Robust Keyword Spotting with Rapidly Adapting Point Process Models

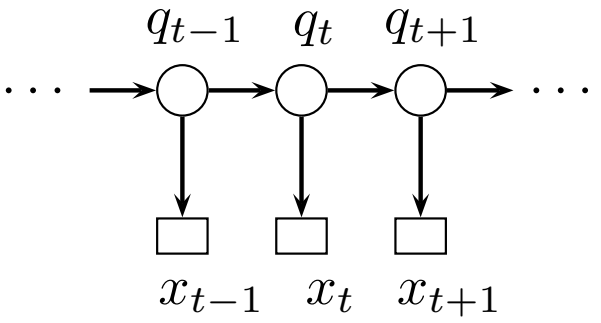
Aren Jansen[1,2] **Partha Niyogi**[1]

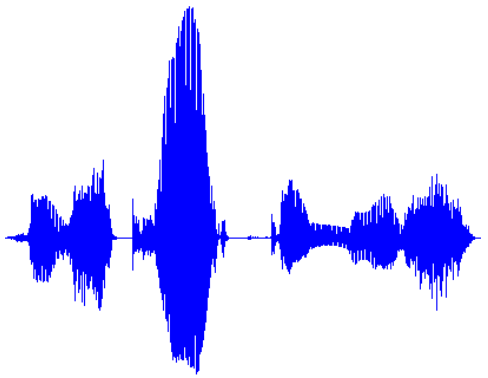
1) *Department of Computer Science*

2) *Human Language Technology Center of Excellence*

Interspeech 2009
Brighton, UK

Are Frames the Optimal Level of Detail?





. . . sh iy hh ae
dcl d y er . . .

?

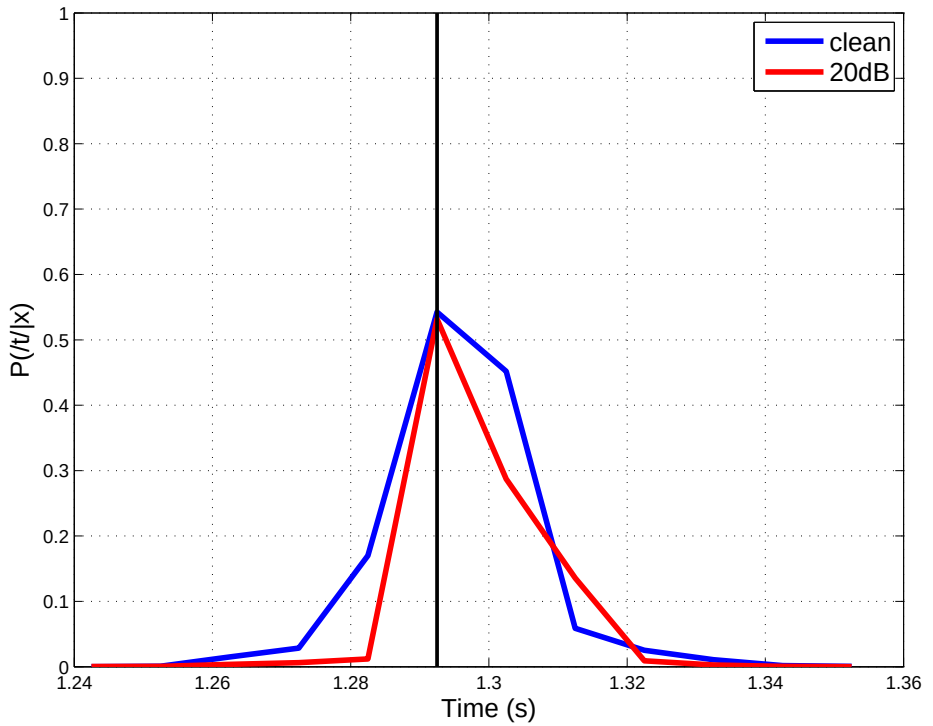
Point Process Models (PPMs)

We consider alternative dynamic models that involve:

1. Transforming the speech signal into **temporal point patterns** of acoustic events that occur
2. Decoding the underlying linguistic message according to the **temporal statistics** of these events

Engineering Benefits

- Point process representation automatically accommodates stream asynchrony
- Temporal coding strategy opens the field up to a large class of alternative statistical approaches
- Temporally sparse representations can reduce computation and increase robustness



