



IBM Research

Speech Activity Detection

Etienne Marcheret, Gerasimos Potamianos

Outline

■ **Measuring performance**

- Metrics.

■ **Techniques**

- Energy.
- Acoustic.
- Fused.
 - > CVC
- Classification.

■ **Evaluations**

- Our own.
- Tuning on development set.

Measuring SpDet performance

- **WER + Decoder CPU**
 - FA rate on pure noise + decoder performance.
 - Low latency < 200 msec.
- **ROC Curves**
 - Forced Alignment as truth N msec gray region.
- **Heuristic from hand labelled data.**
 - FA = 500 msec before start of speech.
 - FR = 100 msec after start of speech.
- **Other Metrics (NIST / CHIL)**
 - Frame level hand labeled for speech detection.

SpDet Methodology

- **Minimize FA rate in noise, no impact on WER.**
- **Non-Stationary noise.**
 - Treat “speech like noise” as noise.
- **Approach:**
 - Exploit CVC structure.
 - Allow disfluent phonemes to eat into word.
 - 3 class:
 - Pure Silence.
 - Disfluent phonemes (unvoiced fricatives, plosives)
 - Voiced phonemes.
 - > SIL, AMN, BRN, VN (silence, background noise, breathe, vocalized noise)
 - > K, S, SH, TS
 - > AA, AE, AH,
 - Fusion with energy. Energy profile helps
- **Use Decoder with disfluent words:**
 - Latency unacceptable.
 - Disfluent words have negative impact on decoder performance.

Silence Detection:

- **Energy based.**
- **Dynamic thresholding on high, mid, low tracks.**

$$e(t) = 10 \log \left(\frac{1}{N} \sum_{i=1}^N y[i]^2 \right)$$

← Bandpass filtered signal

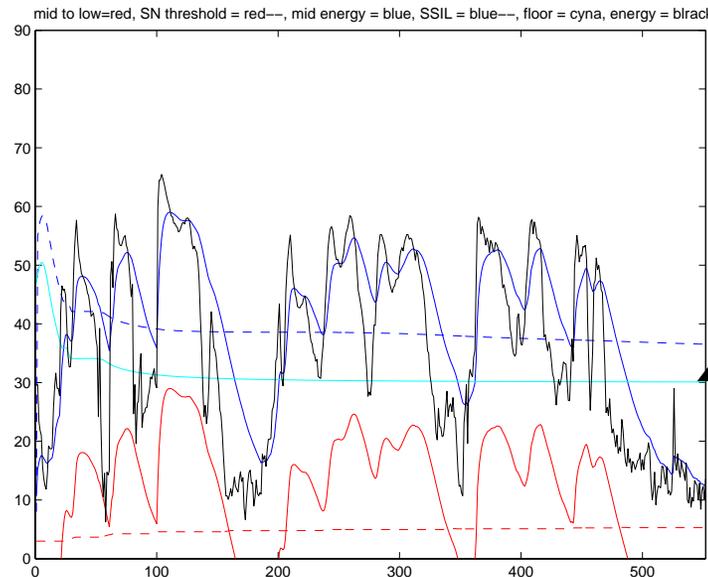
$$rms(t) = 10^{scale * e(t)}$$

$$lt(t) = (1 - \alpha_{l,t}) \times lt(t-1) + \alpha_{l,t} \times rms(t) \quad \leftarrow \quad \alpha_{l,t} = \left(\frac{lt(t-1)}{rms(t)} \right)^2$$

$$mt(t) = (1 - \alpha_m) \times mt(t-1) + \alpha_m \times rms(t) \quad \leftarrow \quad \alpha_m (0.1)$$

$$ht(t) = (1 - \alpha_{h,t}) \times ht(t-1) + \alpha_{h,t} \times rms(t) \quad \leftarrow \quad \alpha_{h,t} = \left(\frac{rms(t)}{ht(t-1)} \right)^2$$

$$m2l(t) = mt(t) - lt(t)$$



Energy based "silence detection"

Is based on these tracks.

