# Distinctive feature detection and landmark-based rescoring

**Amit Juneja**

**University of Maryland College Park**

**juneja@glue.umd.edu**

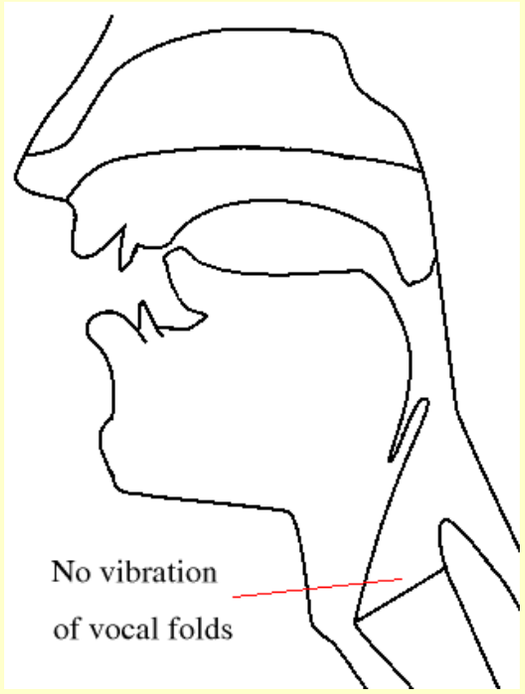
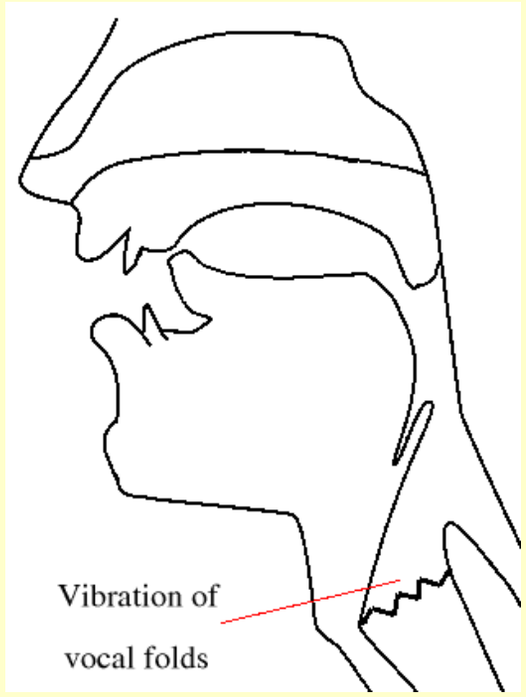
# Agenda

- Phonetic features and acoustic landmarks
- Acoustic features
- Probabilistic landmark detection
- Place and voicing classification
- Application to rescoring
  - DBN and articulatory feature based pronunciation model
  - Constrained landmark and feature detection
  - Queries of feature pairs

