A Hierarchical Point Process Model for Speech Recognition

Aren Jansen **Partha Niyogi**

Department of Computer Science

ICASSP 2008

Motivating Principles

1. **Distinctive Feature vs. Phone-based Representation**
   - Articulatory motivation (e.g. [nasal] and [labial])
   - Acoustic motivation (e.g. [sonorant] and [continuant])
   - Compact phonological rules
   - Account for pronunciation variability

2. **Multiple Acoustic Front Ends**
   - One representation for each distinctive feature
   - Independently selected
   - Time scales and parameter sets may vary

Motivating Principles (cont.)

3. **Point Process Representation**
   - Construct asynchronous detector set, one for each feature
   - Determine set of landmarks for each feature
   - Sparse, point process representation
   - Inspired by selective neuron behavior

4. **Hierarchical Integration**
   - Sonority profile/vowel landmarks segment utterance
   - Syllable-sized analysis units
   - Construct a statistical model of detector firings in each segment

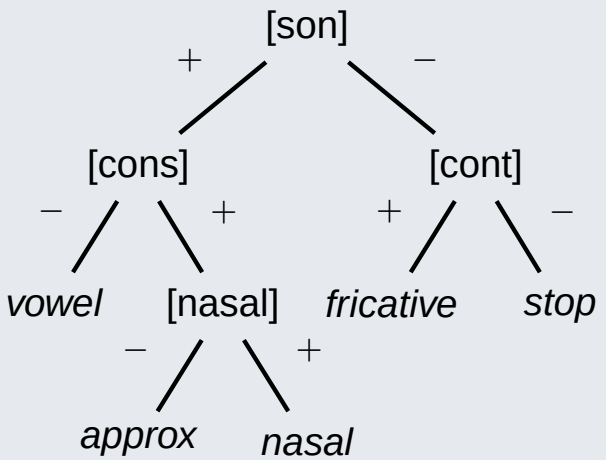
Related Work

- **Distinctive Features and Landmarks** [Stevens, 2002]
  - First to propose an approach to recognition involving hierarchy of distinctive features and context dependent processing at landmarks
  - Computational details not completely specified
  - Some reliance on rules for exceptional input
- **EBS and LBS** [Espy-Wilson et al., 2003 & 2007]
  - Uses similar feature hierarchy to ours
  - One classifier for each feature, all running on a common clock
  - Still a frame-based dynamic model
- **SUMMIT** [Glass, 2003]
  - Preprocessing determines a set of candidate phonetic segments and landmarks
  - Landmarks included in global MAP optimization
  - Temporal dynamics of landmarks not explicitly modelled

Distinctive Feature Representation

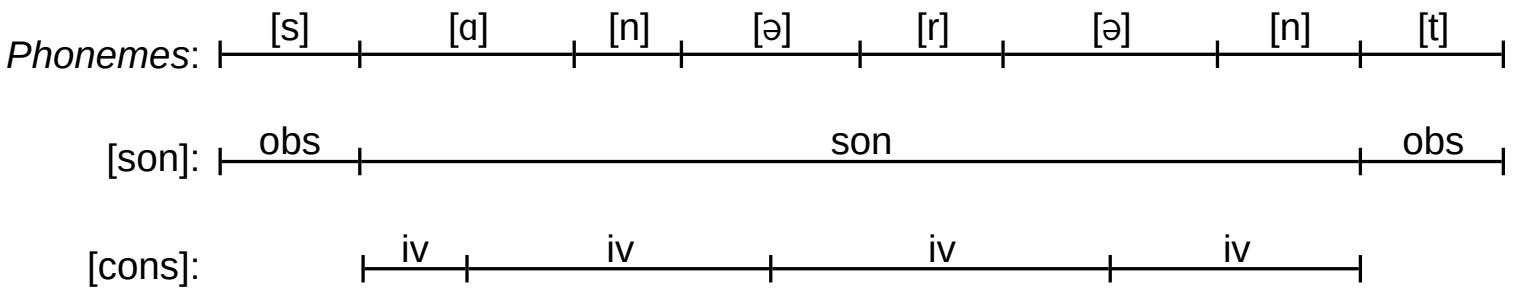
- Each phoneme may be represented as a vector of binary-valued distinctive features
- Features have hierarchical structure:
  - Features higher up in tree are more fundamental/less context dependent
  - Distinct subtrees may be largely independent from each other

Our Hierarchy



- [son]: distinguishes sonorant and obstruent sounds
- [cons]: distinguishes consonant and vowel sounds.
- [cont]: distinguishes stop consonants from everything else
- [nasal]: distinguishes those sounds that couple the nasal cavity from the rest

