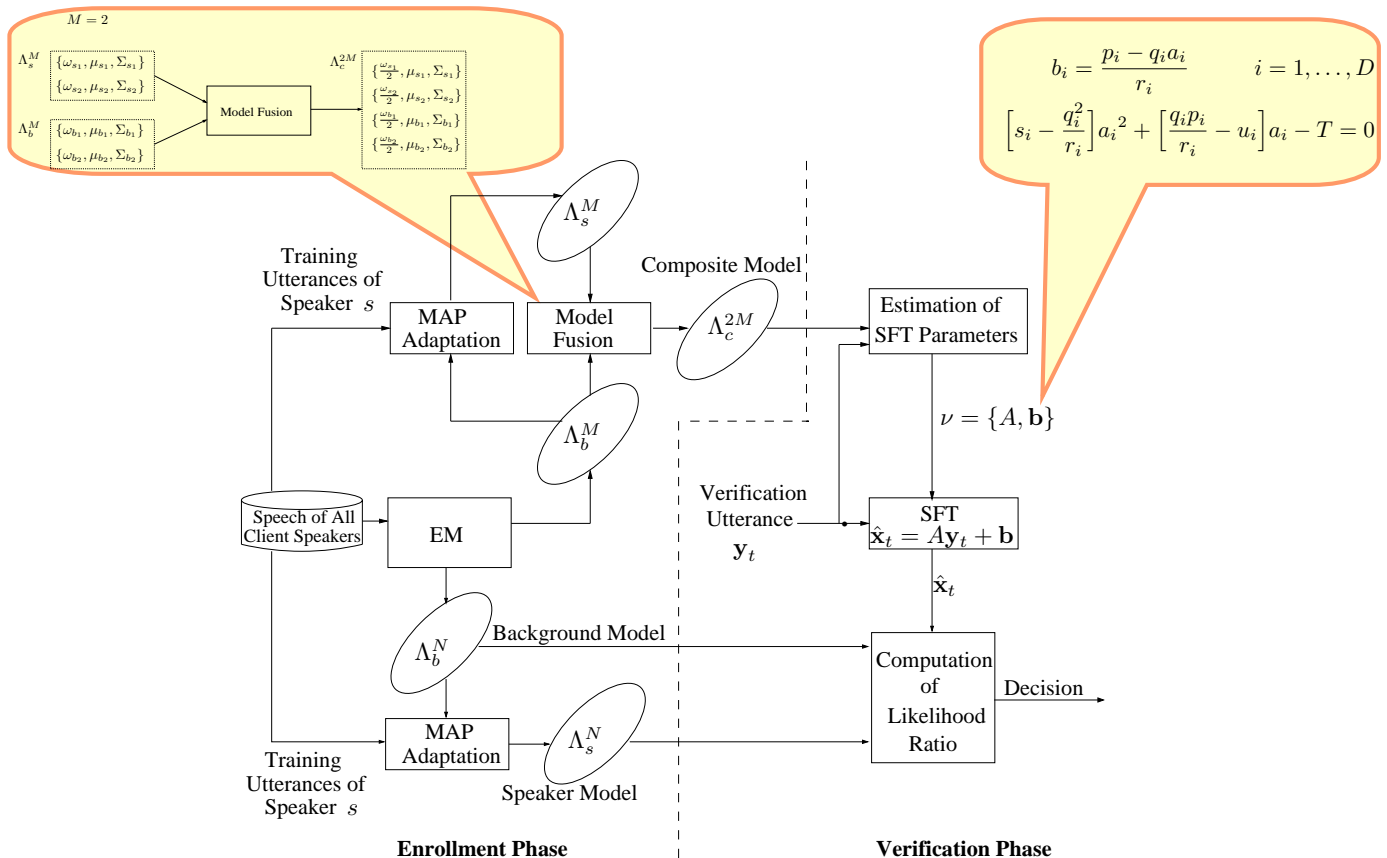


Summary

- To improve the reliability of telephone-based speaker verification systems, we have recently proposed the blind stochastic feature transformation (BSFT).
- However, for short-utterances, the BSFT algorithm may transform the distorted features to an undesirable region, resulting in incorrect verification decisions.
- This paper extends the BSFT algorithm to handle the short-utterance scenario. The idea is to estimate the prior distribution of the transformation parameters. The prior distribution is incorporated into the online estimation of the BSFT parameters via MAP criterion. We refer the extended algorithm to as Bayesian blind stochastic feature transformation (BBSFT).
- Experimental results show that BBSFT performs better than BSFT for short utterances.