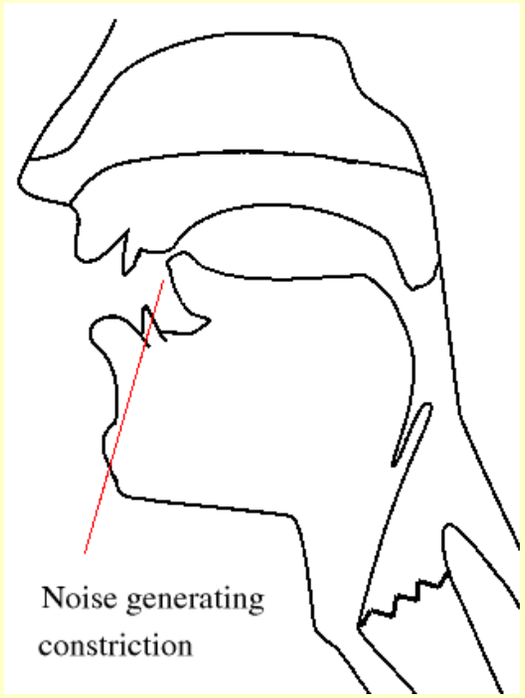
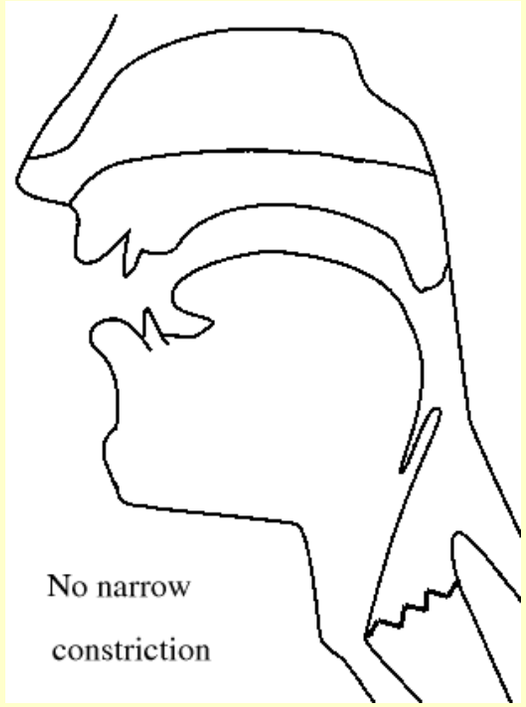
# 1. Phonetic features (Chomsky and Halle, 1968)

- Three kinds of phonetic features characterize speech sounds
  - Source features determine the kind of excitation signal
  - Manner of articulation features determine how open or close is the vocal tract
  - Place of articulation features determine the location of primary constriction
- Phonetic features have articulatory and acoustic correlates, and it is believed that words are stored in memory as bundles of phonetic features (Stevens 2002).

Source feature *voiced*

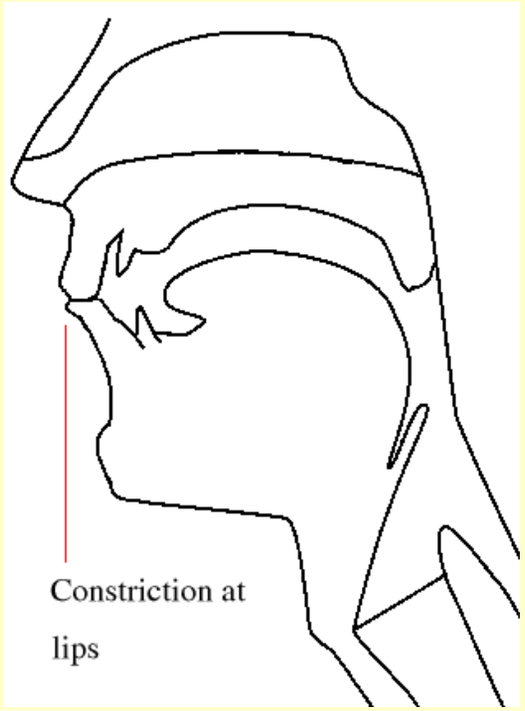
| $-voiced$ | $+voiced$ |
|---|---|
|  |  |
| /s/ | /z/ |