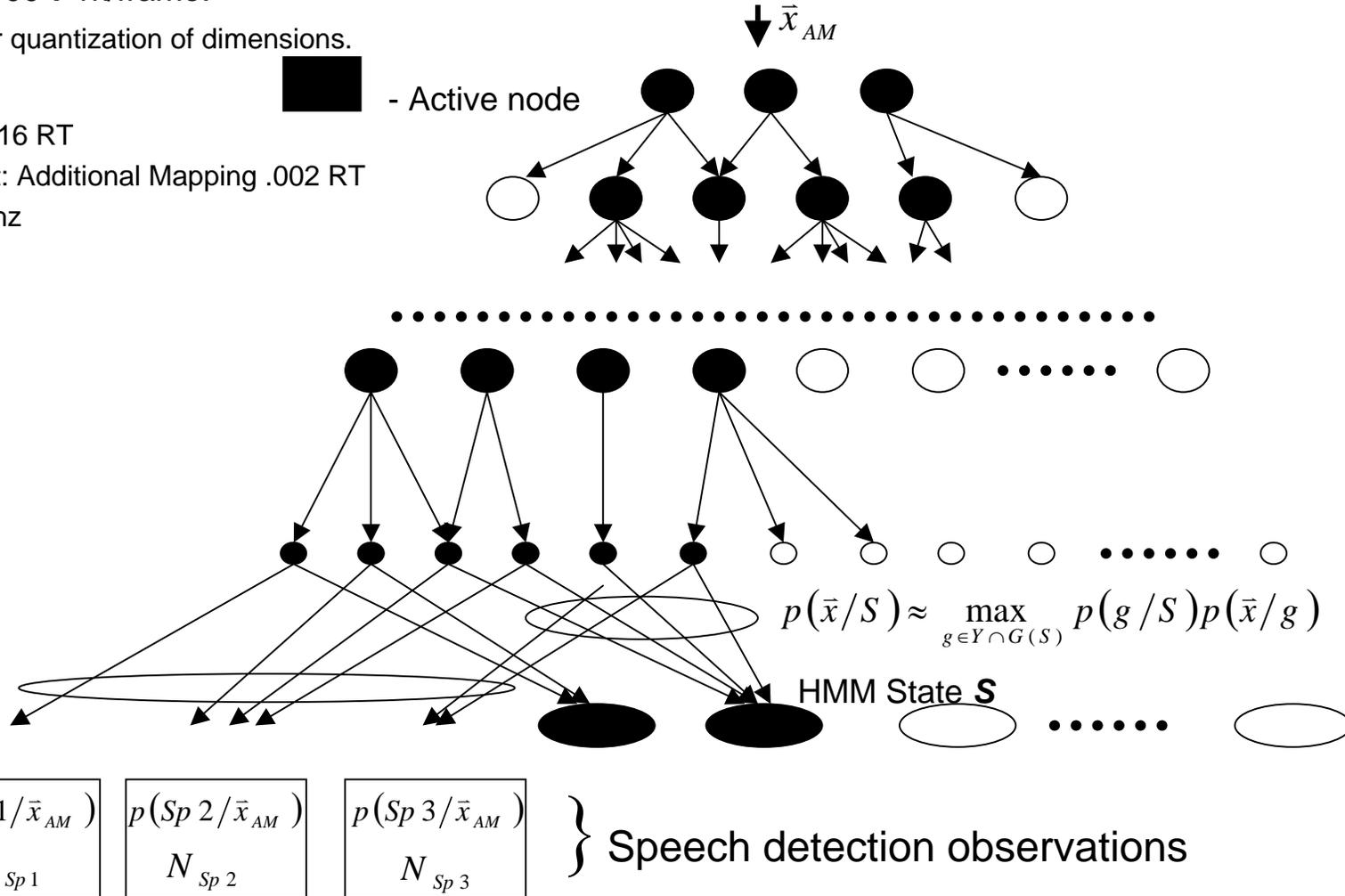
- Limitations in low SNR.