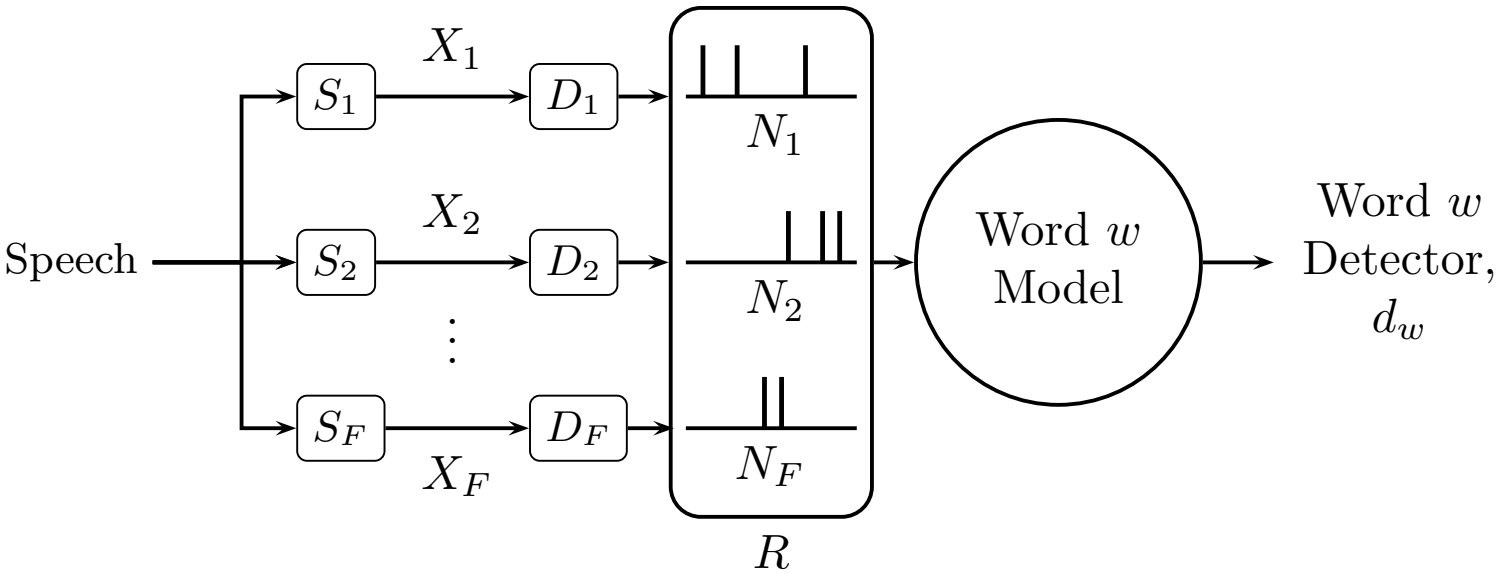
The Keyword Spotting (KWS) Task

The Goal

Given a predefined keyword and a number of training instances, locate all occurrences in a novel utterance with minimal false alarms

- Prevailing frame-based paradigm is the keyword-filler HMM
- Performance metric is the KWS figure-of-merit:
  - FOM $\equiv$ Avg. detection rate allowing $1, 2, \ldots, 10$ false alarms/hr
- Used for audio indexing, command word detection, small vocabulary recognition, and long-range strategies for large vocabulary recognition

PPM-Based Keyword Spotting Architecture



Definitions

- $S_i =$ signal processor for feature i
- $X_i =$ acoustic representation for feature i
- $D_i =$ detector for feature i
- $N_i =$ point pattern (landmark set) for feature i

Phone Detectors

1. Use 39-dim mel-frequency cepstral coefficient (MFCC) front end that defines observation vector space $X = \mathbb{R}^{39}$
2. Build 8-comp. Gaussian mixture model for each phone $p \in \mathcal{P}$:

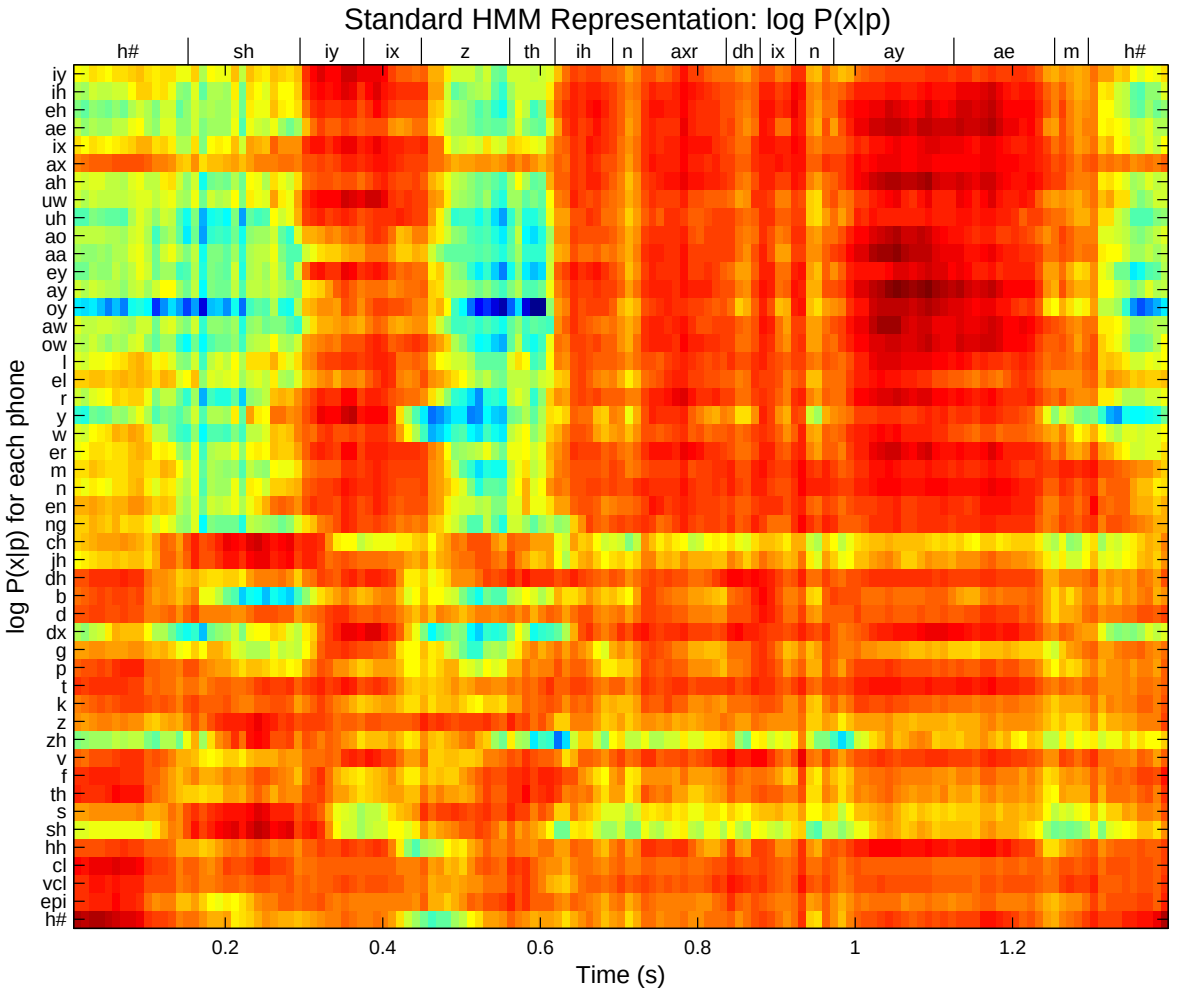
$$P(x \in X|p) = \sum_{c=1}^{8} \omega_{pc} \mathcal{N}(\vec{\mu}_{pc}, \mathbf{\Sigma}_{pc})(x)$$

