labial, dental, alveolar, velar
  - Voicing
  - Lip rounding
- *Prosodic Features*
  - Prosody and stress accent - some syllables are more stressed than others

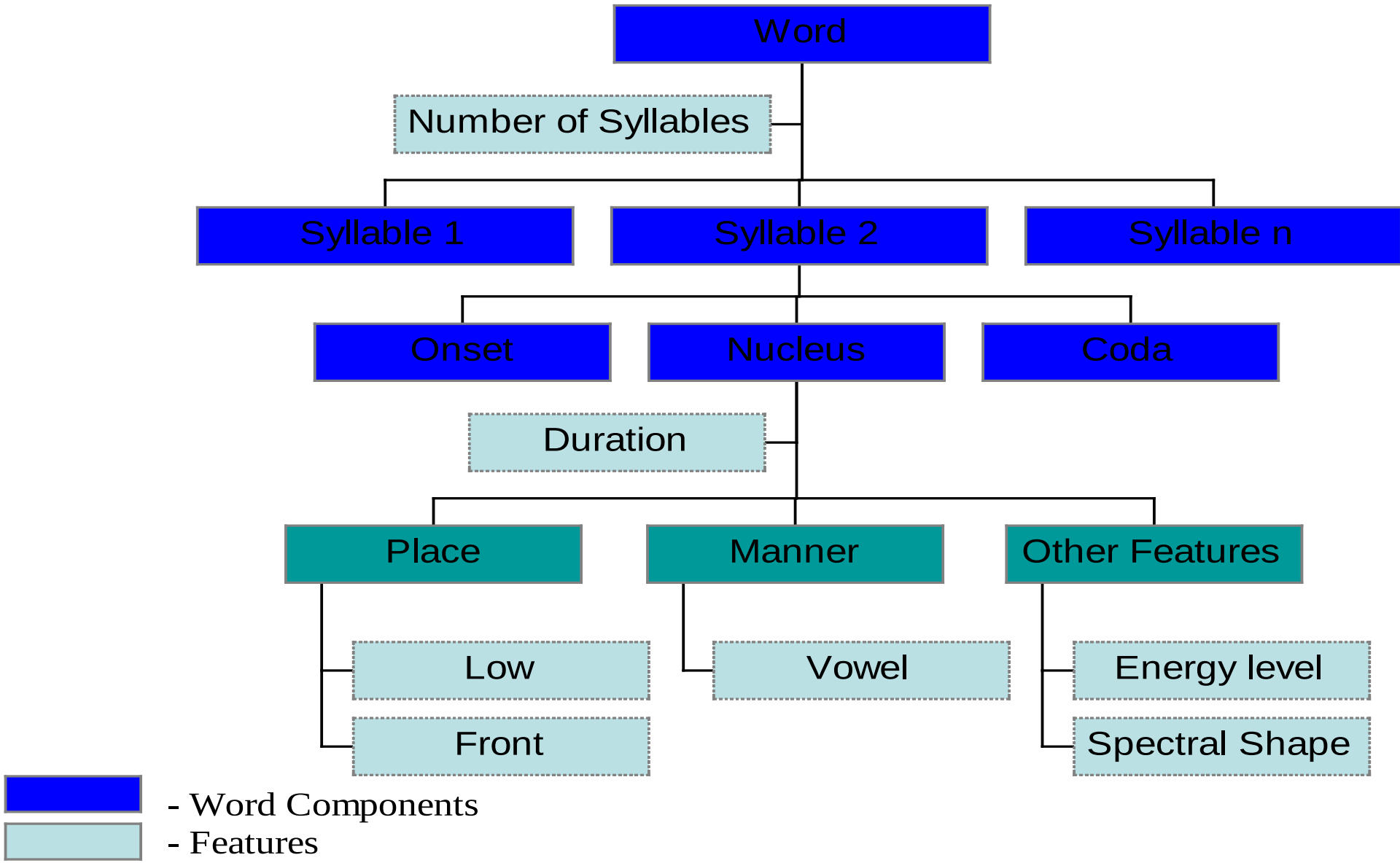
Candidate Features

- *Acoustic Features*
  - Knowledge Based (formant) Acoustic Parameters
  - Neural Firing Rate Features (rate scale)
  - Energy level and modulation
  - Duration - of words, syllables and constituents
- *Other Features*
  - Number of syllables
  - Sensitivity to context (feature weighting)

