Detection-Based Speech Recognition with Sparse Point Process Models

Aren Jansen

Human Language Technology Center of Excellence

JOHNS HOPKINS UNIVERSITY

Partha Niyogi

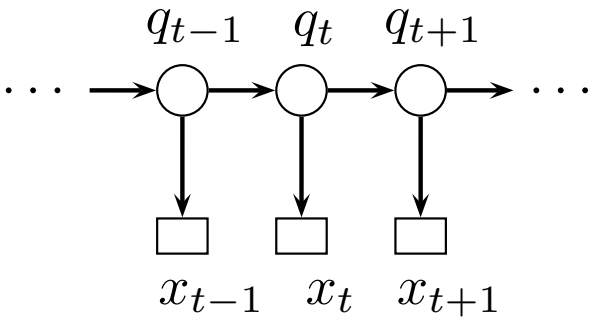
Departments of Computer Science and Statistics

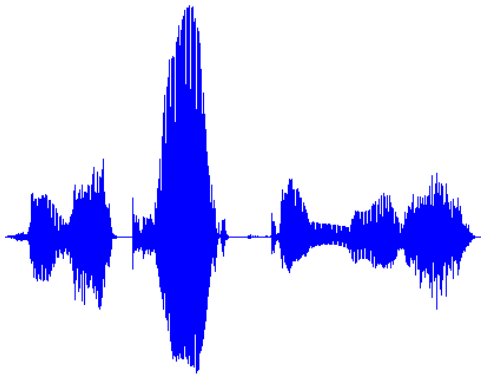
THE UNIVERSITY OF CHICAGO

ICASSP 2010

Dallas, Texas

Are Frames the Optimal Level of Detail?





. . . six zero six
three seven . . .

?

A Unified Event-Driven Approach

Our Strategy: Only model and explain the portions of the signal we are reasonably confident about

Point Process Models (PPM) [Jansen & Niyogi (2009)]

1. Transform the signal into sparse temporal point patterns of acoustic events
2. Decode linguistic objects according to the temporal statistics of these events

Detection-Based ASR Architecture [Ma, Tsao, & Lee (2006)]

1. Run independent detectors for each word in lexicon in parallel
2. Extract word sequence from the combined detector set output

Our Goal: Translate past robustness success of point process word modeling to standard small vocabulary task