Decomposition of the Signal



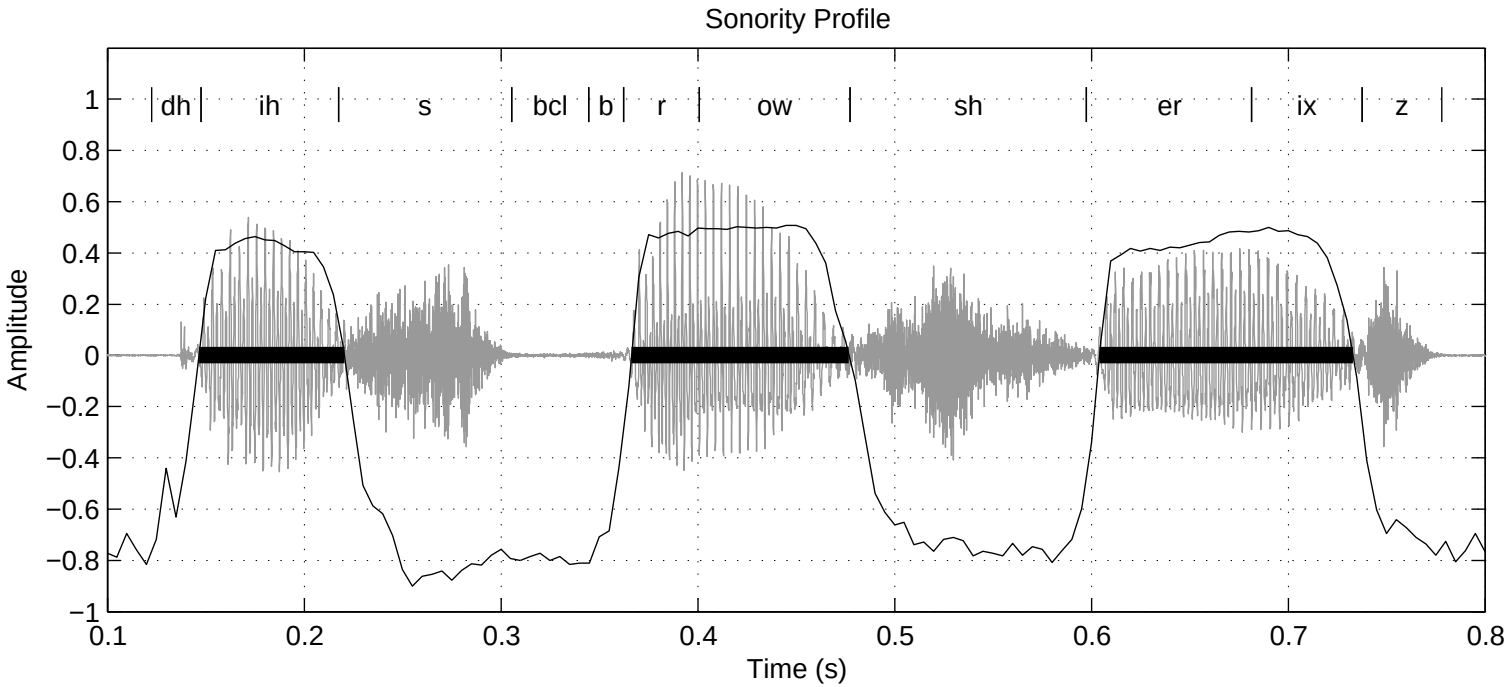
1. Segment utterance into sonorant and obstruent regions
2. Segment sonorant regions about vowel landmarks into sonorant intervocalic regions

Key Points

- Assume segments are independent according to feature hierarchy
- Independent analysis at syllabic time scales
- In TIMIT, only 42/12 possible intervocalic obstruent/sonorant broad class sequences

Sonority Segmentation

- Construct an SVM to classify frames as sonorant or obstruent
- Threshold SVM output to determine segmentation