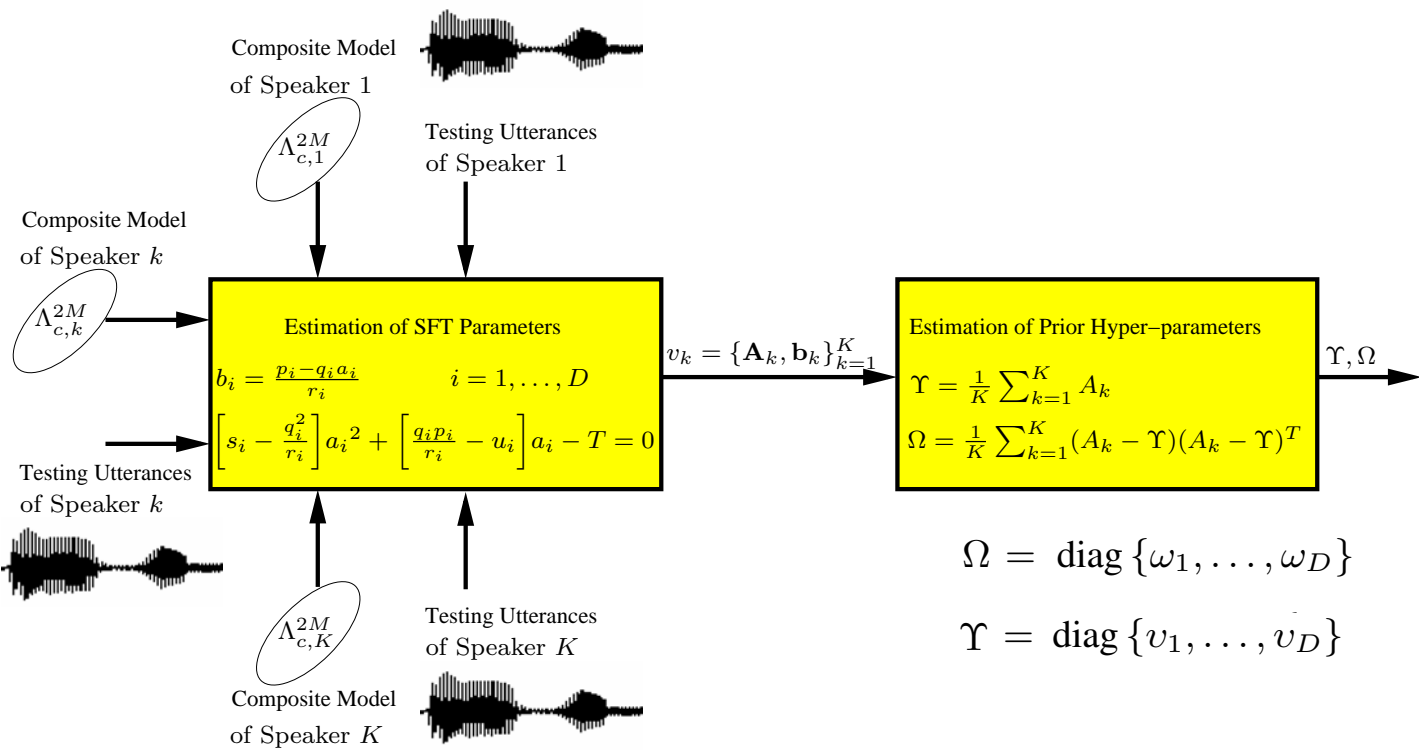
1

Blind Stochastic Feature Transformation (BSFT)