The AURORA2 Evaluation

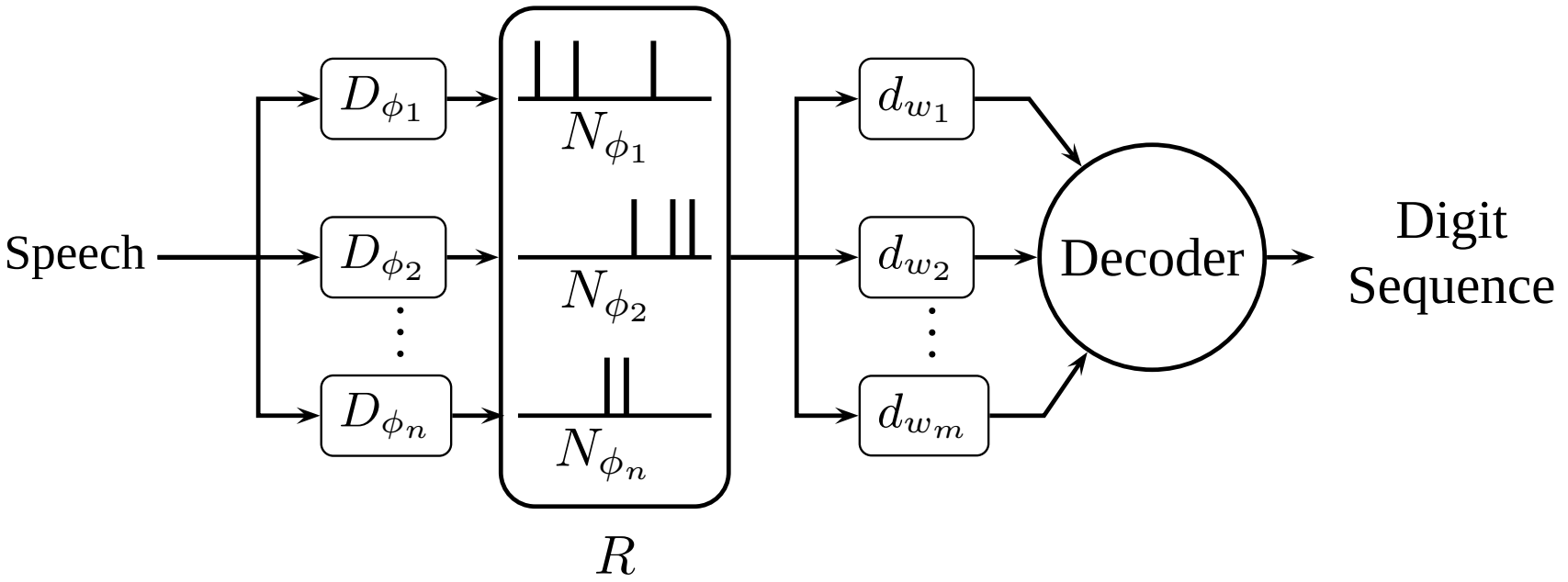
AURORA2 Task

- Spoken digit sequences (8 kHz) both clean and mixed with additive noise at $\{20, 15, 10, 5, 0, -5\}$ dB SNR
- Stationary (mostly): subway, car, exhibition hall, and street noise
- Non-stationary: babble, restaurant, airport, and train station
- We consider clean training evaluation

Baseline HTK 3.4 Recognizer

- MFCCs computed with AURORA Front-End v2.0 (plus vel., acc.)
- 11 digit models $\times$ 16 states/model (left-to-right, no skip transitions) $+$ 3 silence states = **179 states**
- 3 GMM components per digit state, 6 GMM components per silence state

PPM-Based ASR Architecture



Definitions

- D_{ϕ_i} = detector for feature ϕ_i
- N_{ϕ_i} = point pattern (event set) for feature ϕ_i
- d_{w_j} = detector for word w_j

Hidden State Feature Detectors, D_{ϕ_i}

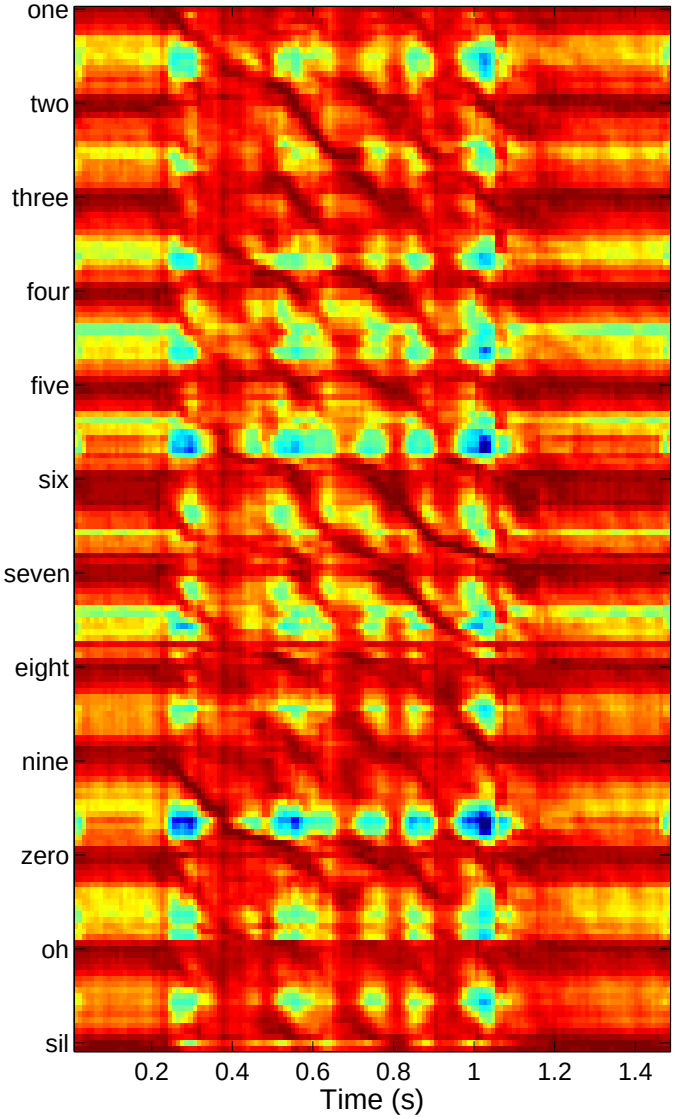
1. Define one feature ϕ_i for each of the 179 HMM states
2. Define detector function for each ϕ_i:

$$g_{\phi_i}(x) = P(\phi_i|x) = \frac{P(x|\phi_i)P(\phi_i)}{\sum_{i=1}^{179} P(x|\phi_i)P(\phi_i)}$$