3. Define detector function for each p:

$$g_p(x) = P(p|x) = \frac{P(x|p)P(p)}{\sum_{p \in \mathcal{P}} P(x|p)P(p)}$$

4. Threshold g_p at δ_p and pick local maxima times as acoustic events for phone p: $N_p = \{t_1, t_2, \dots\}$

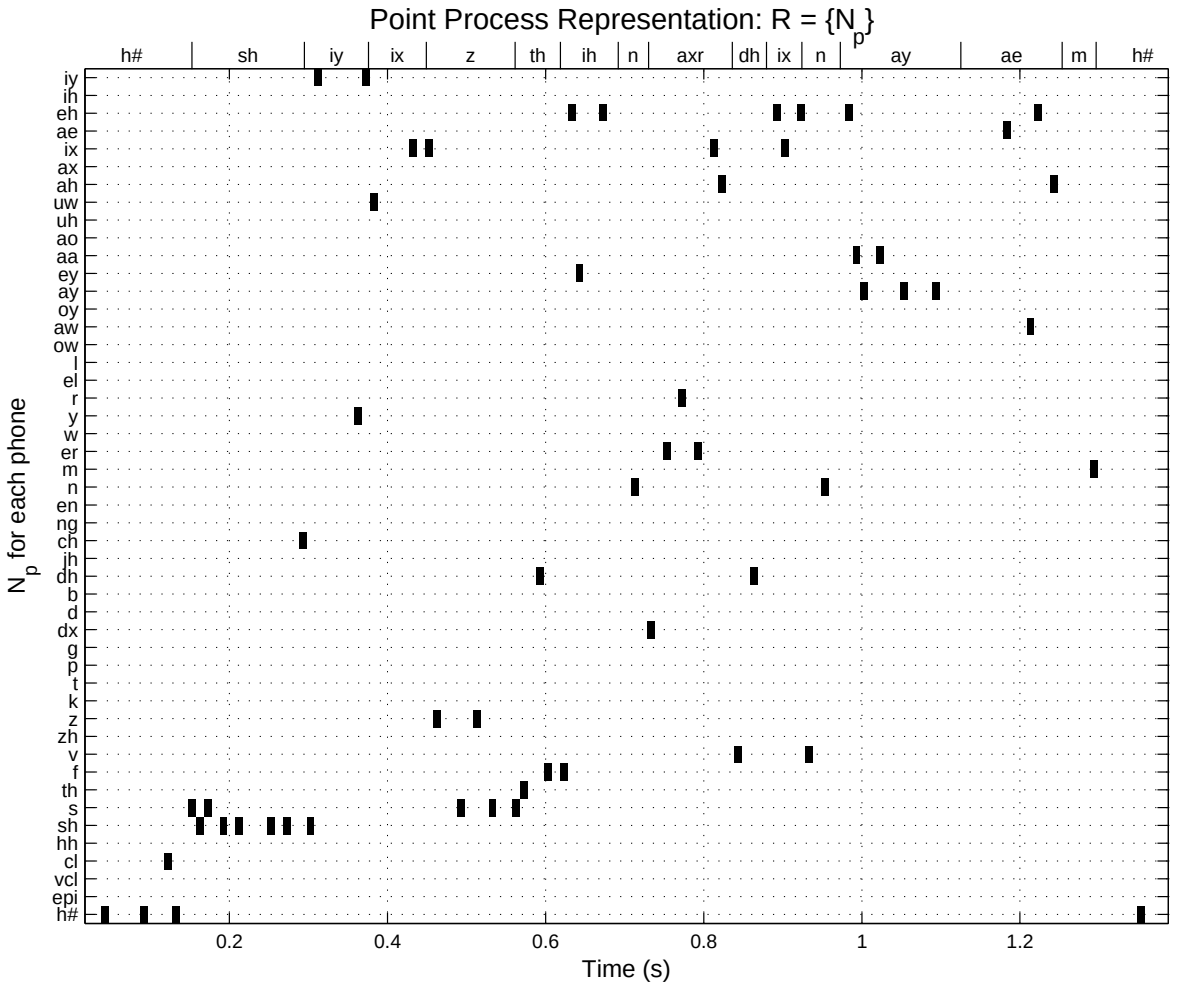
Example: Phone GMM Probability Lattice



"She is thinner than I am."

6672 real-valued likelihoods (48 states $\times$ 139 frames)

Example: Phone Point Process Representation



"She is thinner than I am."