+ Decoder CPU.

Speech Detection (Acoustic Space)

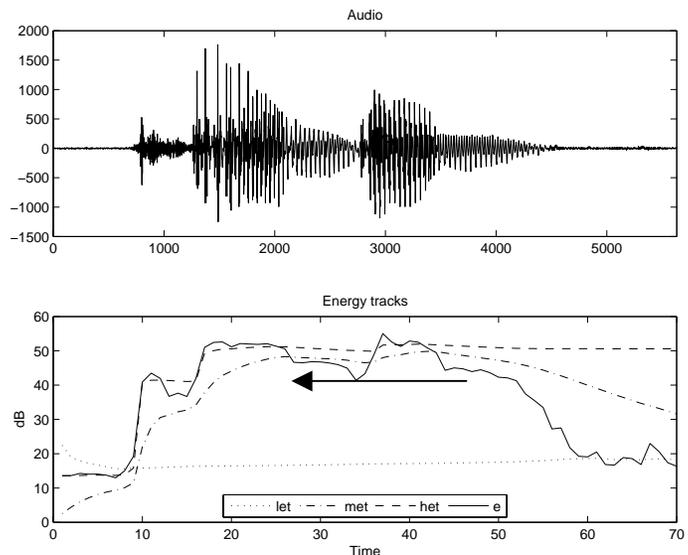
■ **Uses AM directly**

- Pruning at each level: $L_{i,j} = p(x/g_{i,j}) \rightarrow \delta_j = \max_i(L_{i,j}) - \Delta_j$,
- Evaluate 500 → 1k/frame.
 - Scalar quantization of dimensions.
- CPU
 - AM .016 RT
 - SpDet: Additional Mapping .002 RT
 - 3.2 Ghz



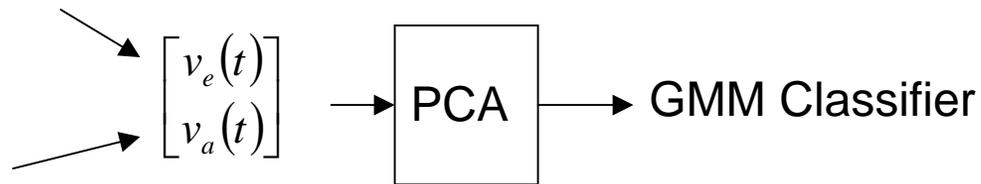
Speech Detection (Fusion)