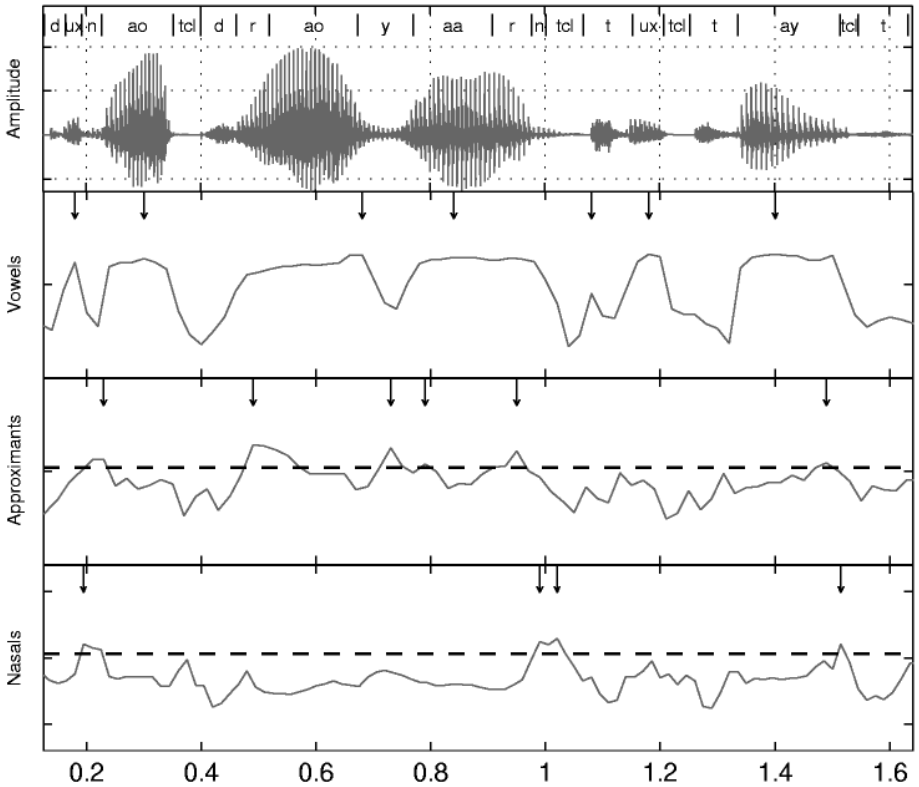
Distinctive Feature Detectors

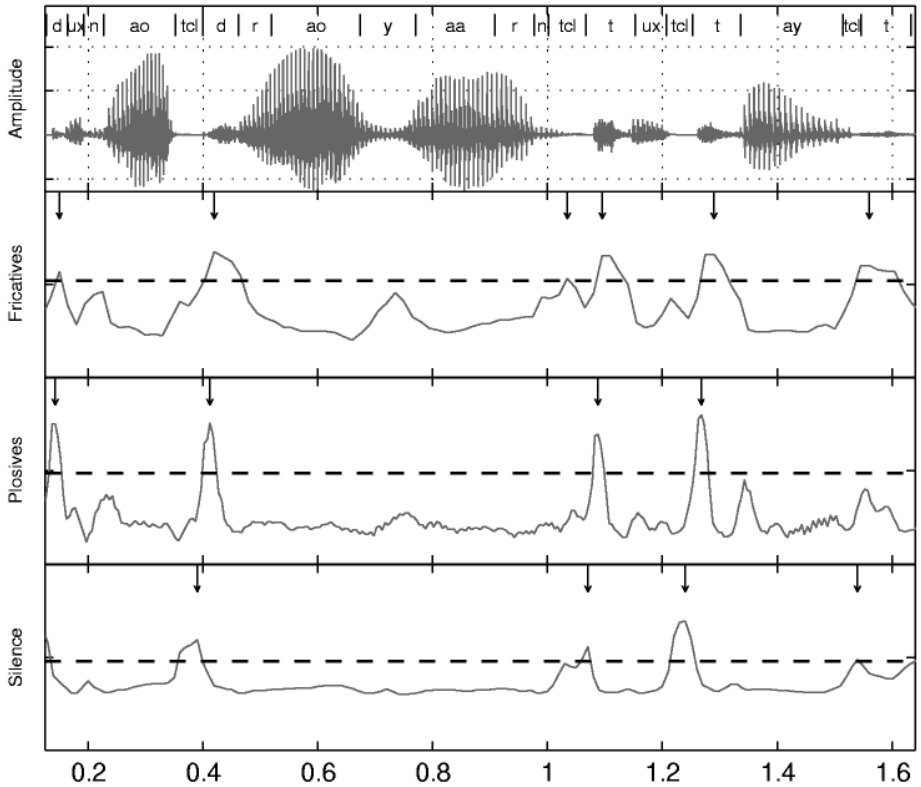
- One detector for each leaf node in feature hierarchy (+silence)
  - Leaf nodes correspond to broad (manner) classes: vowel, approximant, nasal, fricative, stop (plosive)
  - Input: specialized representation
  - Output: a set of landmarks (time-strength pairs)
- Landmarks interpreted as most articulated/perceptually relevant points of signal for each feature