57 real-valued times

Sliding Model Keyword Detectors

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

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

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

Inhomogeneous Poisson Process Definition

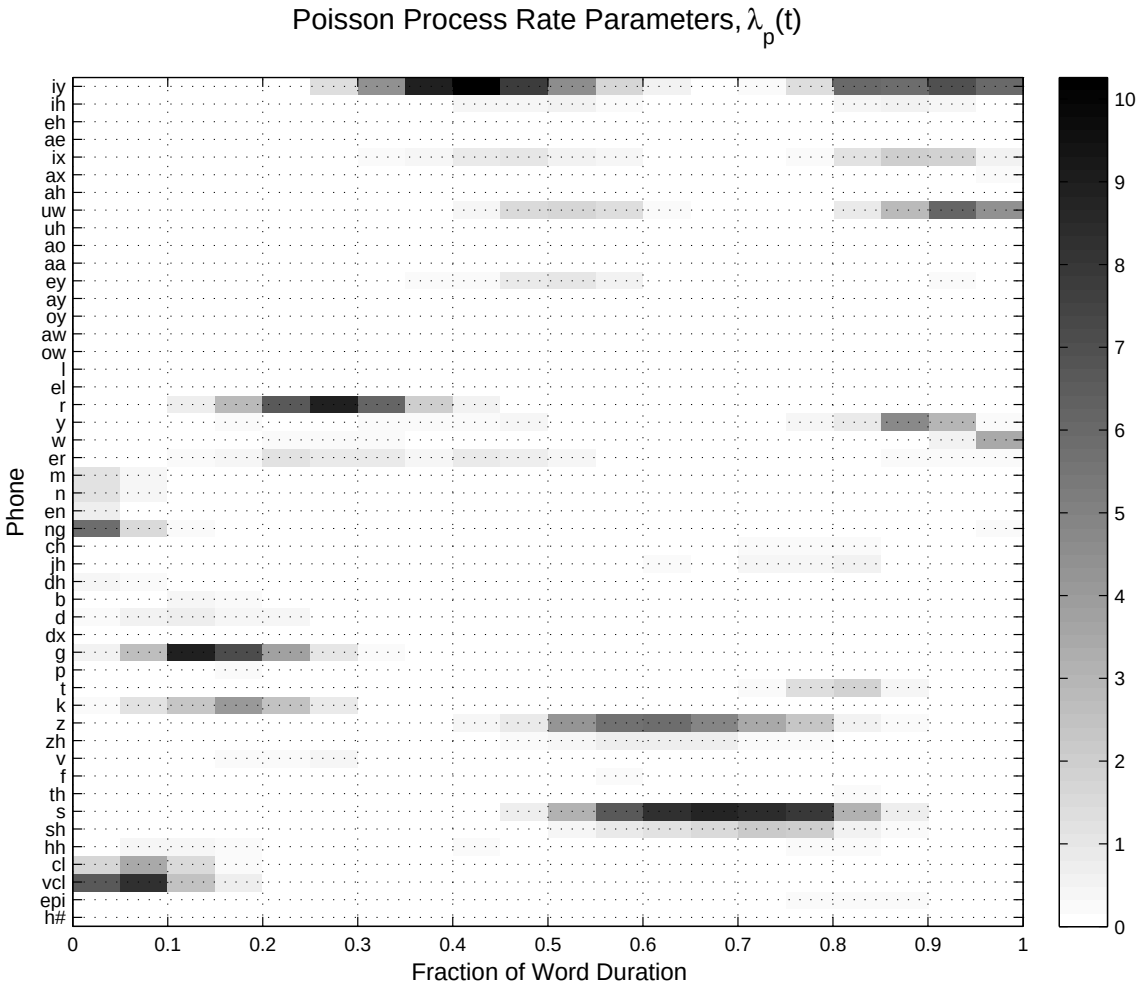
Memoryless point process with p arrival probability $\lambda_p(t)dt$ in differential time element dt at time t

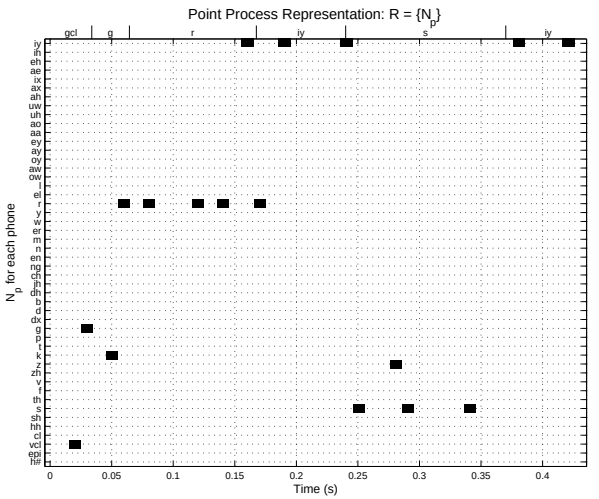
1. Normalize all $t \in R_{t,T}$ to the interval $[0, 1]$, yielding $R' = \{N'_p\}_{p \in \mathcal{P}}$
2. Assume T-independence of R', independent phone detectors, and inhomogeneous Poisson process model for each N'_p:

$$P(R_{t,T}|T, \theta_w(t)=1) = \frac{1}{T^{|R_{t,T}|}} \prod_{p \in \mathcal{P}} e^{-\int_0^1 \lambda_p(s)ds} \prod_{s \in N'_p} \lambda_p(s),$$

3. Rate functions $\{\lambda_p\}_{p \in \mathcal{P}}$ are estimated with kernel smoothing

Example: “greasy” Poisson Process Model





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

Homogeneous Poisson Process Definition

Memoryless point process with **constant** arrival probability $\mu_p dt$ in any differential time element dt

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

$$P(R_{t,T}|T, \theta_w(t)=0) = \prod_{p \in \mathcal{P}} [\mu_p]^{n_p} e^{\mu_p T},$$