Manner feature *sonorant*

| $-sonorant$ | $+sonorant$ |
|---|---|
|  |  |
| /z/ | vowel |

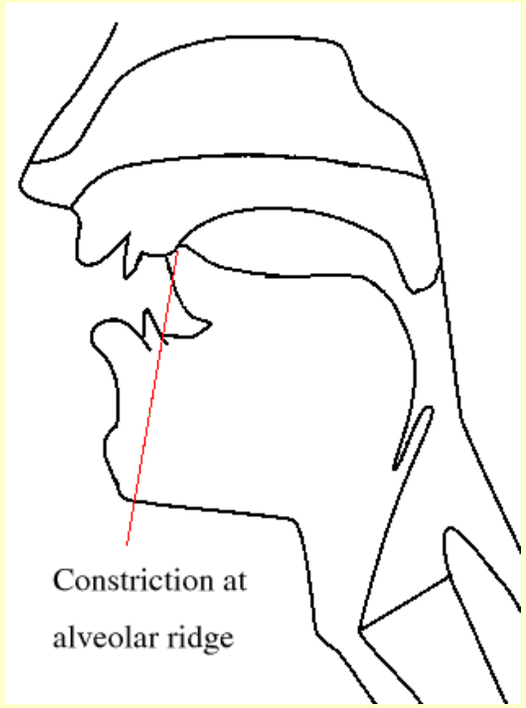
## Stop place features

*+labial*




/p/

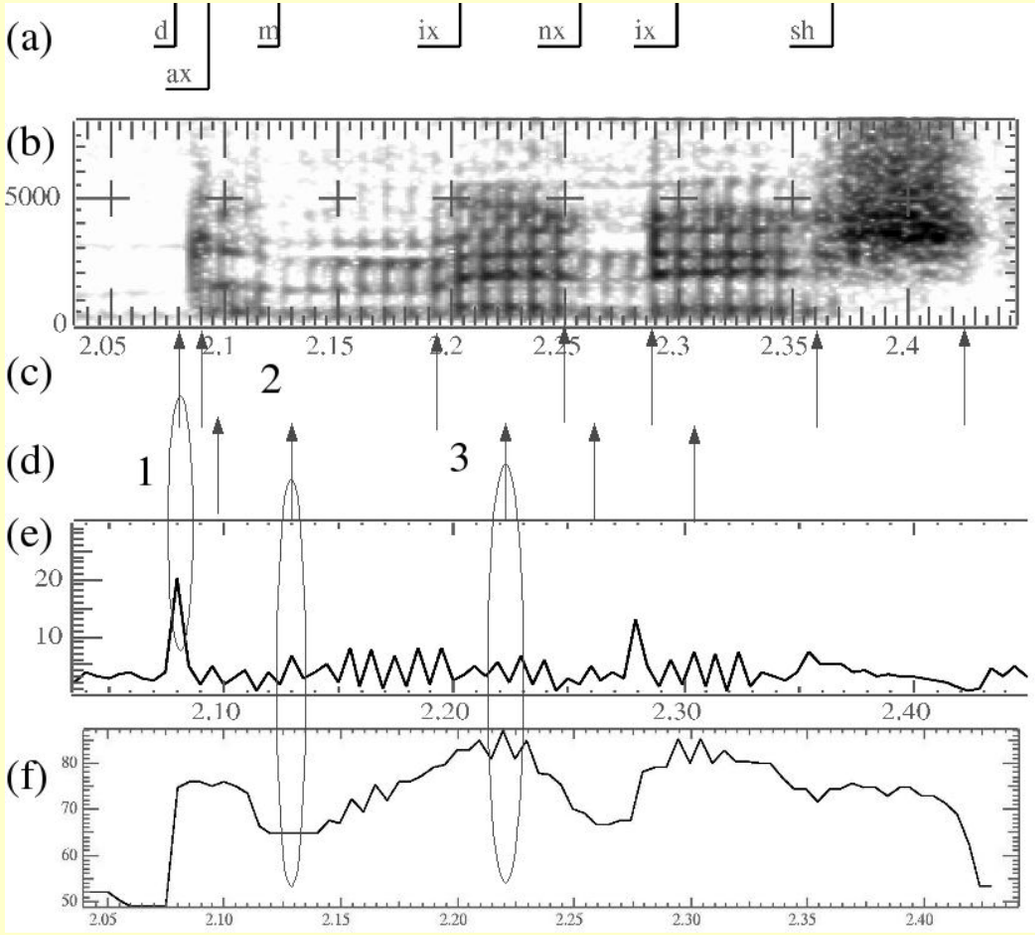
*+alveolar*




/t/

# Acoustic landmarks

(a) phoneme labels, (b) spectrogram, (c) abrupt landmarks, (d) non-abrupt landmarks, (e) energy onset, (f) E[640,2800]