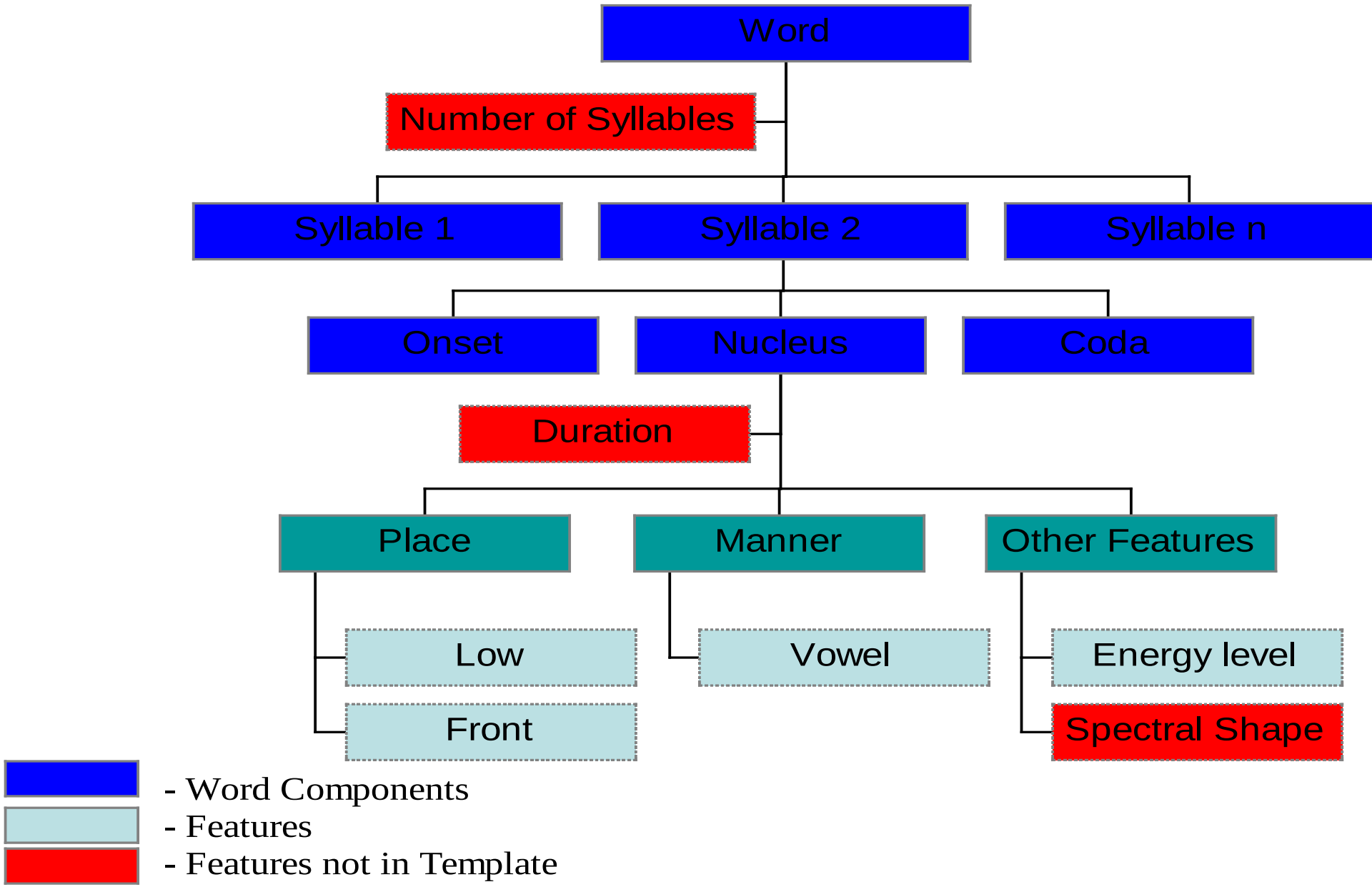
Summary

- Which of the above features are most likely to be preserved in all representations of the word?

Word Template Creation
Structural Components of a Word



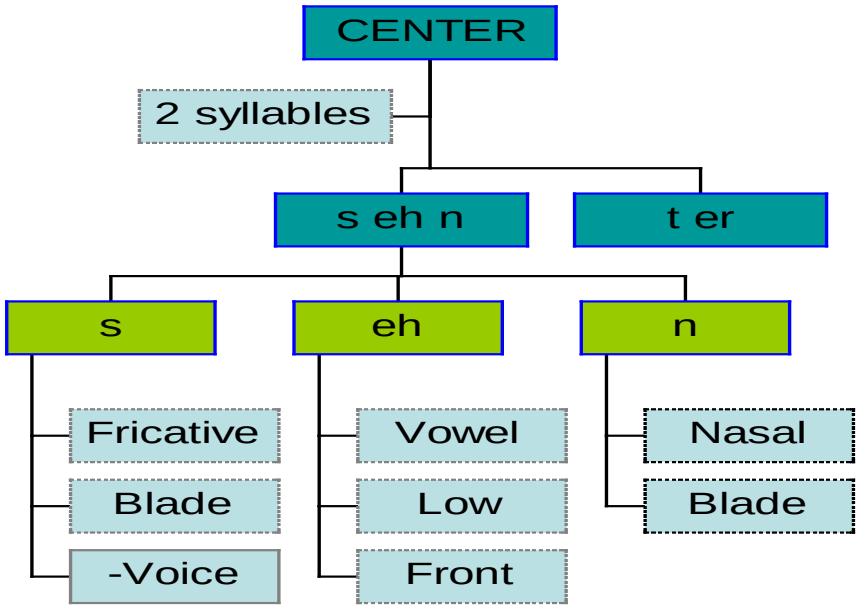
Word Template Creation
Structural Components of a Word