3. Background rate parameters $\{\mu_p\}_{p \in \mathcal{P}}$ are estimated by counting in arbitrary background speech

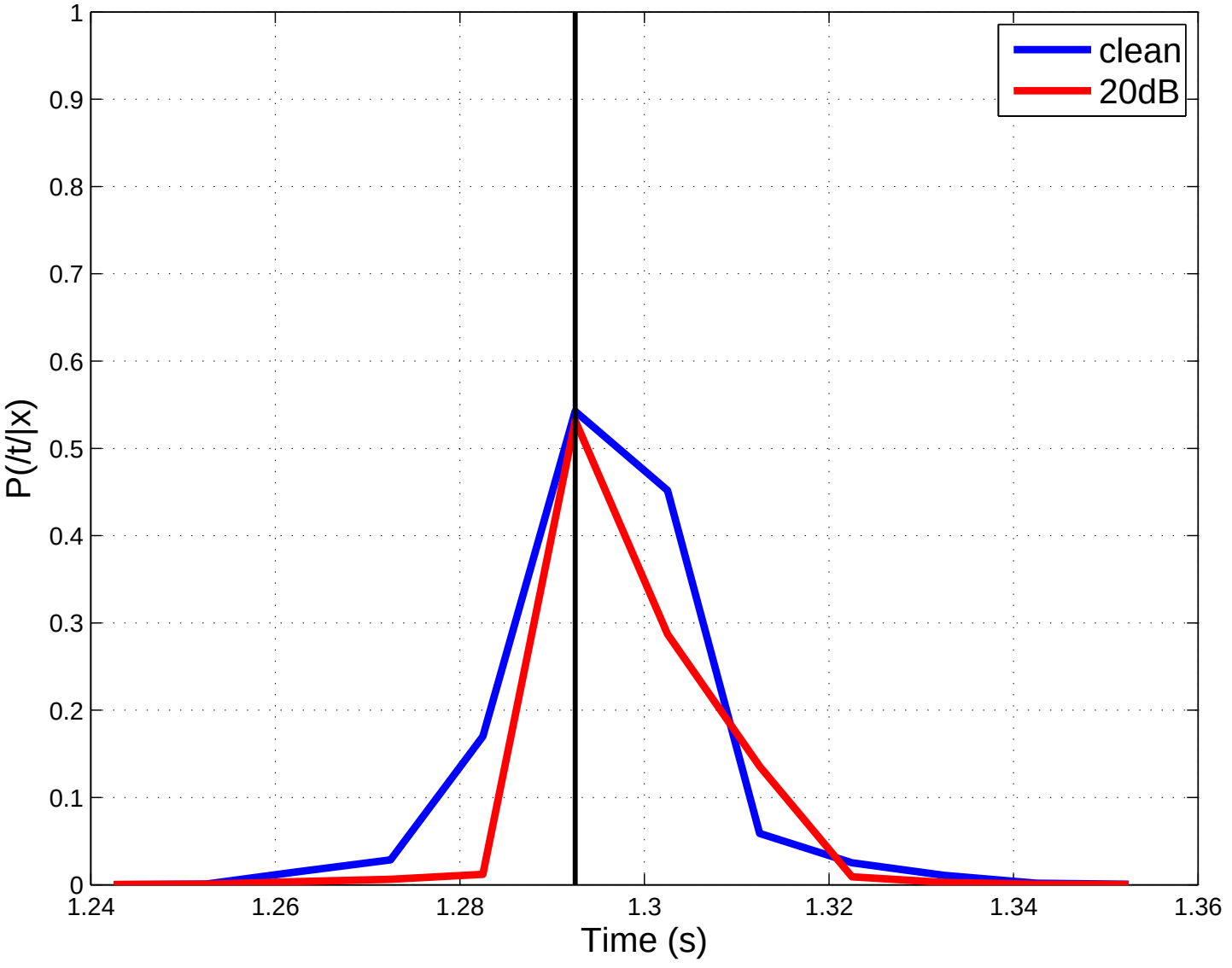
Boston University Radio News Experiments [TASLP 2009]

- **Baseline:** Keyword-filler HMM that uses same monophone model as our phone detectors