2

Bayesian Blind Stochastic Feature Transformation



3

Bayesian Blind Stochastic Feature Transformation

- In Bayesian BSFT and BSFT, the transformed feature vector is

$$\hat{\mathbf{x}} = f_\nu(\mathbf{y}) = \mathbf{A}\mathbf{y} + \mathbf{b}$$

- In Bayesian BSFT, the transformation parameters are modeled by probability density functions

$$p(\mathbf{A}) = \frac{|\Omega|^{-\frac{(D+1)}{2}}}{|\Phi|^{\frac{D}{2}}} \cdot \exp \left\{ -\frac{1}{2} \text{tr}(\mathbf{A} - \Upsilon)^T \Omega^{-1} (\mathbf{A} - \Upsilon) \Phi^{-1} \right\}$$

and

$$p(\mathbf{b}) = (2\pi)^{-\frac{D}{2}} \cdot \left\{ \prod_{i=1}^D \gamma_i \right\} \cdot \exp \left\{ -\frac{1}{2} \sum_{i=1}^D (b_i - \beta_i)^2 \gamma_i^2 \right\}$$

where $\Omega = \text{diag} \{ \omega_1, \dots, \omega_D \}$ and $\Upsilon = \text{diag} \{ v_1, \dots, v_D \}$

4

Bayesian Blind Stochastic Feature Transformation

$$\begin{aligned}
 Q(\nu'|\nu) &= Q(A', \mathbf{b}'|A, \mathbf{b}) \\
 &= \sum_{t=1}^T \sum_{j=1}^M h_j(f_\nu(\mathbf{y}_t)) \log \left\{ \frac{p(f_{\nu'}(\mathbf{y}_t)|\mu_j, \Sigma_j)}{|J_{\nu'}(\mathbf{y}_t)|} \right\} \\
 &\quad + \log p(A')p(\mathbf{b}')
 \end{aligned}$$

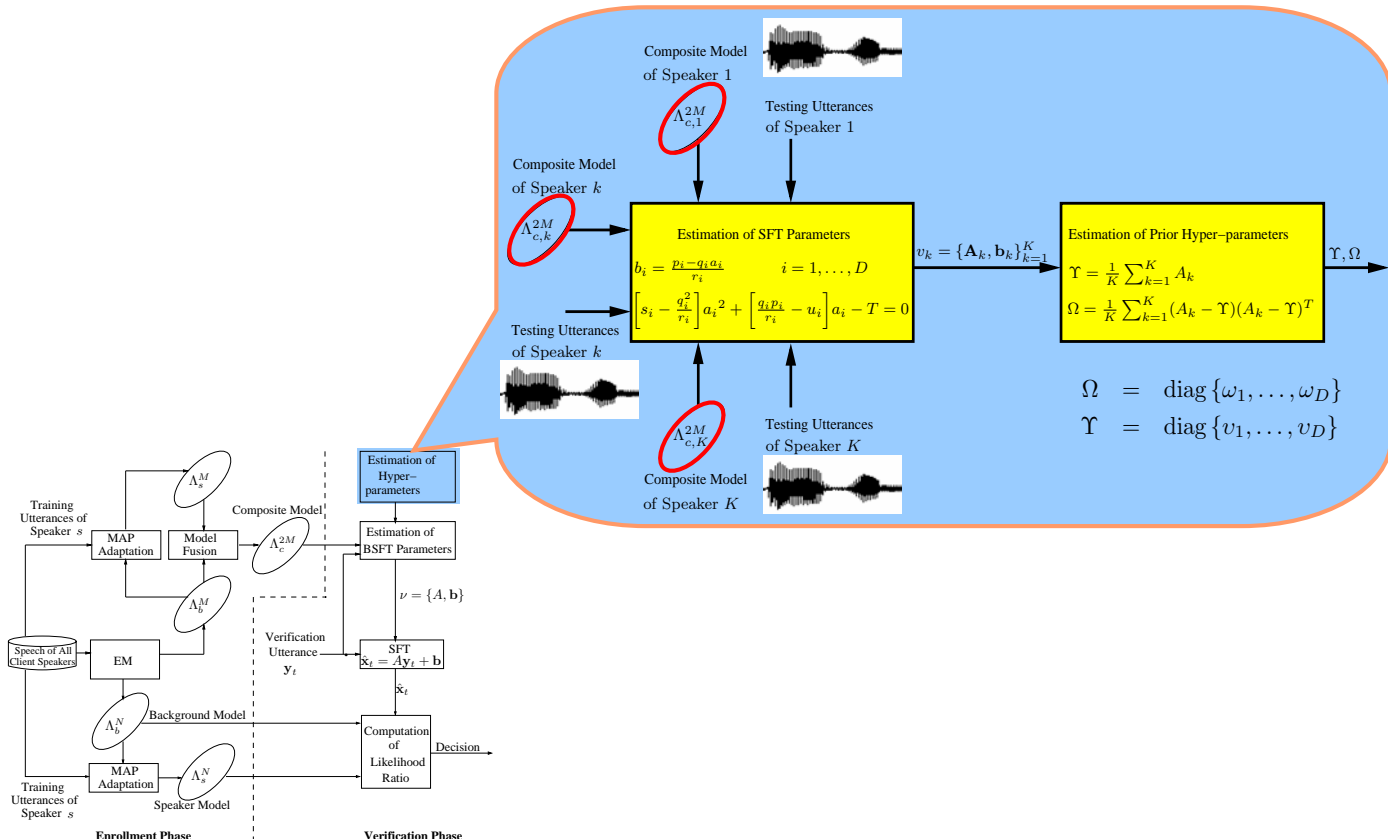
$$\frac{\partial Q(\nu'|\nu)}{\partial a'_i} = 0 \qquad \frac{\partial Q(\nu'|\nu)}{\partial b'_i} = 0$$