3. Threshold g_{ϕ_i} at δ_{ϕ_i} and pick local maxima times as acoustic events for feature ϕ_i: $N_{\phi_i} = \{t_1, t_2, \dots\}$

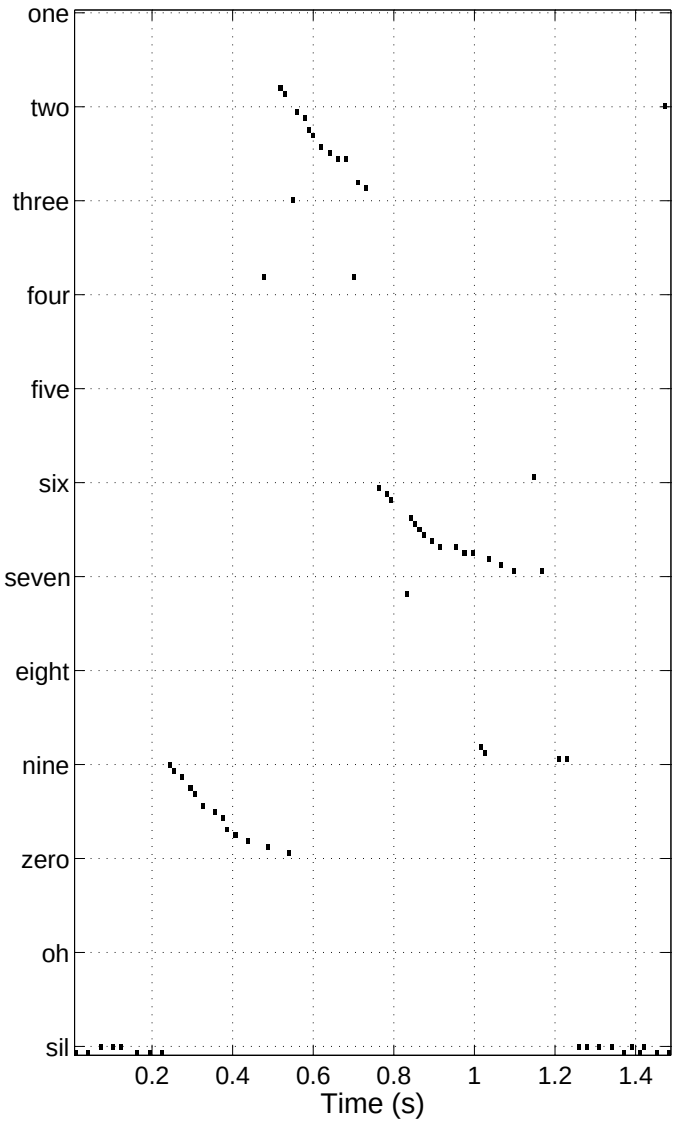
Point Process Representation Example

HTK Lattice: 926



26134 real-valued likelihoods
(179 states $\times$ 146 frames)

Point Process Representation: 926



69 real-valued times

Sliding Model Word Detectors, d_{w_j}

1. Let $\theta_w : \mathbb{R} \rightarrow \{0, 1\}$ be indicator function of word occurrence
2. Define LLR detector function $f_w(t) = \log \left[\frac{P(R|\theta_w(t)=1)}{P(R|\theta_w(t)=0)} \right]$
3. Introduce duration latent variable T:

$$P(R|\theta_w) = \int P(R|T, \theta_w) P(T|\theta_w) dT$$

4. Partition R into three subsets: $R_l = R|_{(0,t]}$, $R_{t,T} = R|_{(t,t+T]}$, and $R_r = R|_{(t+T,L]}$. Then,

$$f_w(t) = \log \int \frac{P(R_{t,T}|T, \theta_w(t)=1)}{P(R_{t,T}|T, \theta_w(t)=0)} P(T|\theta_w(t)=1) dT.$$