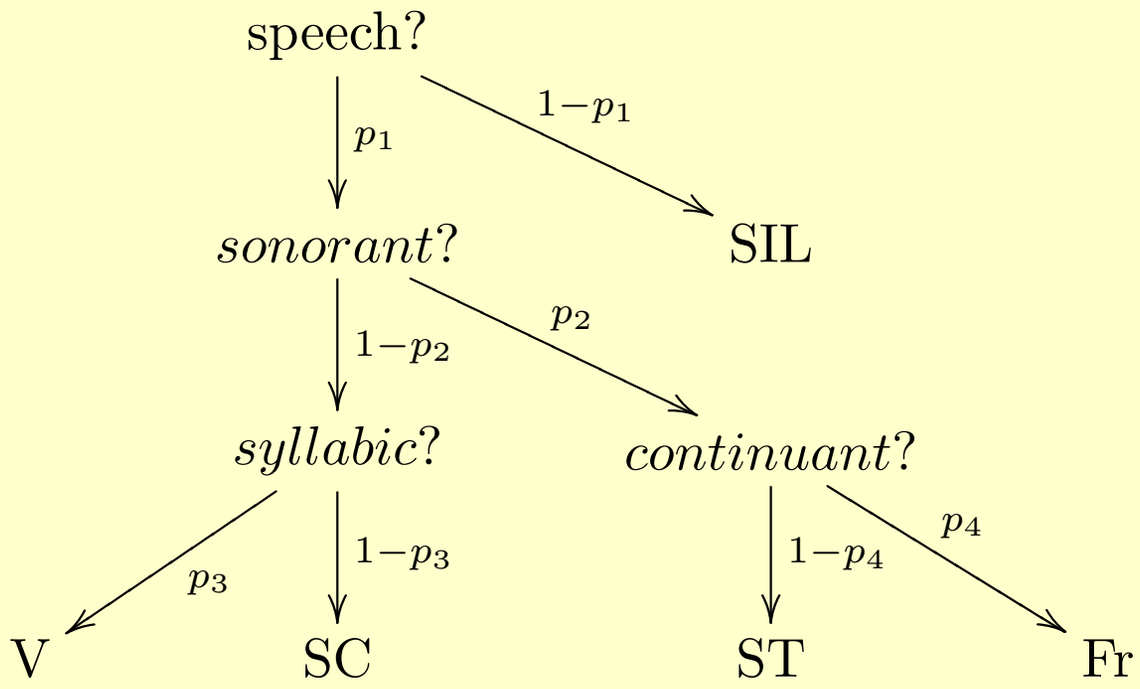
# 2. Acoustic features

- Multiscale spectro-temporal modulation features (Mesgarani, 2004)
- Long-window mel-frequency cepstral coefficients (MFCCs)
- Parameters extracted from speech on the basis of acoustic phonetics knowledge (Bitar and Espy-Wilson, 1996)
- Formants and related measurements like formant amplitude and bandwidth (Zheng and Hasegawa-Johnson, 2003)
- Very short window MFCCs (Hasegawa-Johnson et al 2004)

# 3. Probabilistic landmark detection

Event-based system (EBS) (Juneja and Espy-Wilson 2004)

- SVM-based classifiers for the manner phonetic features - *sonorant*, *syllabic*, *continuant* and *silence* - are applied in each frame of speech
- SVM outputs are converted to posterior probabilities using a histogram method
- The probabilities are combined to get probabilities of manner classes - vowel (V), sonorant consonant (SC), fricative (Fr), stop release (ST) and silence (SIL)
- A probabilistic segmentation algorithm (similar to Lee 1998) is applied to these probabilities to obtain the manner change landmarks

