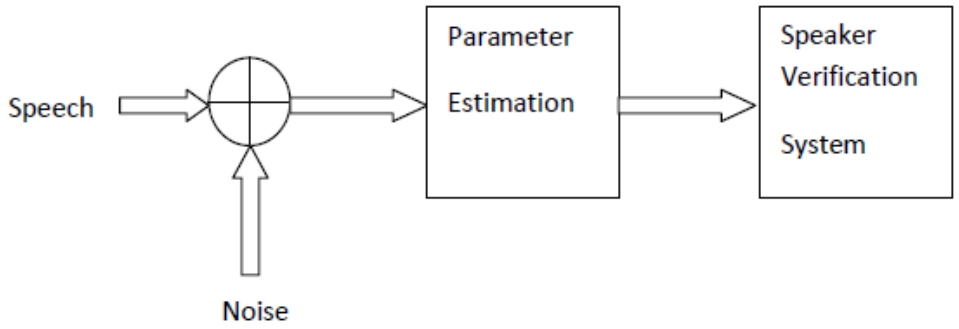
Log Spectra Enhancement using Speaker Dependent Priors for Speaker Verification

Ciira wa Maina and John MacLaren Walsh
Drexel University
Department of Electrical and Computer Engineering
Philadelphia, PA 19104
cm527@drexel.edu, jwalsh@ece.drexel.edu

Key Idea

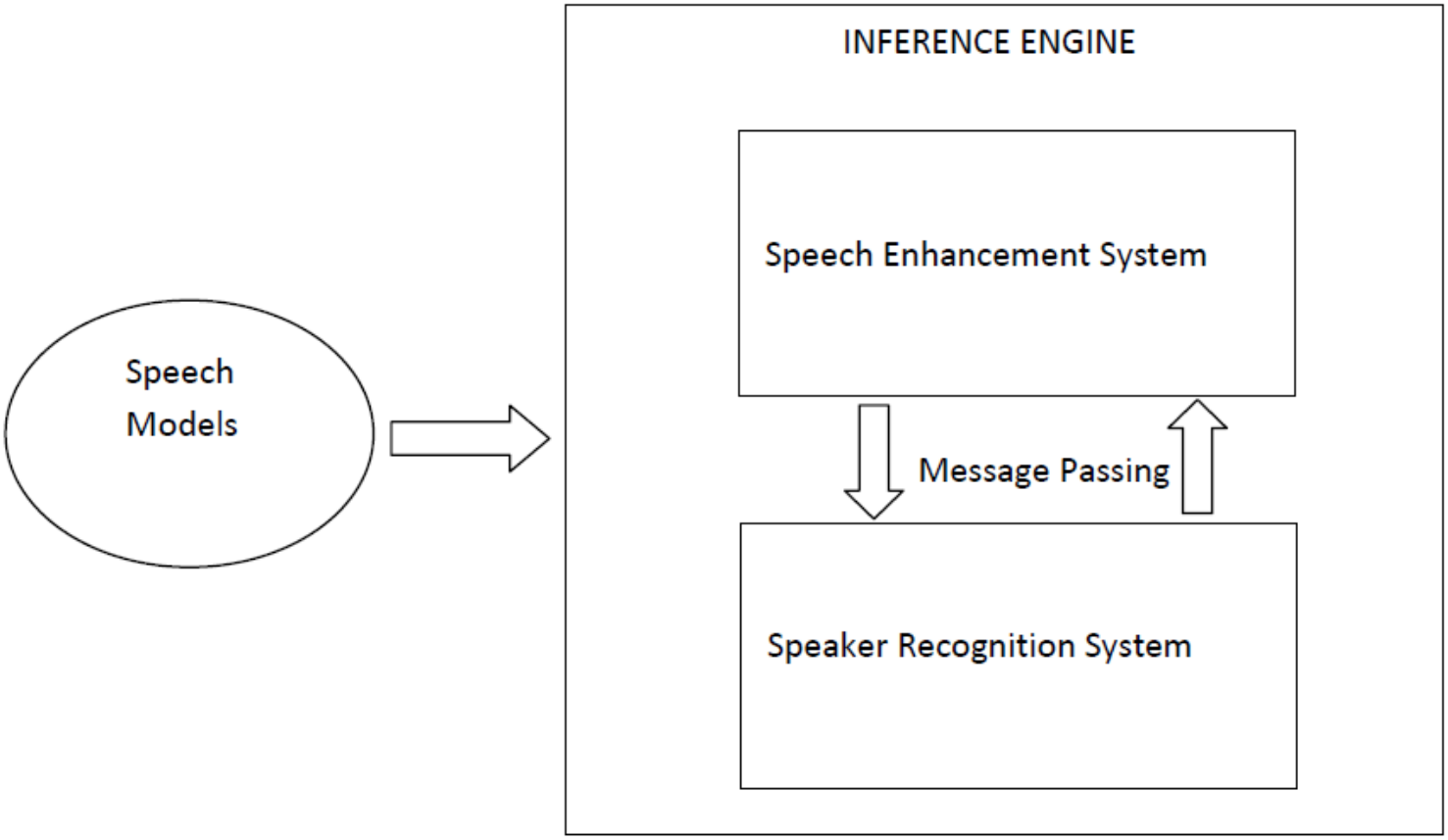
- Use Bayesian approaches to improve speaker verification.
- How
  1. Use Bayesian approaches for parameter estimation
  2. Bayesian approaches provide a principled way to account for parameter uncertainty