Word Model, $P(R_{t,T}|T, \theta_w(t)=1)$

Inhomogeneous Poisson Process Definition

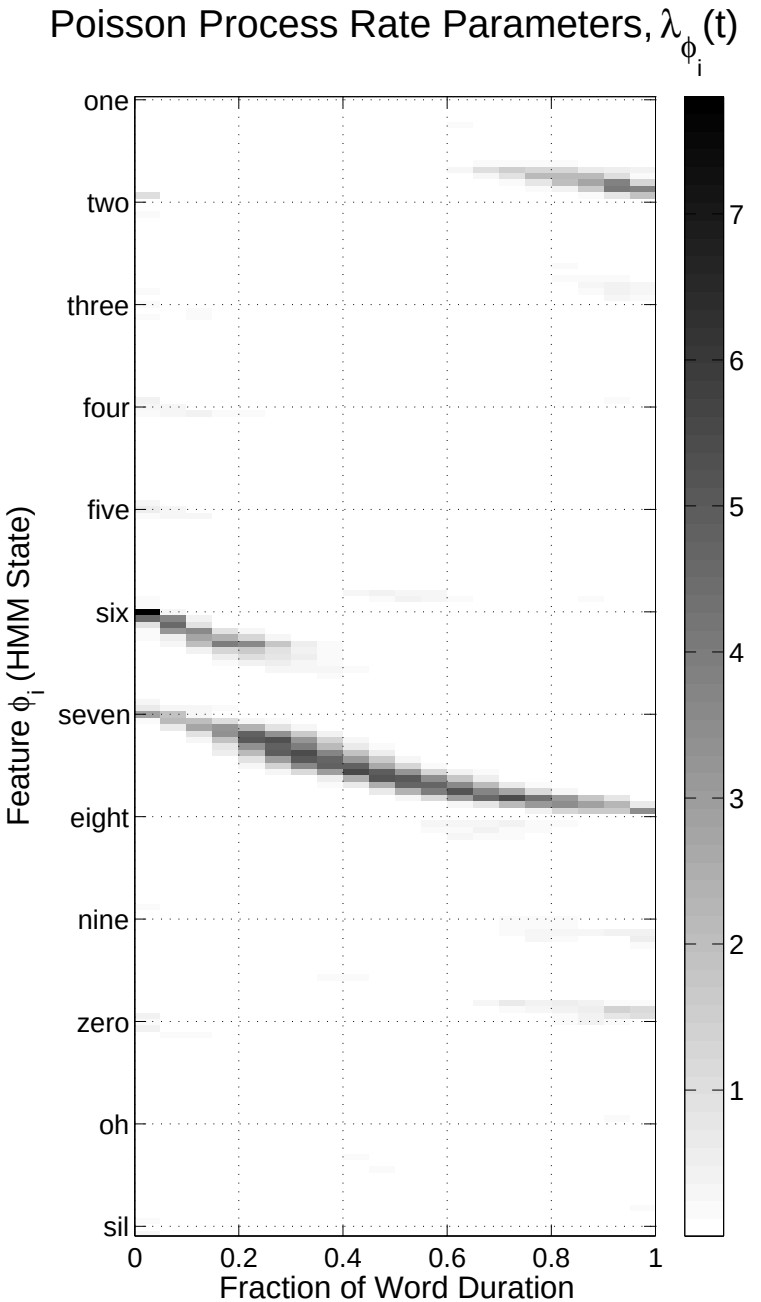
Memoryless point process with feature ϕ_i arrival probability $\lambda_{\phi_i}(t)dt$ in differential time element dt at time t

1. Normalize all $t \in R_{t,T}$ to the interval $[0, 1]$, yielding $R' = \{N'_{\phi_i}\}_{i=1}^{179}$
2. Assume T-independence of R', independent feature detectors, and inhomogeneous Poisson process model for each N'_{ϕ_i}:

$$P(R_{t,T}|T, \theta_w(t)=1) = \frac{1}{T^{|R_{t,T}|}} \prod_{i=1}^{179} e^{-\int_0^1 \lambda_{\phi_i}(s)ds} \prod_{s \in N'_{\phi_i}} \lambda_{\phi_i}(s),$$