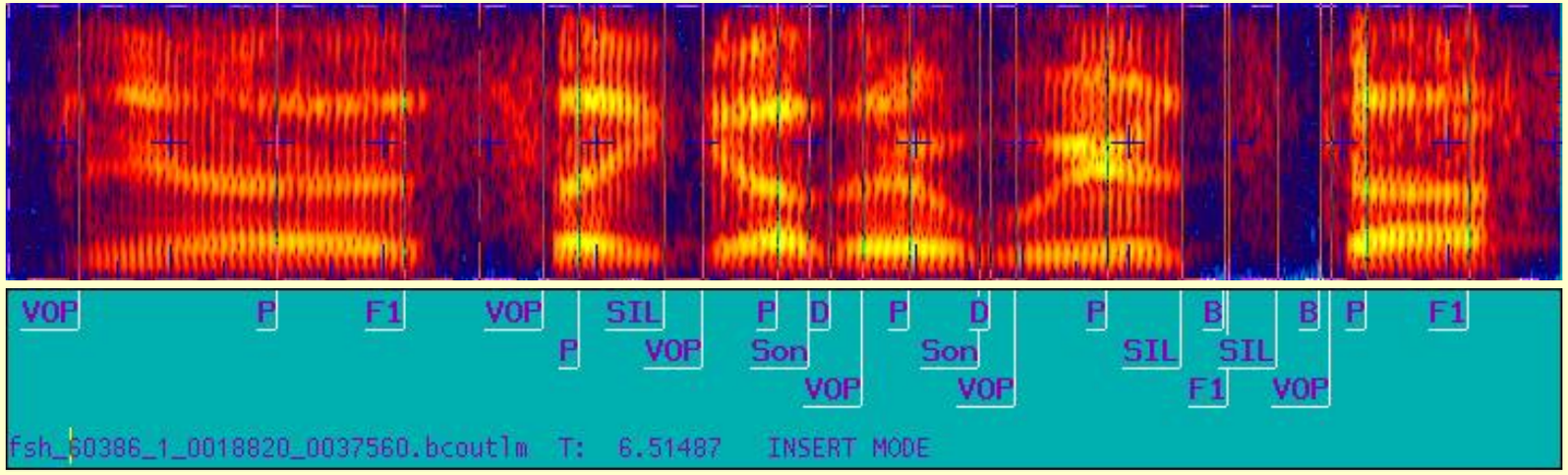



Probabilistic phonetic feature hierarchy

# Landmark output example

F1: fricative onset, Son: sonorant consonant onset, P: Vowel nucleus, D: syllabic dip, SIL: silence, B: stop burst, VOP: vowel onset point

"yeah it's like other weird stuff"



# 4. Place and voicing classification at landmarks

All results in percent

| Feature | Accuracy (before WS04) | Accuracy (now) |
|---|---|---|
| Stop Alveolar | 64.02 | 70.00 |
| Fricative Anterior | 75.71 | 77.14 |
| Nasal Labial | 67.30 | 81.18 |
| Rhotic for vowels | - | 88.16 |
| Flap | - | 86.2 |

# 5 (a) Unconstrained outputs

- Outputs of SVMs are supplied to the DBN based pronunciation model in two ways
- Method 1:
  Probabilities of the manner phonetic features are provided in each frame and the probabilities of place and voicing features are provided at landmarks obtained from the probabilistic segmentation algorithm
- Method 2:
  Each manner, place and voicing SVM is applied in each frame, and the probabilities of each phonetic feature is given to the DBN in each frame