Our Implementation

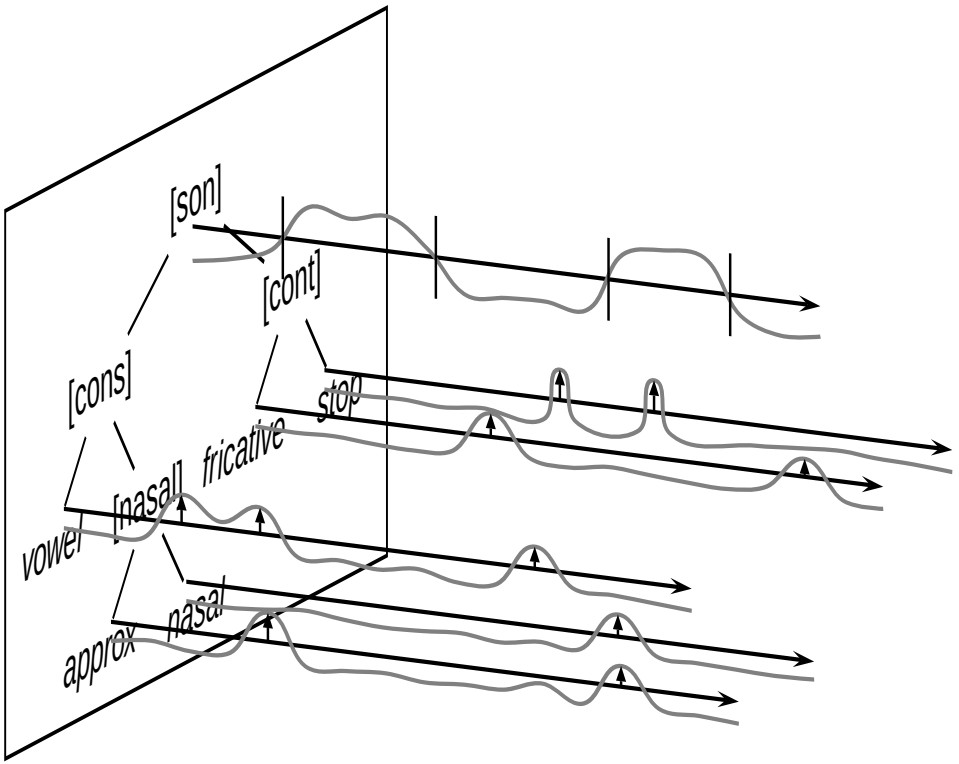
- Construct one SVM classifier for each broad class
- Threshold and pick local maxima as landmarks

Feature Detectors: Example





Timing Tier Representation



- Hierarchically structured marked point processes
- Sparse (not frame-based) and asynchronous
- Assume linguistic content is encoded in temporal dynamics of point process representation
- Feature hierarchy implies intervocalic sonorant and obstruent segments are independent

The Goal

Map this timing tier representation into a broad class sequence.

Sonority Segment Decoding (SSD)

Definition

Given segment (T_1, T_2) of duration $T = T_2 - T_1$ and N detectors

- $O = \{T, O_1, \ldots, O_N\}$ are observables
  - $O_i = \{(t_1^{(i)}, f_1^{(i)}), \ldots, (t_{L_i}^{(i)}, f_{L_i}^{(i)})\}$ are class C_i landmarks
  - $t_k^{(i)} \leftarrow (t_k^{(i)} - T_1)/T =$ normalized time, $f_k^{(i)} =$ strength
- $H = \{H_1, \ldots, H_N\}$ are landmark indicator variables, where
  - $H_i = \{h_1^{(i)}, \ldots, h_{L_i}^{(i)}\}$ are class C_i variables
  - $h_k^{(i)} = \begin{cases} 1 & \text{if } k^{\text{th}} \text{ detection of class } C_i \text{ is a true positive} \\ 0 & \text{otherwise} \end{cases}$
- $B \in \Sigma^*$ where $\Sigma = \{V, A, N, P, F, sil\}$

- Segment-level MAP Estimate:

$$(B_{\text{opt}}, H_{\text{opt}}) = \arg\max_{B,H} P(B, H|O)$$
$$= \arg\max_{B,H} P(O|B, H)P(H|B)P(B)$$

Sonority Segment Decoding (cont.)