- Two main causes of performance degradation in speaker verification
  1. Noise
  2. Mismatch

Key Idea cont.

- Exploit the link between enhancement and speaker recognition.
- Make use of speaker specific priors

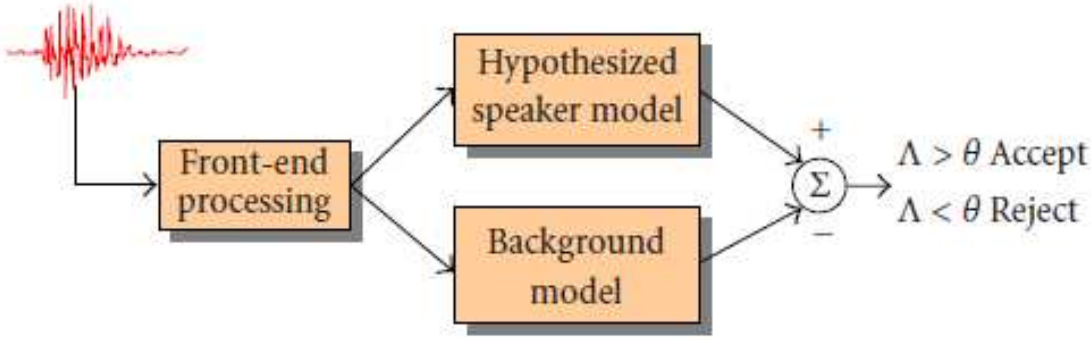


Outline

- **Speaker Verification**
- **Bayesian Inference**
- **Variational Bayesian inference**
- **Probabilistic model**
- **Experimental results**

Background: Speaker Verification

- The basic task is to determine whether a given speaker is speaking in a particular speech segment [Bimbot *et al.* 2004].
- Thus given a speech segment X we test the following hypotheses
  - H0: X is from speaker S
  - H1: X is not from speaker S



- Target speakers are modelled using speaker specific GMMs
- A universal background model (UBM) is used to test the alternate hypothesis H1.

Speaker verification cont.

- The likelihood ratio is compared to a threshold in order to determine which hypothesis is correct.
- For each trial we compute the score

$$\text{Score} = \log p(\mathbf{X}|\text{TargetModel}) - \log p(\mathbf{X}|\text{UBM}).$$

  where $\mathbf{X}$ are the features computed from the test utterance. One of the most popular parameterizations are Mel Frequency cepstral coefficients (MFCCs) [Reynolds and Rose 1995]
- Threshold in the hypothesis test determines the decision
- Trade off between probability of false alarm and probability of missed detection
- Detection Error tradeoff curves popular in the speaker verification community
- Performance metric is the equal error rate (EER).