- Energy feature space

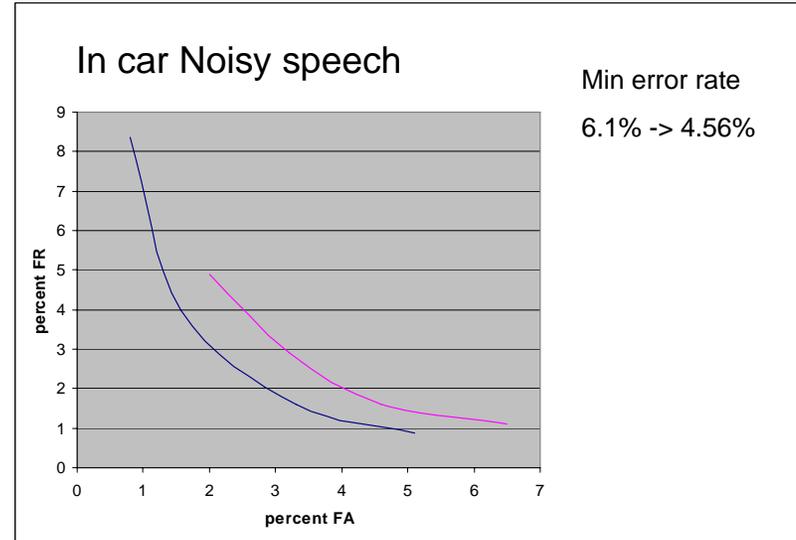
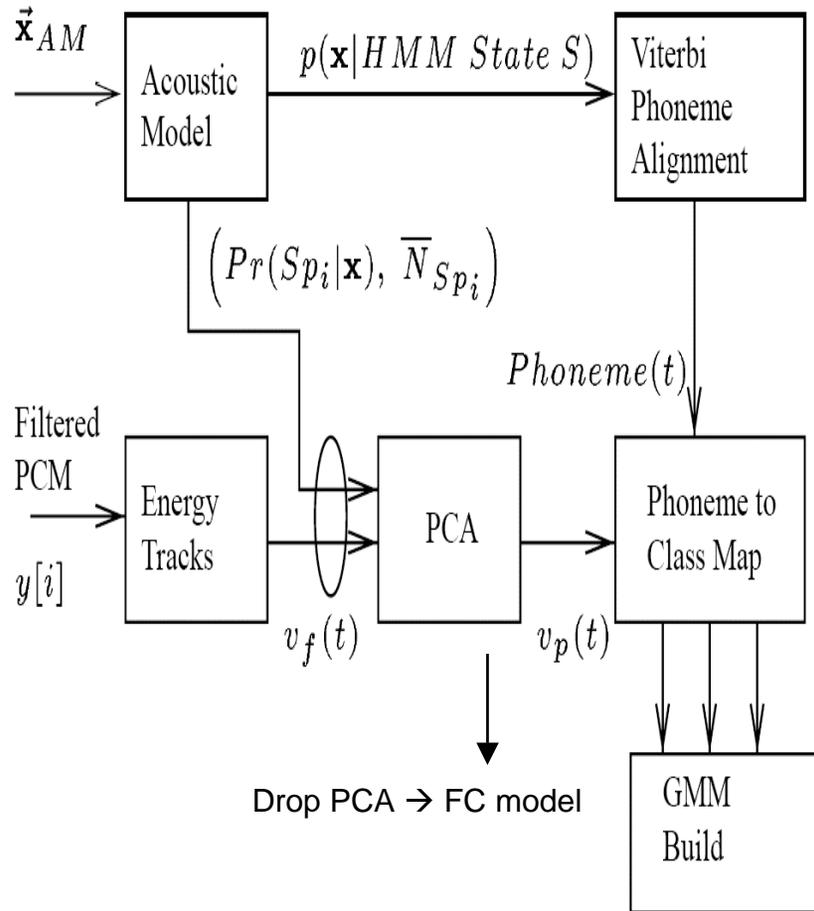


$$v_e(t) = [e(t) \quad lt(t) \quad mt(t) \quad ht(t) \quad m2l(t)]$$

$$v_a(t) = [\log(\Pr(Sp_i | x)) \quad \log(\bar{N}_{Sp_i})]$$



SpDet Training



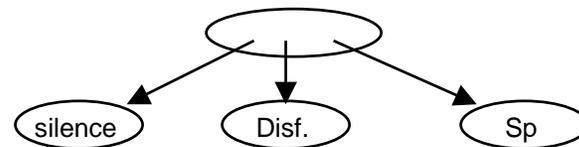
Class dependent

- (1) Class dependent LBG cluster and split to initial size.
- (2) Iterate: EM and split.



Class Independent + dependent

- (1) Class independent LBG cluster and split/EM to initial size.
- (2) Class dependent single EM step.



Speech/Silence Decode.

- **Smoothed 100msec.**
- **LRT**
 - max(pure silence (C1), disfluent (C2)) vs. (voiced (C3))
 - Classified C2 mapped to C3 if
 - ..C1...C2..C3..
 - ..C3...C2..C1...
- **Latency: LDA + 100msec.**