- For our experiments, we assume independent detectors, correctness patterns, landmarks, times and strengths:

$$(B_{\text{opt}}, H_{\text{opt}}) = \arg\max_{B,H} P(T|B)P(B)\times$$
$$\prod_{i=1}^{N}\left[P(H_i|B)\prod_{k=1}^{L_i}P(t_k^{(i)}|B,h_k^{(i)})P(f_k^{(i)}|B,h_k^{(i)})\right]$$

Model Training

- Histogram method used to estimate discrete variable distributions.
- Uniform kernel density estimation used for scalar variable distributions (introduces kernel bandwidths).

Sonority Segmentation and Detector Performance

Front End Construction

- *Sonority Segmentation:* 39-dim MFCCs (0-8kHz), 10/5 ms window/step size
- *Feature Detectors:* 39-dim MFCCs w/ varying construction; stop detector: E[0-8kHz], E[3-8kHz], Wiener entropy

- Segmentation performance:
  - 6.44% frame-level test error
  - 95.0% sonorants & 89.3% obstruents in correct segment
- Feature Detector Performance:

Detector	Window	Step	Freq. Range	Error Rate
Vowel	40 ms	20 ms	0-4 kHz	15.1%
Approx.	20 ms	20 ms	0-8 kHz	28.2%
Nasal	30 ms	15 ms	0-8 kHz	10.0%
Fricative	30 ms	15 ms	0-8 kHz	11.4%
Stop	35 ms	5 ms	N/A	17.4%
Silence	20 ms	10 ms	0-8 kHz	7.8%

Sonority Segment Decoding Performance

Experimental Setup

- Train/test using actual segments from transcription
- 1000 training sentences, all 1344 test sentences

	Obstruent			*Intervocalic*		
Method	*Acc*	*Corr*	*Ins*	*Acc*	*Corr*	*Ins*
Baseline	42.0	79.2	37.2	25.5	54.0	28.5
Standard	77.0	83.0	6.0	53.0	69.9	16.9
Rank ≤ 2	89.2	92.1	2.9	85.1	90.4	5.3
Rank ≤ 3	93.8	94.7	0.9	95.1	96.8	1.7

- SSD maintains correctness while reducing insertions
- N-best estimates exceedingly good (language model promising)

Overall Performance: Setup

Our Experimental Setup

- Train/test using all system components
- 1000 training sentences, all 1344 test sentences

HMM Experimental Setup

- CMU Sphinx-3 HMM system
- Test both context dependent (CD) and context independent (CI) decoding
- Test both broad class (BC) and phoneme (Ph) 3-state models
- 39-dim MFCCs, 8 Gaussian mixture observation densities, no skip transition
- 3696 training sentences, 1344 test sentences
- *No language model*

Overall Performance: Results

System	*Accuracy*	*% Correct*	*% Ins*	*% Del*	*% Sub*
Our System	70.3	76.0	5.7	11.3	12.7
HMM, CI/BC	65.5	68.4	2.9	17.6	13.9
HMM, CD/BC	65.1	90.5	25.4	1.4	8.1
HMM, CI/Ph	69.0	79.7	10.7	6.0	14.2
HMM, CD/Ph	72.2	91.5	19.3	1.6	6.9

- Our accuracy exceeds 3 of 4 HMM variants
- SSD is partially CD (model short phone sequences, but no inter-segmental dependence)
- Our complexity is most similar to CD/BC HMMs
- Our system is a conservative guesser, largely due to detector thresholding.
- Significant room for improvement of our implementation

Phonetic Recognition: Preliminary Results

Experimental Setup

- *Hidden Markov Model* [Sha & Saul, 2007]
  - 39-dim MFCC front-end (25/10 ms window/step)
  - 1 state/phone, EM-trained C-comp GMM for each state
  - Transition probabilities estimated from frame-level counts
- *Poisson Process Model* [Jansen & Niyogi, 2008]
  - Identical front end as HMM
  - Phone detector set computed from GMM emit probabilities
  - One inhomogeneous Poisson process model for each possible sonorant consonant (61) or obstruent (385) phone sequence

Recognition Accuracies

	Obstruent		*Intervocalic*	
C	*Poisson*	*HMM*	*Poisson*	*HMM*
1	56.6	51.1	72.4	60.5
2	60.3	57.5	75.2	64.3
4	62.5	61.3	76.2	67.8
8	63.2	63.3	78.0	70.7

Future Directions

1. **Full Phonetic Recognition**
   - Phone models
   - Context-dependent transcription refinement at landmarks
2. **Alternative Point Process Modelling**
   - Inhomogeneous Poisson process model
   - Spiking neural networks
3. **Improving System Robustness**
   - Adaptive thresholds
   - Acoustic parameter front end
   - Neurobiologically plausible detector sets
4. **Probabilistic Segmentation**