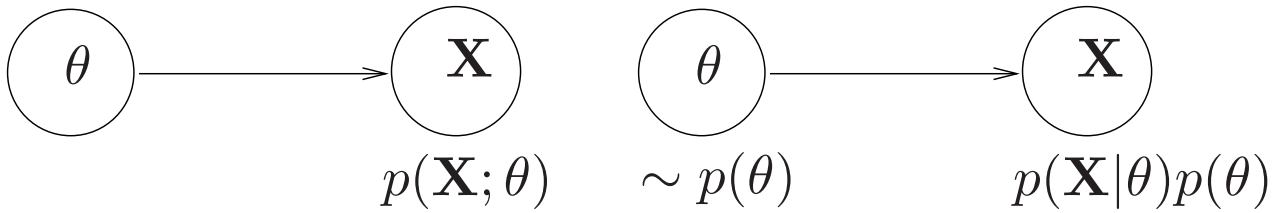
Outline

- Speaker Verification
- **Bayesian Inference**
- Variational Bayesian inference
- Probabilistic model
- Experimental results

Background: Bayesian Inference

- **Two main approaches to parameter estimation**

  1. Maximum likelihood: Parameter an unknown constant
  2. Bayesian Inference: Parameter a random variable



(a) Maximum likelihood (b) Bayesian Inference

- **Key quantities**

  1. Maximum likelihood: The likelihood $p(\mathbf{X};\theta)$
  2. Bayesian Inference: The posterior $p(\theta|\mathbf{X}) \propto p(\mathbf{X}|\theta)p(\theta)$

Background: Bayesian Inference

- In the Bayesian framework, the parameters of our probabilistic model are treated as random variables governed by a prior $p(\Theta)$.
- The posterior is a central quantity in Bayesian inference and is given by

$$p(\Theta|\mathbf{X}) = \frac{p(\mathbf{X}|\Theta)p(\Theta)}{\int p(\mathbf{X}|\Theta)p(\Theta)d\Theta}$$

- Bayesian estimators are derived to minimize various expected costs

$$\hat{\Theta} = \arg\min_{\Theta} \int \mathcal{C}(\Theta - \hat{\Theta})p(\Theta|\mathbf{X})d\Theta$$

- We can obtain parameter estimates such as $\hat{\Theta}_{\text{MMSE}} = \int \Theta p(\Theta|\mathbf{X})d\Theta$. which corresponds to $\mathcal{C}(\Theta - \hat{\Theta}) = \|\Theta - \hat{\Theta}\|^2$
- These integrals are often intractable.

Variational Bayesian (VB) Inference

- VB is an approximate Bayesian inference technique [Bishop 2006].
- We use VB to obtain an approximation $q(\Theta)$ to the intractable posterior $p(\Theta|\mathbf{X})$ which minimizes the Kullback-Leibler (KL) divergence between $q(\Theta)$ and $p(\Theta|\mathbf{X})$ with $q(\Theta)$ constrained to lie within a tractable approximating family.

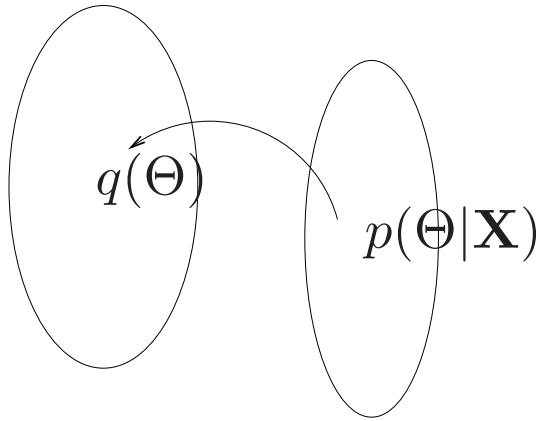


Figure 1: Approximating the intractable posterior

Variational Bayesian (VB) Inference cont.