Keyword	Train	Med. T	PPM	HMM
Massachusetts	334	710 ms	**98.5**	98.0
program	44	510 ms	97.9	**98.1**
Boston	272	470 ms	**89.3**	85.7
president	52	490 ms	**83.0**	71.8
thousand	56	490 ms	78.9	**85.7**
congress	13	550 ms	**74.4**	45.0
official	7	410 ms	71.3	**79.3**
percent	80	450 ms	**71.3**	65.9
committee	41	380 ms	66.1	**70.3**
public	68	340 ms	60.1	**60.6**
yesterday	90	550 ms	59.6	**89.0**
government	43	440 ms	52.4	**72.6**
city	41	320 ms	**45.6**	38.0
hundred	121	310 ms	34.1	**49.5**
year	144	230 ms	**33.1**	20.3
seven	39	370 ms	31.3	**46.5**
about	116	250 ms	27.9	**30.2**
state	273	300 ms	**26.6**	23.3
time	82	320 ms	**25.7**	21.8
by	337	180 ms	9.0	**9.8**
		Averages:	56.8	58.1

PPM and HMM systems make different errors

What About Robustness?

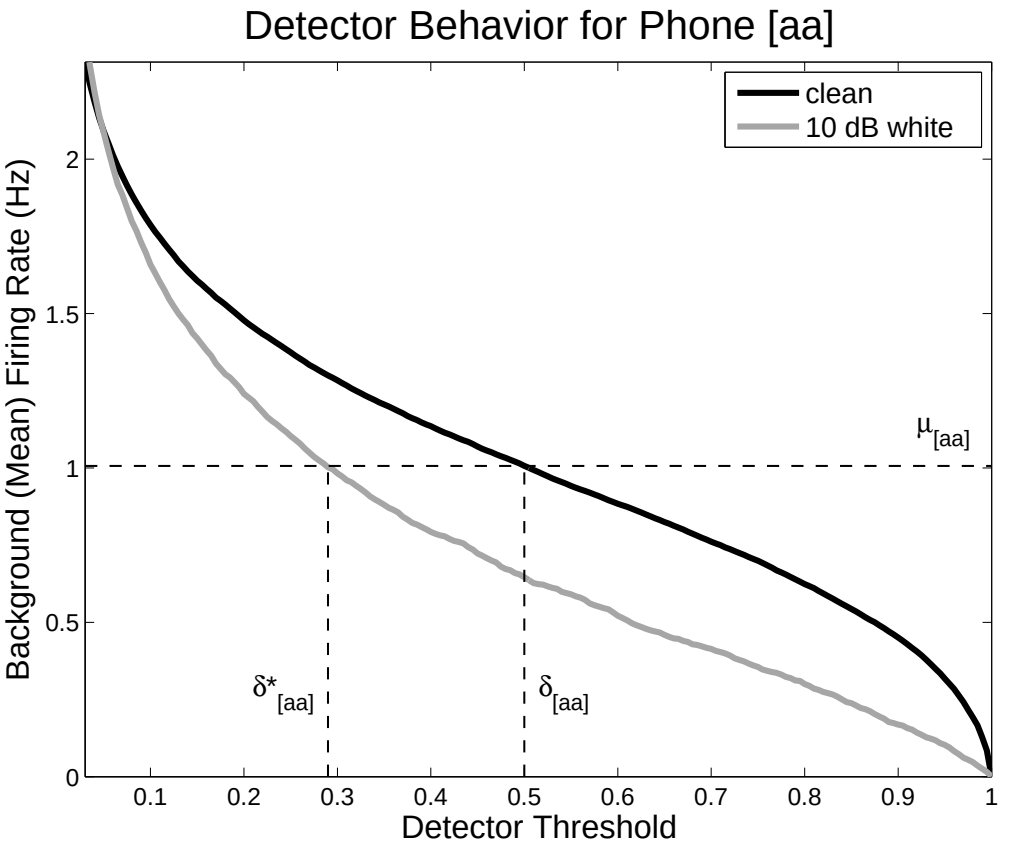


Phone Detector Threshold Adaptation

1. Find phone detector threshold δ_p^* 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**

Maximum Likelihood Linear Regression Adaptation

The Goal

Find the linear transformation that, when applied to the Gaussian means, will maximize the likelihood of the adaptation data.