$$\Rightarrow \begin{cases} b'_i = \frac{p_i - q_i a'_i + \beta_i \gamma_i^2}{r_i + \gamma_i^2} \\ \left[1 + \omega_i s_i - \frac{\omega_i q_i^2}{r_i + \gamma_i^2} \right] a_i'^2 + \left[\frac{\omega_i q_i (p_i + \beta_i \gamma_i^2)}{r_i + \gamma_i^2} - \omega_i u_i - v_i \right] a'_i - \omega_i T = 0 \end{cases}$$

$$\begin{aligned}
 p_i &= \sum_{t=1}^T \sum_{j=1}^M h_j(f_\nu(\mathbf{y}_t)) \mu_{ji} \sigma_{ji}^{-2}, \\
 q_i &= \sum_{t=1}^T \sum_{j=1}^M h_j(f_\nu(\mathbf{y}_t)) y_{ti} \sigma_{ji}^{-2}, \\
 r_i &= \sum_{t=1}^T \sum_{j=1}^M h_j(f_\nu(\mathbf{y}_t)) \sigma_{ji}^{-2}, \\
 s_i &= \sum_{t=1}^T \sum_{j=1}^M h_j(f_\nu(\mathbf{y}_t)) y_{ti}^2 \sigma_{ji}^{-2}, \text{ and} \\
 u_i &= \sum_{t=1}^T \sum_{j=1}^M h_j(f_\nu(\mathbf{y}_t)) \mu_{ji} y_{ti} \sigma_{ji}^{-2}.
 \end{aligned}$$