# 5 (b) Rescoring using EBS

Feature bundle representation

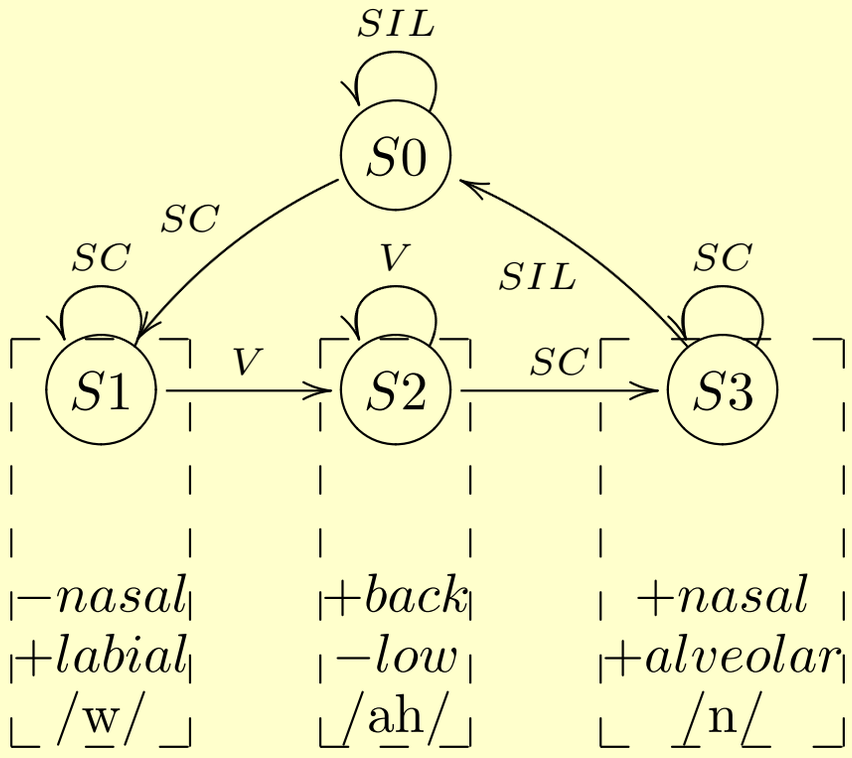
| /z/ | /I/ | /r/ | /o/ | /w/ |
|---|---|---|---|---|

$U \Rightarrow$

| $u_1$ | $u_2$ | $u_3$ | $u_4$ | $u_5$ |
|---|---|---|---|---|
| $-sonorant$ | $+sonorant$ | $+sonorant$ | $+sonorant$ | $+sonorant$ |
| $+continuant$ | $+syllabic$ | $-syllabic$ | $+syllabic$ | $-syllabic$ |
| $+strident$ | $-back$ | $-nasal$ | $+back$ | $-nasal$ |
| $+voiced$ | $+high$ | $+rhotic$ | $-high$ | $+labial$ |
| $+anterior$ | $+lax$ | | $+low$ | |