- To ensure tractability we assume that the posterior can be written as a product of factors depending on disjoint subsets of $\Theta = \{\theta_1, \ldots, \theta_M\}$.
- $$q(\Theta) = \prod_{i=1}^{M} q_i(\theta_i). \tag{1}$$
- We determine the optimal form of $q_j(\theta_j)$ denoted by $q_j^*(\theta_j)$ that minimizes $D(q||p)$
- $$\log q_j^*(\theta_j) = \mathbb{E}\{\log p(\mathbf{X}, \Theta)\}_{q(\Theta \setminus j)} + const. \tag{2}$$

Outline

- Speaker Verification
- Bayesian Inference
- Variational Bayesian inference
- **Probabilistic model**
- Experimental results

Speaker verification

- We work in the log spectral domain
- The observed signal is clean speech corrupted by additive noise

$$y[t] = s[t] + n[t]$$

- To obtain the log spectrum we take the DFT

$$Y[k] = S[k] + N[k]$$

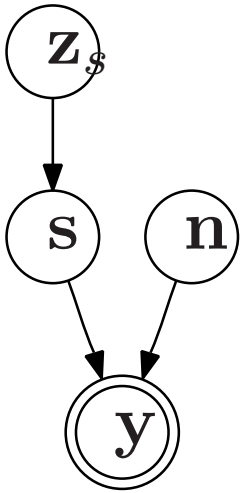
  and take the logarithm of the power spectrum $\mathbf{y} = \log |Y[:]|^2$
- It can be shown [Frey *et al.*]

$$\mathbf{y} \approx \mathbf{s} + \log(\mathbf{1} + \exp(\mathbf{n} - \mathbf{s}))$$

- And

$$p(\mathbf{y}|\mathbf{s}, \mathbf{n}) = \mathcal{N}(\mathbf{y}|\mathbf{s} + \log(\mathbf{1} + \exp(\mathbf{n} - \mathbf{s})), \boldsymbol{\psi}).$$

Speaker verification



- The joint distribution of this model is

$$p(\mathbf{y}, \mathbf{s}, \mathbf{z}_s, \mathbf{n}) = p(\mathbf{y}|\mathbf{s}, \mathbf{n})p(\mathbf{s}|\mathbf{z}_s)p(\mathbf{z}_s)p(\mathbf{n}).$$

- The prior over $\mathbf{s}$ is a speaker dependent GMM.

$$p(\mathbf{s}|\ell) = \sum_{m=1}^{M_s} \pi_{\ell m}^s \mathcal{N}(\mathbf{s}; \boldsymbol{\mu}_{\ell m}^s, \boldsymbol{\Sigma}_{\ell m}^s)$$

  where $\ell \in \mathcal{L} = \{\mathrm{TargetSpeaker}, \mathrm{UBM}\}$

- $\mathbf{z}_s$ is an indicator variable to indicate 'speaker' and mixture component.

VB Algorithm

- We assume an approximate posterior $q(\Theta)$ that factorizes as follows

$$q(\Theta) = q(\mathbf{s})q(\mathbf{z}_s)q(\mathbf{n}).$$

- The forms of the factors are
  - $q^*(\mathbf{s}) = \mathcal{N}(\mathbf{s}; \boldsymbol{\mu}^*_{\mathbf{s}}, \boldsymbol{\Sigma}^*_{\mathbf{s}})$
  - $q^*(\mathbf{n}) = \mathcal{N}(\mathbf{n}; \boldsymbol{\mu}^*_{\mathbf{n}}, \boldsymbol{\Sigma}^*_{\mathbf{n}})$
  - $q^*(\mathbf{z}_s) = \prod_{i=1}^{M_s|\mathcal{L}|}(\gamma_i)^{z_{s,i}}$
- We enhance the observed log spectra by computing the posterior mean of the clean log spectra.
- Derive MFCCs from the enhanced log spectra for verification

Outline

- Speaker Verification
- Bayesian Inference
- Variational Bayesian inference
- Probabilistic model
- **Experimental results**

Speaker verification Experiments

- We use the TIMIT data set and the MIT Mobile Device Speaker Verification Corpus (MDSVC).
- Train a UBM using 300 speakers
- The initial verification experiments were performed with the test utterances corrupted by additive white Gaussian noise at various input SNRs.
- Two true trials and 20 impostor trials per speaker
- Since there are 630 speakers we get 1260 true trials and 12600 impostor trials
- For each trial compute the score

$$\text{Score} = \log p(\mathbf{X}|\text{TargetModel}) - \log p(\mathbf{X}|\text{UBM}).$$

Speaker verification Experiments cont.