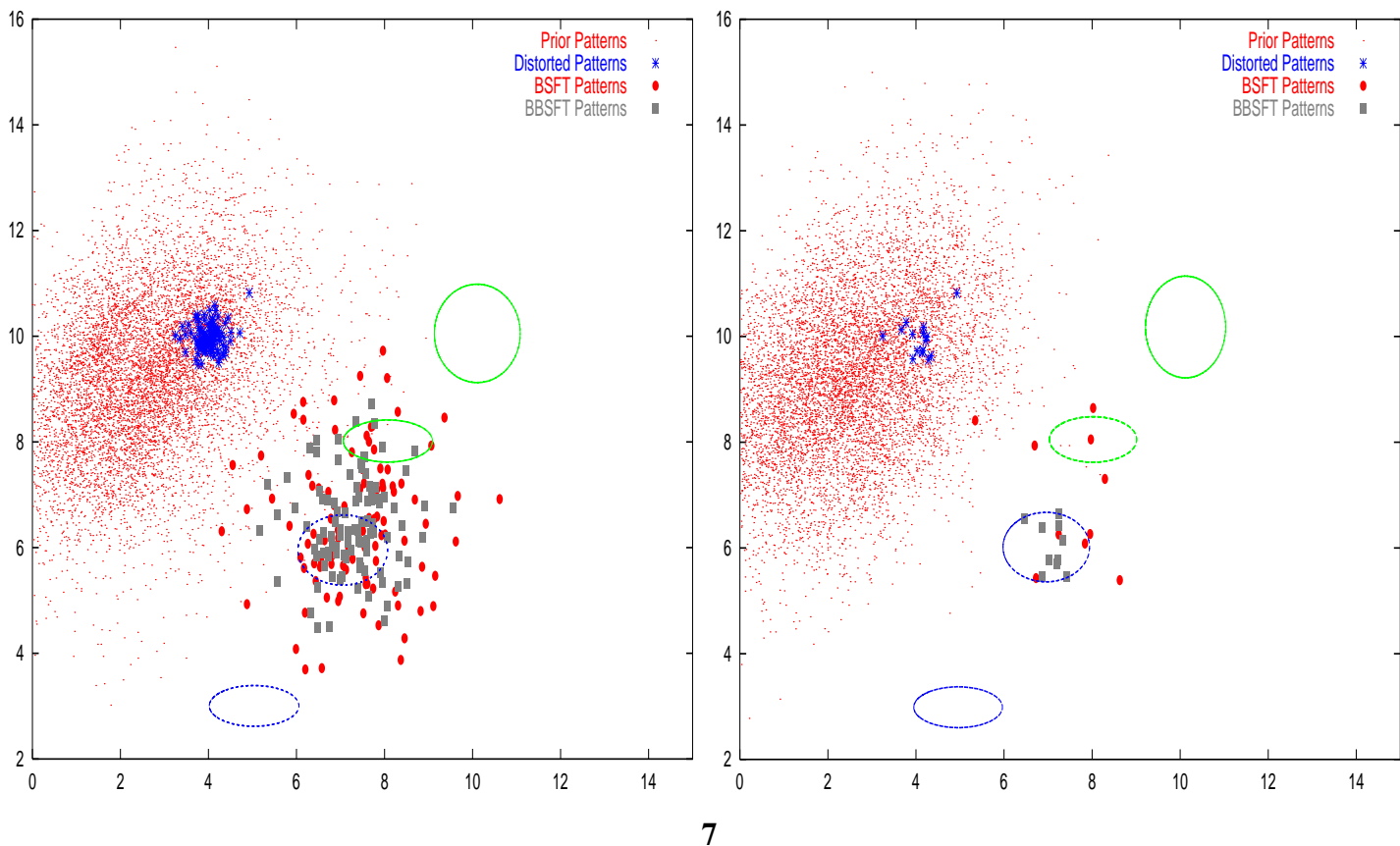
5

Bayesian Blind Stochastic Feature Transformation



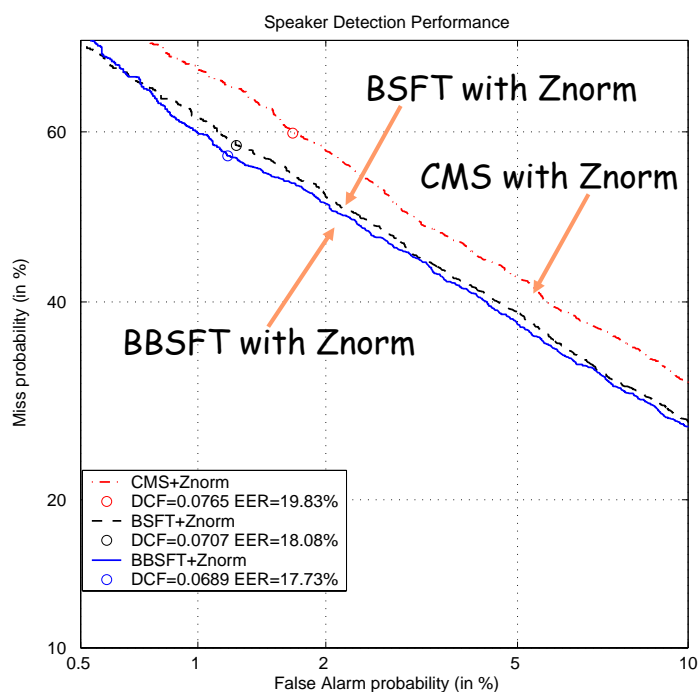
6

A Two-Dimensional Example



Results – equal error rate & DET plots

Verification Utterance Length (sec.)	CMS	BSFT	Bayesian BSFT
1	24.87	23.19	22.96
2	19.83	18.08	17.73
4	14.96	13.18	12.98
8	12.05	10.43	10.65
16	11.33	9.21	9.34
whole utterance	10.79	8.91	9.05