Word Template Representation

Summary

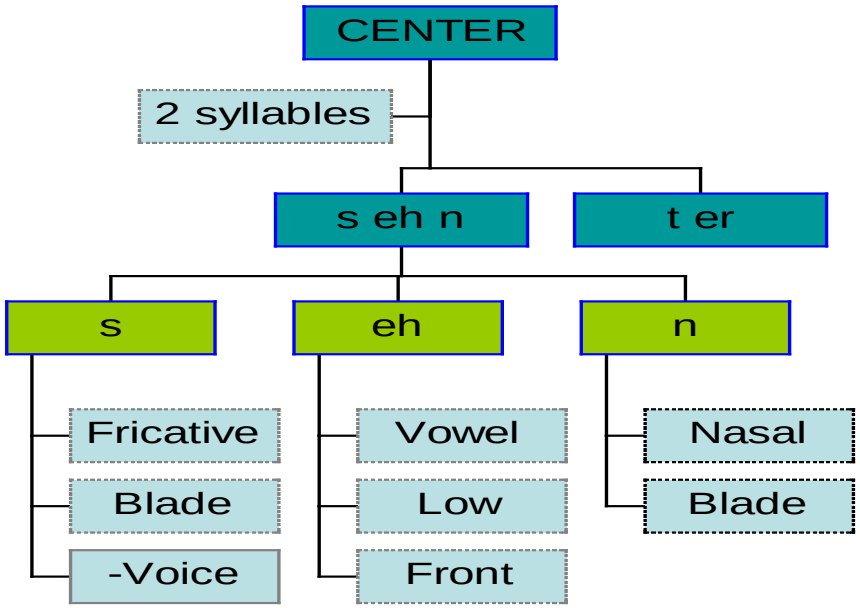
- Template - features selected and weighted according to their importance in the identification of the word.
- Thus the system might learn the following representation of a word, say, *center*,

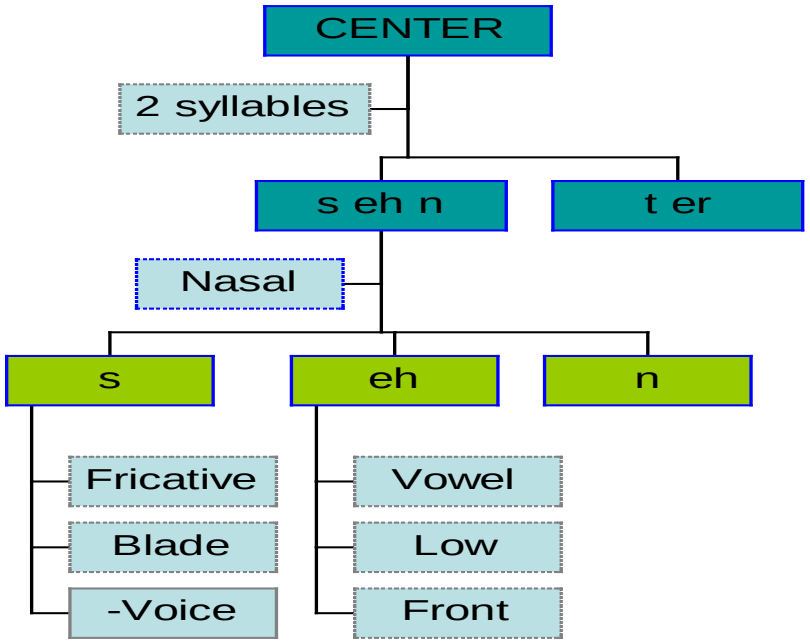


Word Template Representation

Summary

- Template - features selected and weighted according to their importance in the identification of the word.
- Thus the system might learn the following representation of a word, say, *center*,





Methodology

- **Choosing the specific words to study**
  - Common confusable words from Switchboard
- ***Broadly Three Step Process***
  - Feature Generation
  - Word Template Creation
  - Evaluation
- ***Feature Generation***
  - Use the phonetic classifiers to generate features for the whole corpus
  - Stress Accent Detector for Syllables - Accuracy of 79% on manually transcribed corpus of Switchboard has been obtained at WS ‘04