- We compute separate transformations for each phone GMM
- Since we are using full covariance matrices, we limit the search to diagonal linear transformations to allow a closed form solution
- In the context of noise adaptation, we found unsupervised MLLR to harm performance
- We provide phonetic identity of each frame of the noisy adaptation data

This method is fully **supervised**

Average Robustness Levels

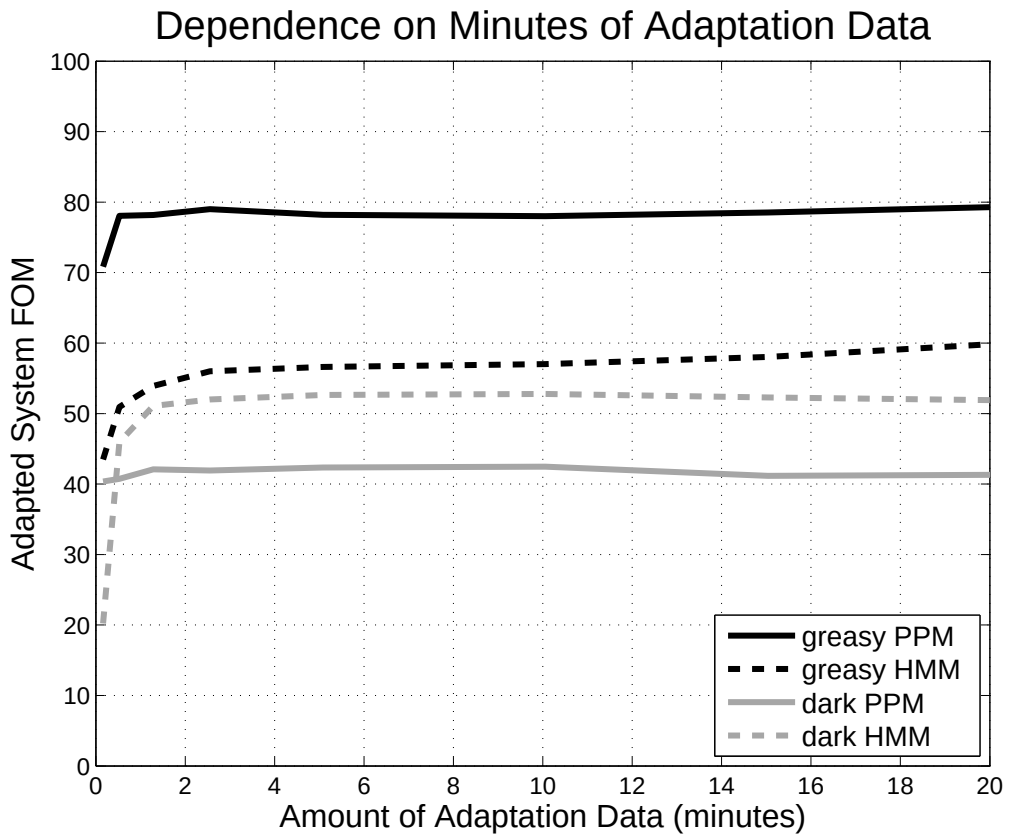
Noise Type	SNR	Adapted PPM	Adapted HMM	Non-adapted PPM	Non-adapted HMM
white	30 dB	**92.4**	72.7	84.9	84.6
	20 dB	**65.8**	50.5	47.5	44.4
	10 dB	**21.3**	16.5	7.0	5.2
	0 dB	**2.2**	0.2	0.0	0.0
pink	30 dB	**95.5**	82.0	90.7	82.3
	20 dB	**71.7**	65.5	71.6	58.6
	10 dB	**26.7**	21.1	10.2	10.5
	0 dB	**2.5**	0.0	0.3	0.0
babble	30 dB	102.3	85.3	**103.7**	95.0
	20 dB	**80.9**	73.7	76.8	63.4
	10 dB	32.5	31.9	28.2	**32.9**
	0 dB	**3.7**	3.5	1.6	2.1

Note: Individual keyword performance values in paper

- Average robustness = performance relative to clean FOM, averaged over 4 TIMIT keywords
- Adapted PPM robustness matches or outperforms adapted HMM
- Non-adapted PPM inherently more robust to 20-30 dB babble
- PPM adaptation improvement stable down to 10 seconds of data
- MLLR harms HMM performance below one minute

How Much Adaptation Data Is Needed?

Word/Noise	Adapted		Non-adapted	
	PPM	HMM	PPM	HMM
greasy, white 20 dB	79.3	59.9	46.0	56.0
dark, pink 20 dB	41.3	51.9	24.8	25.1



- PPM adaptation improves performance down to 10 seconds
- MLLR harms HMM performance below one minute

Conclusions

1. Poisson process modeling of a highly sparse, phone-based point process representation can spot keywords in clean speech as well as an equivalent frame-based HMM
2. Unsupervised PPM threshold adaptation outperforms supervised HMM MLLR adaptation in white and pink noise
3. Non-adapted PPM is inherently robust to non-stationary babble

An Advertisement

See me during the break for a real-time demo of the keyword spotting system