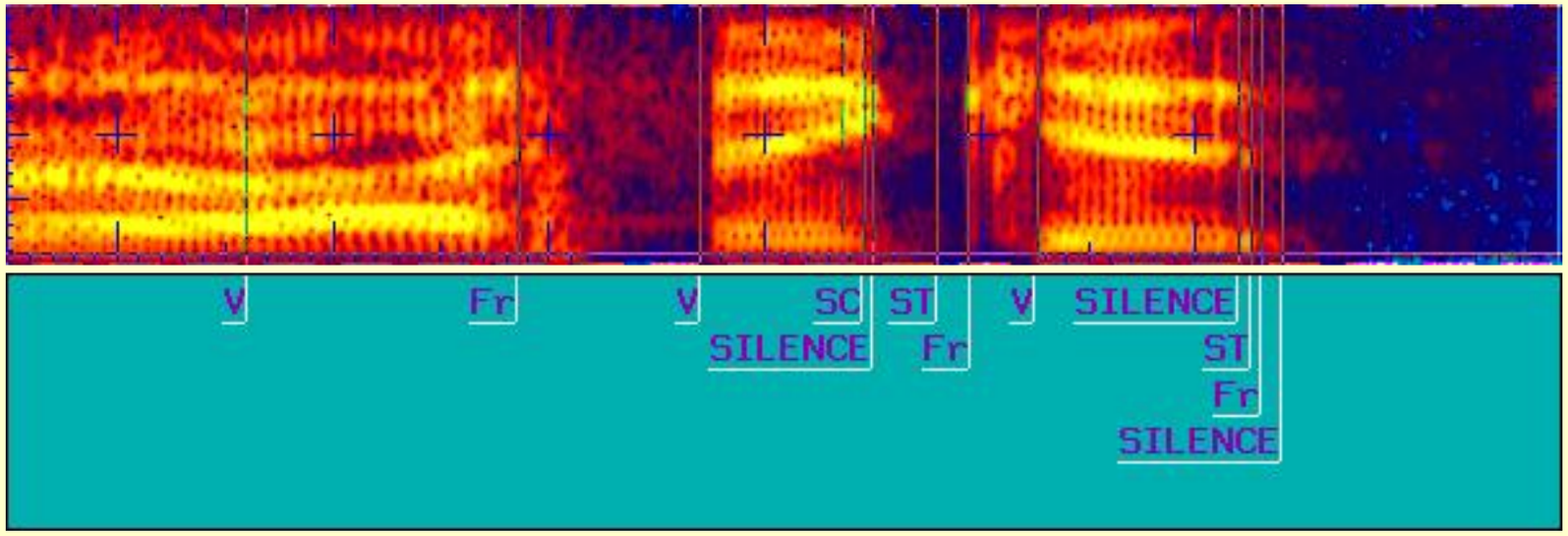
$L \Rightarrow$

| $l_1$ | $l_2$ | $l_3$ | $l_4$ | $l_5$ |
|---|---|---|---|---|
| Fricative onset | Vowel onset | SC onset | Vowel onset | SC onset |
| Fricative offset | Syllabic peak | Syllabic dip | Syllabic peak | Syllabic dip |
| | | SC offset | | SC offset |




Probability of a word is computed as $P(U/O) = P(L/O)P(U/LO)$, where $U$ : sequence of bundles of distinctive features, $L$ : sequence of landmarks, $O$ : acoustic observations

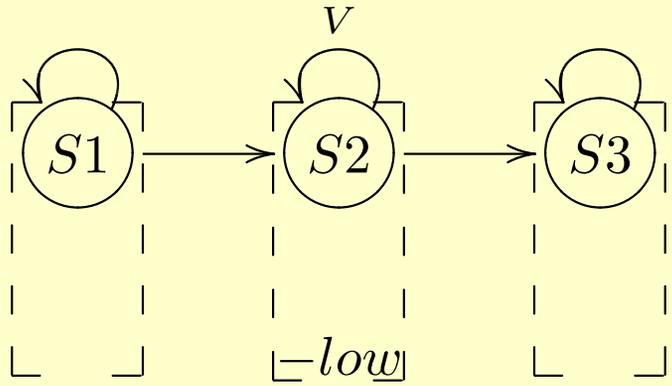
"i_think_it"




A stream weight of $10^{-5}$ has been obtained on RT03 development set, up from $-10^{-8}$

# 5 (c) Computation of score for a pair of features

- Detection can be constrained to get probabilities of a pair of features
- For example, a broad class vowel (V) and its place *low*



# SUMMARY

- Landmarks can be extracted from telephone speech with reasonable reliability using support vector machines
- Combination a variety of different acoustic observations and focus on training methods has significantly dropped the classification error on phonetic feature classification
- Further improvement is required in place of articulation classification
- Frame-based and landmark-based probabilities of phonetic features have been used with DBN-based pronunciation model, a discriminative pronunciation model and landmark based rescoring

# Future Work

A number of experiments are planned for the near future

- Improvement in phonetic feature classifiers
- Integration of prosodic features and syllable structure (Greenberg 2004)
- Assessment of the system in noise
- Comparison of feature classification accuracies with benchmark phoneme classification accuracies on telephone speech
- Modification of classifiers for better integration with DBN and MaxEnt model
- Improvement of probabilistic modeling for event-based rescoring