3. Rate functions $\{\lambda_{\phi_i}\}_{1=1}^{179}$ are estimated with parametric model or KDE (examples from HMM force-align)

Example: “seven” Poisson Process Model



Background Model, $P(R_{t,T}|T, \theta_w(t)=0)$

Homogeneous Poisson Process Definition

Memoryless point process with **constant** feature ϕ_i arrival probability $\mu_{\phi_i} dt$ in any differential time element dt

1. No interval normalization necessary
2. If n_{ϕ_i} is the number of events of type ϕ_i in $R_{t,T}$, then

$$P(R_{t,T}|T, \theta_w(t)=0) = \prod_{i=1}^{179} [\mu_{\phi_i}]^{n_{\phi_i}} e^{\mu_{\phi_i} T},$$

3. Background rate parameters $\{\mu_{\phi_i}\}_{i=1}^{179}$ are estimated by counting in arbitrary background speech

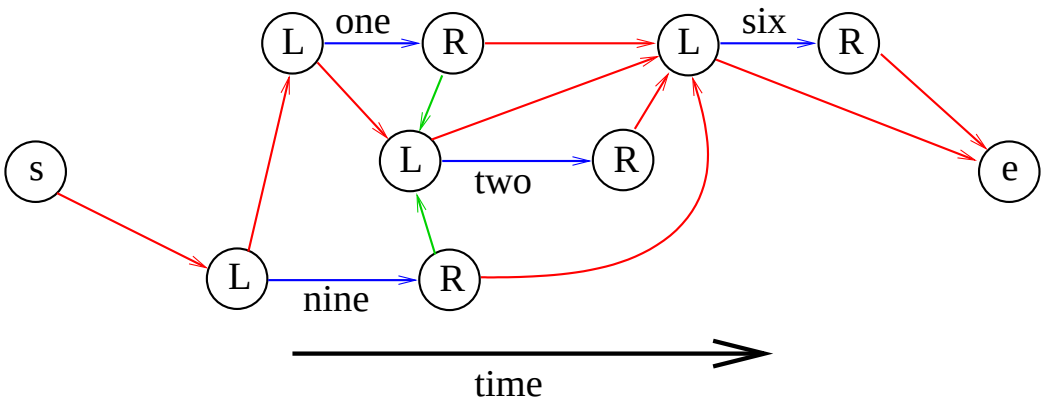
Graph-Based Decoder [Ma, Tsao, & Lee (2006)]

Input: Digit detectors produce candidate detect set, along with confidence scores (f_w) and durations ($\arg\max_T$ of integrand)