Evaluation

FA vs. WER

- **Maintain Robustness to noise, don't damage recognition.**

- **2773 Noise samples (recorded in car).**

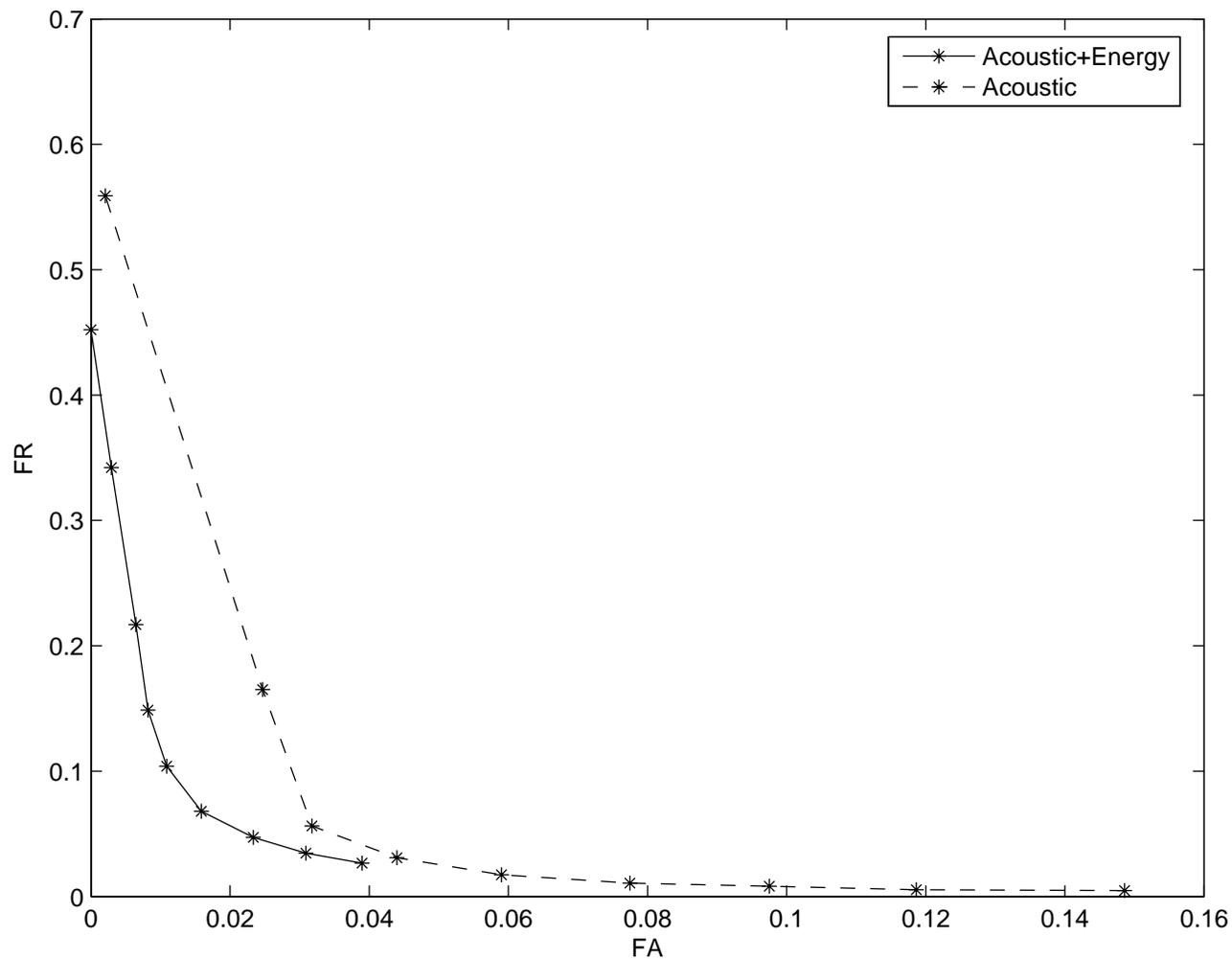
- Two class (Pure Silence) vs. (All Sp phones): 22.09%
- Two class fused with energy: 14.82%
- Three class: 4.8%

(Points for equivalent WER/SER on clean)

- **33000 sentence clean test set**

- Baseline (energy based) WER/SER = 3.20/9.97
- Three class (4.8% FA) WER/SER = 3.29/10.13
 - Heuristic Backoff of 250 msec.
 - 6k sec → 3k sec. processing time.