Methodology

- ***Word Template Creation***
  - Features of a particular word are selected according to their information content.
  - Mutual Information between a word and a feature captures the notion of information content quantitatively
  - $I(W; F) = H(W) - H(W/F)$

    $= \sum \sum p(w, f) \log[p(w|f)/p(w)]$
  - The importance of a feature will be weighted according to the mutual information value, accuracy of the classifier and the stress accent pattern of the syllable.
  - $\sum I(W; F) >= I(W; F_1, F_2, \ldots F_n)$

    There will definitely be dependencies between the features so methods for their careful selection would be used.
  - Features can be decided upon by discriminative analysis as well, especially in the cases with data sparsity

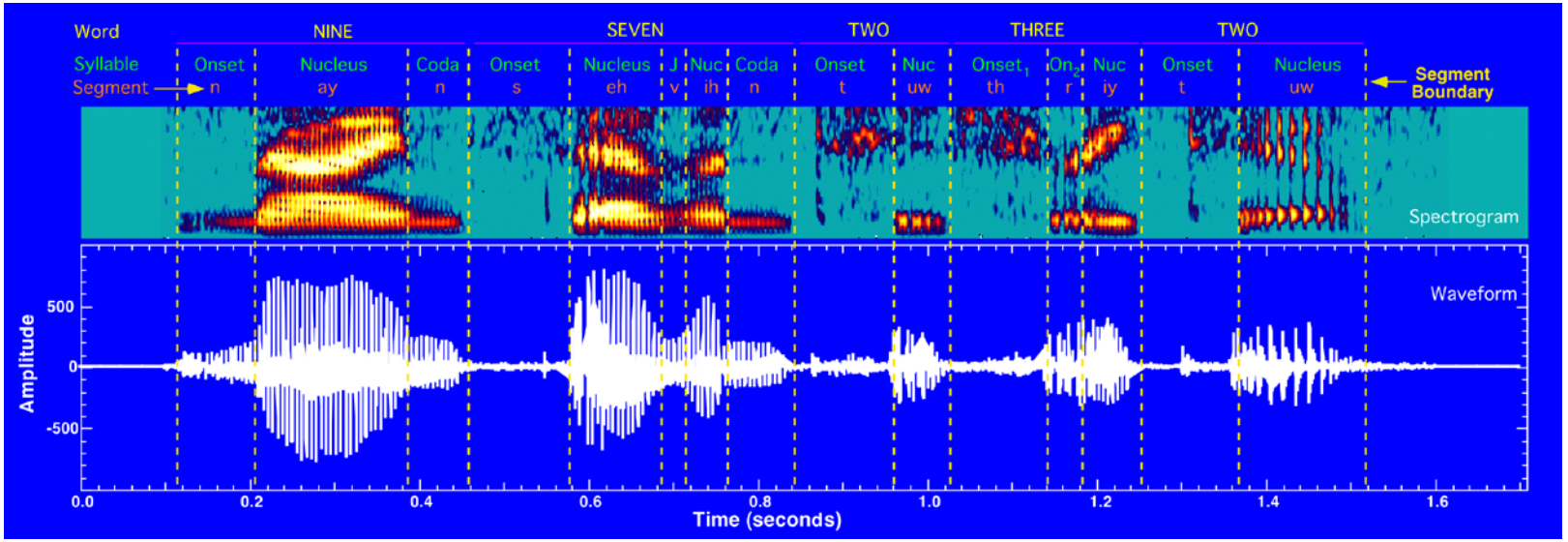
Methodology

- ***Evaluation Task***

  The feature set will be evaluated on two tasks,

  – Word Identification

    - Use this template to find the temporal bounds of a particular word in an utterance
    - Evaluation Metric: Equal Error Rate
    - Utterances will be chosen from the TIMIT and Switchboard corpus
    - Segment the utterances using a syllable classifier



  – Word Classification

    - Develop classifiers that could be used to classify word confusion pairs that exist in a lattice.

Summary

Collaborators

Mark Hasegawa Johnson, *University of Illinois*
Kemal Sonmez, *SRI*
Steve Greenberg, *University of California*
Johns Hopkins Supervisors:
Sanjeev Khudanpur
Izhak Shafran

Summary

- The objective of this proposal is to obtain a *minimal set representation* of a word with respect to its acoustic/articulatory/prosodic features using *mutual information* to choose the features.
- Phonetic classifiers developed/tuned this summer will be used for this purpose.
- Initial word identification experiments will be used to test the proof of concept.
- Given time and efficacy of the method, it will be integrated with the current LVCSR system to distinguish between confusable pairs of words.
- The project will hopefully provide interesting insights as to what the key features of a word are.