- We compare the VB algorithm to feature domain intersession compensation (FDIC) [Castaldo 2007].
- This is a feature-domain technique
- Observed features are projected onto a session independent subspace
- This compensates for mismatch between training and testing conditions.
- The projection matrix is determined from training data.

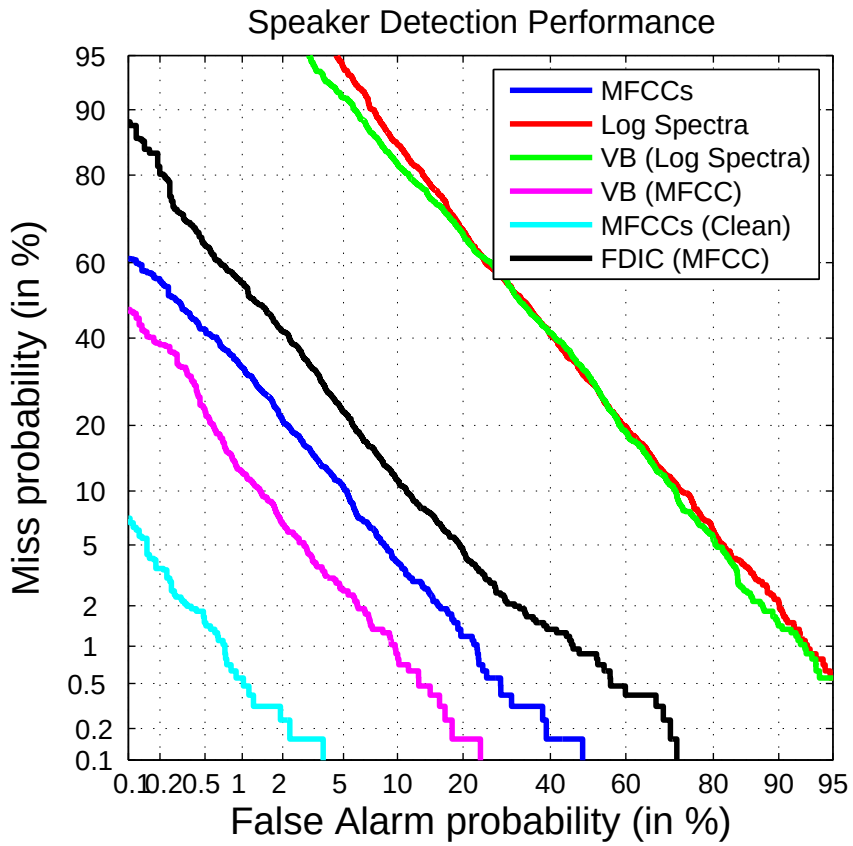
$$\hat{\mathbf{o}}_t = \mathbf{o}_t - \sum_{m=1}^{M} \gamma_m(\mathbf{o}_t)\mathbf{U}_m\mathbf{x}$$

Speaker verification Experiments cont.

Table 1: Speaker verification EER (%) for the entire TIMIT data set

System	SNR (dB)						
	10	20	22	24	26	28	30
MFCCs (Baseline)	43.49	23.25	18.97	15.24	11.98	9.21	6.83
VB (MFCC)	26.51	11.83	9.44	7.46	6.27	4.84	3.65
FDIC	33.25	20.56	17.94	15.63	14.84	12.62	10.56
Log Spectra	49.68	45.16	43.89	43.17	42.06	40.79	40.48
VB (Log Spectra)	44.68	43.57	42.78	42.22	41.51	40.40	40.71

Speaker verification Experiments cont.



- Speaker verification performance for the entire TIMIT data set at 30dB
- Equal Error rate (EER) is reduced by about half from 6.83% to 3.65%.

Speaker verification Experiments cont.

Table 2: Speaker verification EER (%) for the entire TIMIT data set in factory noise

System	SNR (dB)						
	0	5	10	15	20	25	30
MFCCs (Baseline)	46.79	39.13	27.78	15.95	7.54	2.94	1.67
VB (MFCC)	35.48	23.49	11.90	6.11	3.17	2.06	1.51
Log Spectra	47.22	46.35	44.05	40.85	37.54	35.40	34.84
VB (Log Spectra)	44.84	42.06	39.92	37.78	35.87	35.08	35.48

Speaker verification Experiments cont.