Experiments

- 2001 NIST speaker recognition evaluation set.
- 174 target speakers (74 male and 100 female).
- Enrollment: approximately 2 minutes of speech.
- Verification: 2038 utterances (850 male and 1188 female).
- 5 sets of test utterances were extracted from the original verification utterances. The 5 sets consists of speech data of duration 1, 2, 4, 8, 16 seconds.
- Speaker and background models:
 - MFCC + Δ MFCC
 - A universal background model with 1024 centers.
 - Speaker models: adapted from the gender-independent background model using MAP.

References

M. W. Mak, C. L. Tsang, and S. Y. Kung, “Stochastic feature transformation with divergence-based out-of-handset rejection for robust speaker verification,” *EURASIP J. on Applied Signal Processing*, vol. 4, pp. 452–465, 2004.

K.K. Yiu, M.W. Mak, and S.Y. Kung, “Blind stochastic feature transformation for channel robust speaker verification,” *J. of VLSI Signal Processing*, to appear.

D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

“The NIST year 2001 speaker recognition evaluation plan,” in <http://www.nist.gov/speech/tests/spk/2001/doc>.

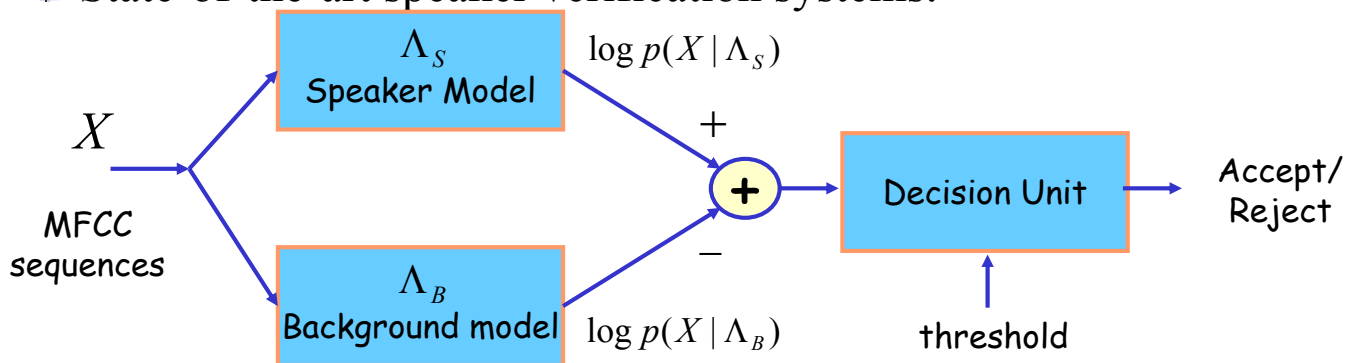
References

- M. W. Mak, C. L. Tsang, and S. Y. Kung, “Stochastic feature transformation with divergence-based out-of-handset rejection for robust speaker verification,” *EURASIP J. on Applied Signal Processing*, vol. 4, pp. 452-465, 2004.
- K. K. Yiu, M. W. Mak, and S. Y. Kung, “Blind stochastic feature transformation for channel robust speaker verification,” *J. of VLSI Signal Processing*, to appear.
- D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing*, vol. 10, pp 19-41,2000.
- “The NIST year 2001 speaker recognition evaluation plan,” in <http://www.nist.gov/speech/tests/spk/2001/doc>.

11

Spectral Feature-Based System

- ✚ State-of-the-art speaker verification systems:



- ✚ Telephone-based speaker verification system

- ✚ Suffer from performance degradation because of handset mismatch: different handsets will be used in enrollment and verification sessions
- ✚ The stochastic feature transformation (SFT) proposed by *Mak and Kung, ICASSP'02*, is adopted to improve the system robustness

CHANNEL ROBUST SPEAKER VERIFICATION VIA BAYESIAN BLIND STOCHASTIC FEATURE TRANSFORMATION

K. K. Yiu, M. W. Mak



Center for Multimedia Signal Processing
Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University

S.Y. Kung



Dept. of Electrical Engineering
Princeton University
USA