Decoder DAG Definition

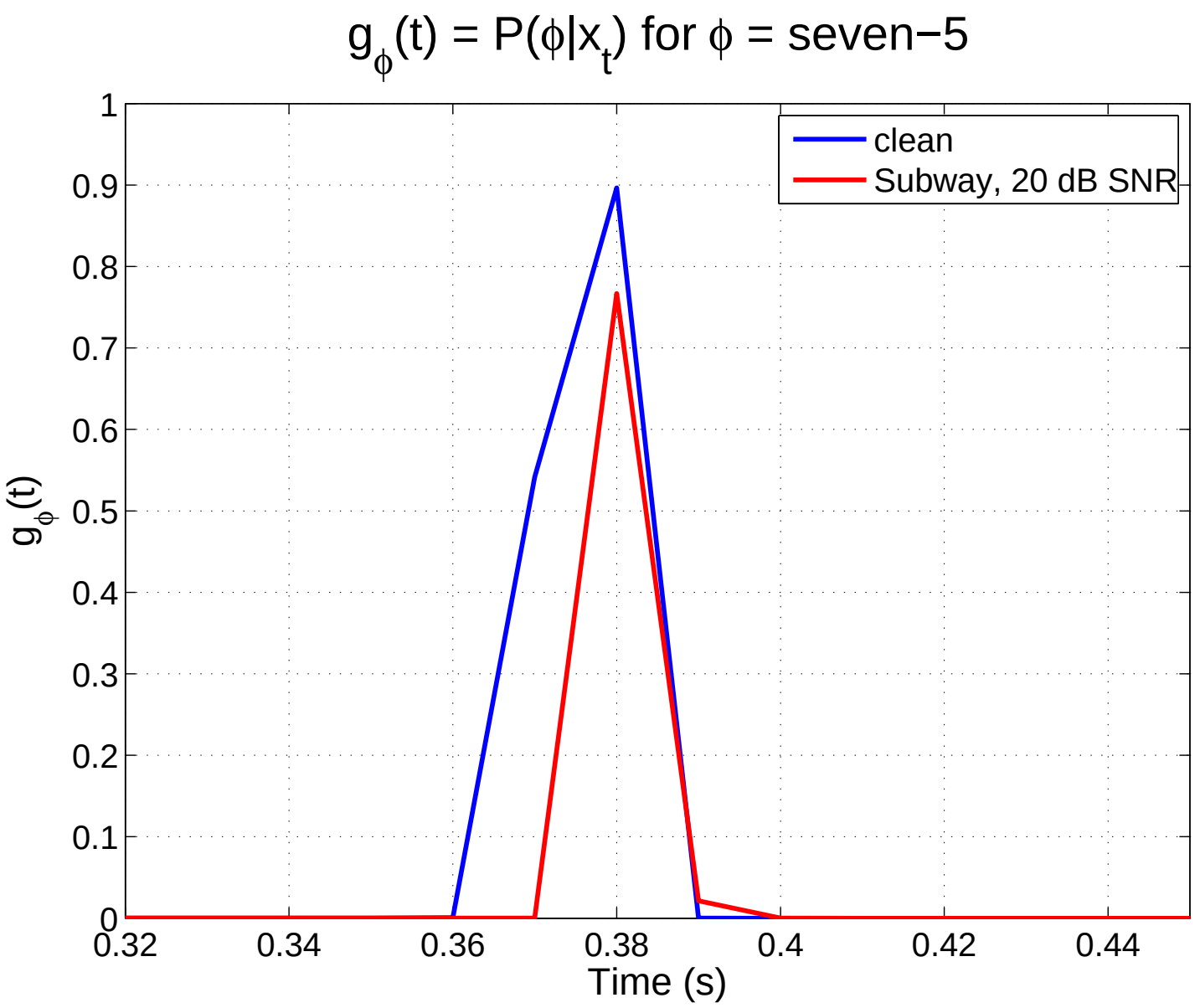
Vertices: *start* at $t = 0$, *end* at $t = \infty$, two for each digit detect (left and right boundary)

1. Connect each vertex to next left boundary vertex with weight 0
2. Connect each left boundary vertex to its right boundary vertex with weight $-f_w(t)$
3. Connect each right boundary vertex to *all* left boundaries within 20 ms prior with weight 0 (no cycles)



Decode: Min-cost path from *start* to *end* with Dijkstra's algorithm

What About Robustness?