- The MIT Mobile Device Speaker Verification Corpus (MDSVC).
  - Speech recorded in an office, hallway, and street intersection.
  - 48 Target speakers.
  - 40 impostors.
  - Speaker models trainied using office data.

Speaker verification Experiments cont.

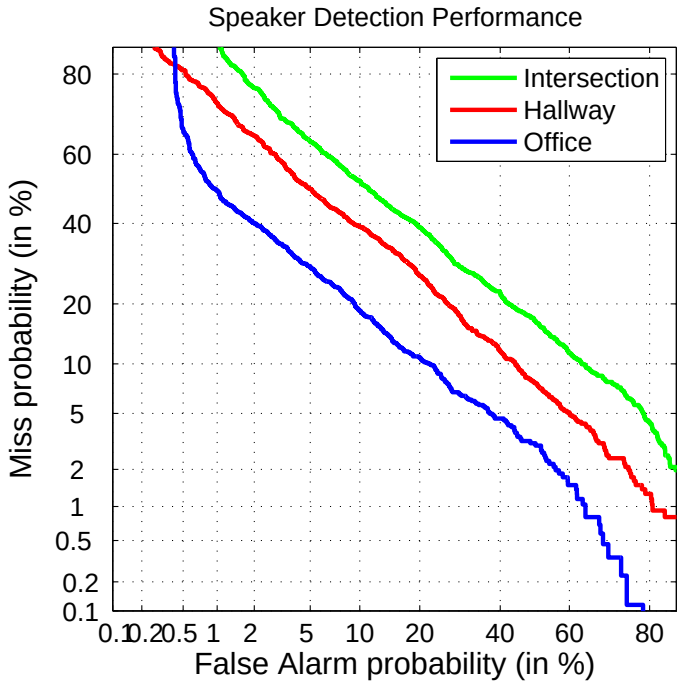


Table 3: Speaker verification results for MDSVC test data in the three different environments

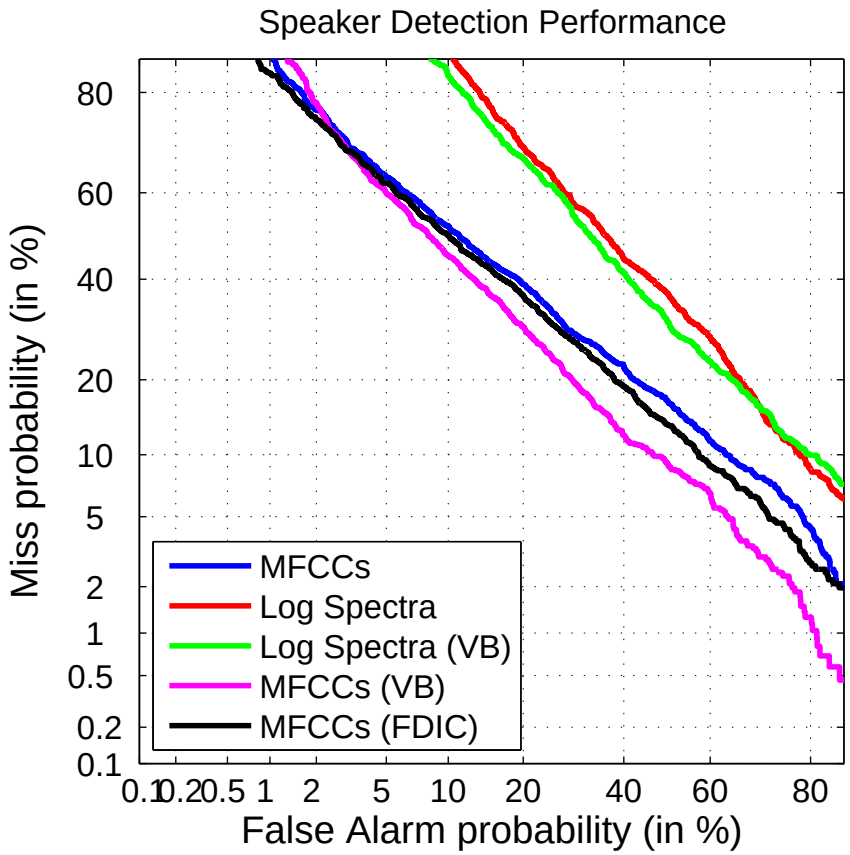
Location	EER (%)
Office	14.24
Hallway	22.92
Intersection	28.82

Speaker verification Experiments cont.