Benefits of Energy Fusion



Truth = Viterbi Alignment

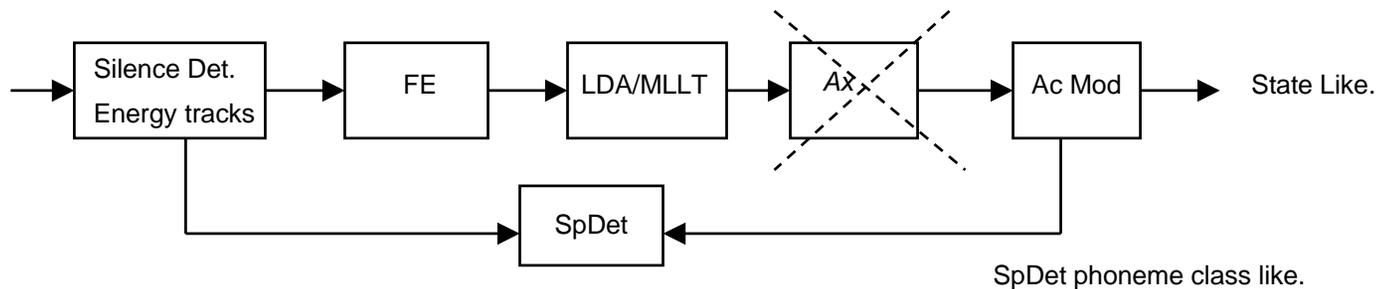
100msec no penalty

Acoustic = 7.49%

Acoustic + Energy = 6.56%

System

- AM: 13 Dim PLP, 40 dim LDA, 200k gauss, 6k states, SAT (J. Huang)
- + Quantization: 3 levels, 256 \rightarrow 4096 \rightarrow 200k, 256 quantization levels/dim.



- SpDet training:

CHIL06 (3 hrs.) + CHIL 04 Summer (4hrs.) + 5% (5 hrs.) of the rest (see J. Huang)

Model A: Diagonal, 11 \rightarrow 8 Dim PCA, 8 mixture/class

Model B: Full Covariance, 7 dim 2 mixture/class.

Model C: Full Covariance, 7 dim 4 mixture/class.

Development Data Set

■ **Tuning:**

- 7 sems (hand-labeled by UPC – refined over ELDA).
 - 3 UKA
 - 1 each of IBM, ITC, AIT, UPC.
- 3-UKA subset of above.
- ALL-DEV06 (as labeled by ELDA).

Tuning

- Speaker Diarization errors – (NIST tool)

param1: Lead/Lag Sil→Sp, Sp→Sil msec

param2: Remove SIL segments < param2 msec.

- Model A (PCA, Diagonal)

<i>param1</i>	<i>param2</i>	7-sems	3-UKA	DEV06
300	100	9.92	1.13	9.37
300	150	9.77	1.10	9.25
300	250	9.62	1.17	9.12

- Model B (FC-2mix.) [300,250] 9.69, 1.92, 9.25
- Model C (FC-4mix.) [300,250] 8.84, 1.23, 8.94

SDM → MDM

- Simple majority rules at the msec level.
- ROVER A: Smooth individual channels → Rover.
- ROVER B: Rover → Smooth.

	<i>param1</i>	<i>param2</i>	7-sems	3-UKA	DEV06
Model A	300	250	8.84	1.23	8.94
* Rover A	300	250	8.61	0.56	8.57
Rover B	300	250	8.72	0.53	8.52

Evaluations

- **Lecture only**

- MDM Best result: 8.02% (2.8 FR / 5.2 FA)
- SDM Middle of pack: 12.15% (5.7 FR / 6.5 FA)
- Large Gain from Rover.

Observations.

- **Other forms of speech detection benefit from fusion with energy profile.**
- **Incremental FMLLR significant gains for SpDet.**
- **Moving from feature → Model Fusion gives gains.**