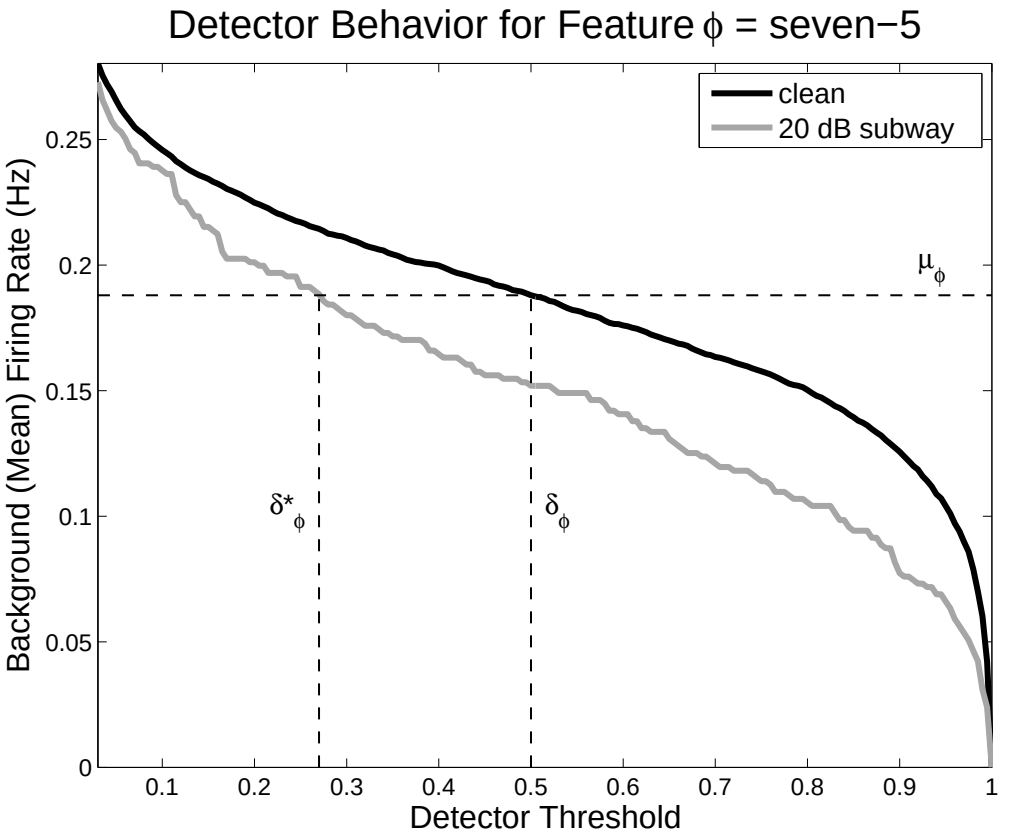


Feature Detector Threshold Adaptation

1. Find feature detector threshold $\delta^*_{\phi_i}$ that maintains background firing rate from clean speech
2. Use clean word/background models with adapted phone detector threshold

Underlying Assumptions

1. Times/relative strengths of local maxima preserved
2. Background rate is adequate statistic



This method is entirely **unsupervised**

Clean Speech Performance

HTK (% Acc)	**PPM** (% Acc)
99.0	98.3

- Only 0.7% WER increase after a $\sim$400X reduction in representational data
- Possible explantion: forced aligned digit training examples were imperfect

Non-Stationary Noise Performance

Train: Clean, Test: Babble

SNR	HTK	PPM	Adapt PPM
20 dB	90.2	92.2	93.6
15 dB	73.8	83.1	89.7
10 dB	49.4	65.0	80.2
5 dB	26.8	43.8	62.5
0 dB	9.3	22.3	35.8
-5 dB	1.6	10.0	16.4
Avg. (0-20)	49.9	61.3	72.4

Train: Clean, Test: Airport

SNR	HTK	PPM	Adapt PPM
20 dB	90.6	92.6	94.3
15 dB	77.0	84.3	90.4
10 dB	53.9	68.8	82.2
5 dB	30.3	46.0	66.7
0 dB	14.4	24.1	40.6
-5 dB	8.2	11.5	20.0
Avg. (0-20)	53.2	63.2	74.8

- Non-adapted PPM system is significantly more robust than the HMM system to non-stationary noise
- Unsupervised feature detector threshold adaptation provides further gains

Stationary Noise Performance

Train: Clean, Test: Subway

SNR	HTK	PPM	Adapt PPM
20 dB	97.1	94.1	94.6
15 dB	93.5	86.0	89.6
10 dB	78.7	67.4	79.4
5 dB	52.2	40.4	61.3
0 dB	26.0	18.8	34.6
-5 dB	11.2	8.5	15.6
Avg. (0-20)	69.5	61.3	71.9

Train: Clean, Test: Car

SNR	HTK	PPM	Adapt PPM
20 dB	97.4	94.6	95.1
15 dB	90.0	85.9	89.9
10 dB	67.0	63.0	76.8
5 dB	34.1	33.3	57.1
0 dB	14.5	13.6	29.0
-5 dB	9.4	6.5	12.6
Avg. (0-20)	60.6	58.1	69.6

- Non-adapted PPM less robust than HTK system to stationary noise
- Suboptimal feature detector threshold is culprit
- Unsupervised threshold adaptation improves robustness over HTK levels at lower SNRs

Conclusions

1. Discarding 99.7% of the HMM lattice results in negligible loss in small vocabulary recognition accuracy
2. Sparse point process word models + detection-based ASR architecture improves robustness to all non-stationary noise sources in AURORA2
3. Unsupervised PPM adaptation (only 1 minute of data) improves robustness to all noise sources
4. Our system is compatible with other noise robustness techniques (both front end and GMM adaptation)
5. Sparse point process representations may supply the computational efficiency required to scale up detection-based ASR systems