Table 4: Speaker verification EER (%) for the MDSVC data set

System	Intersection EER
MFCCs (Baseline)	28.82
VB (MFCC)	24.54
FDIC	27.89
Log Spectra	42.71
VB (Log Spectra)	40.63

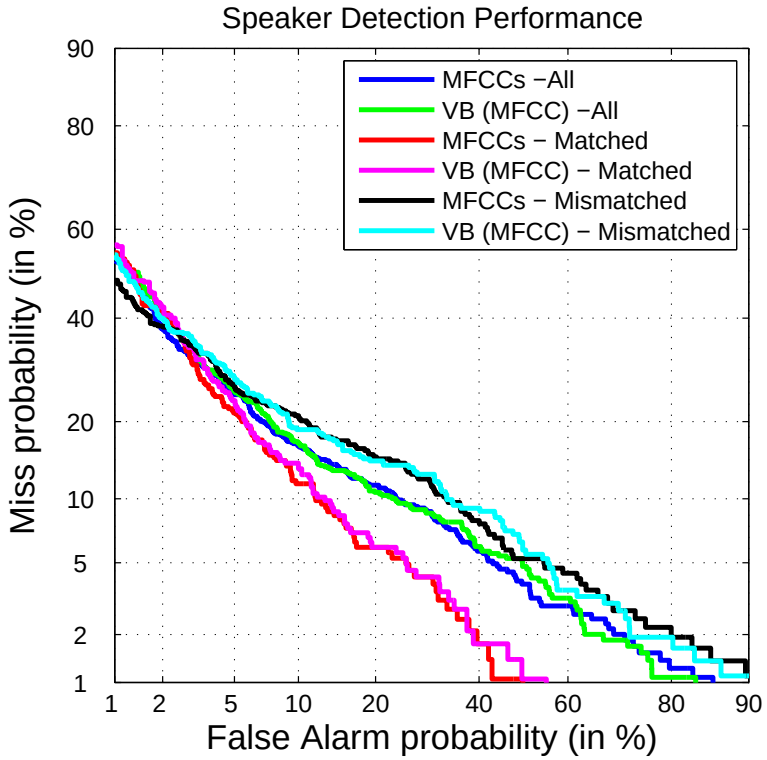
Speaker verification Experiments cont.



Speaker verification Experiments cont.

- For the SRE data experiments we use the 2004 corpus
- Results are shown for the 1side-1side condition
- We determine the improvement in performance in trials with telephone type mismatch between training data and testing data
- Gender dependent UBMs with 512 mixture coefficients were trained using approximately 20 hours of speech.
- Speaker models were then obtained using MAP adaptation with only the means of the UBM being adapted.
- We use 19 dimensional MFCCs extracted using a 20ms window with 50% overlap.
- RASTA processing and CMS is performed.

Speaker verification Experiments cont.



- our baseline system has an EER of 13.89% and the VB system has an EER of 13.43%.
- Mismatched: EER reduces from 16.53% to 15.70%, Matched: EER reduces from 11.58% to 11.23%

Conclusion

- We have presented a log spectral VB algorithm for speaker verification
- Performance gains are reported in additive noise
- Mismatch compensation has been demonstrated
- The technique requires clean speech to train models
- Only modest gains obtained for SRE data

References

[1] D. Reynolds and R. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Trans. Speech Audio Processing*, 3(1):72–83, 1995.

[2] B. J. Frey, T. T. Kristjansson, L. Deng, and A. Acero. ALGONQUIN Learning dynamic noise models from noisy speech for robust speech recognition. In *Advances in Neural Information Processing Systems 14*, pages 1165–1172, January 2002.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[4] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, C. Vair. Compensation of Nuisance Factors for Speaker and Language Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):1969 – 1978, September 2007.