Missing-Feature Approaches in Speech Recognition

ShihHsiang 2006

References

- B. Raj and R.M. Stern, “Missing-Feature Approaches in Speech Recognition “, IEEE Signal Processing Magazine, 2005
- M.P. Cooke, P.G. Green, and M.D. Crawford, “Handling missing data in speech recognition” in Proc. ICSLP, 1994
- M.P. Cooke, A. Morris, and P.D. Green, “Missing data techniques for robust speech recognition,” in Proc. ICASSP,1997.
- M.P. Cooke, P.G. Green, L. Josifovski, and A. Vizinho, “Robust ASR with unreliable data and minimal assumptions,’’ in Proc., Robust’99, 1999
- M.P. Cooke, P.G. Green, L. Josifovski, and A. Vizinho, “Robust Automatic Speech Recognition with missing and unreliable acoustic data,’’ Speech Communication,, 2000.
- B. Raj, “Reconstruction of incomplete spectrograms for robust speech recognition,”, Ph.D. dissertation, ECE Dept., Carnegie Mellon Univ., Pittsburgh, PA, Apr. 2000.
- B. Raj, M.L. Seltzer, and R.M. Stern, “Reconstruction of missing features for robust speech recognition,’’ Speech Communication, 2004.

Outline

- Introduction
- Missing feature approaches
  - Classifier-compensation method
    - Data Imputation
    - Marginalization
  - Feature-compensation method
    - Correlation-based reconstruction
    - Cluster-based reconstruction
- Identification of unreliable components
- Conclusions

Introduction (1)

- Automatic speech recognition (ASR) system is still adversely affected by noise and other sources of acoustical variability
  - There are a lots of algorithms that have been developed to cope with the effect of additive noise
    - The drawback of these approaches
      - Most of them assume the noise is stationary
      - Only effective in the context of their intended purposes
- Conventional environmental compensation provides only limited benefit for these problems even today

Introduction (2)

- The missing feature approaches is based on the exploitation of the inherent redundancy in the speech signal
  - Speech that has undergone excision of spectral bands or short temporal regions remains intelligible
- The advantages of Missing-Features approaches
  - Make no assumptions about the corrupting noise
  - Do not need to have a knowledge about noise
  - Remarkable robust to high levels of noise corruption

The problems in missing feature

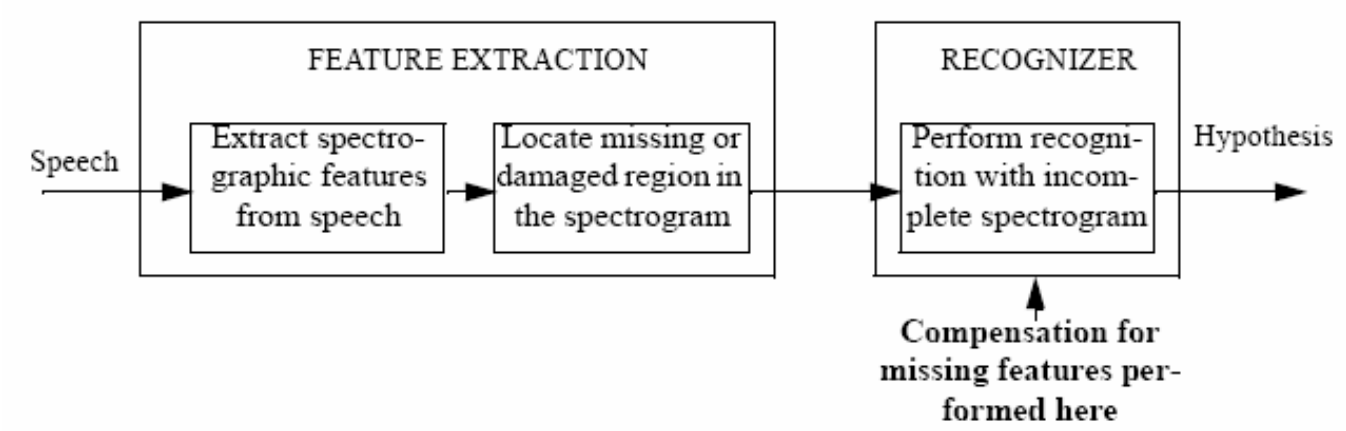
- There are two problems in missing feature approaches
  - The identification of unreliable components
    - Derived from computational auditory scene analysis of the signal
    - Depend in some manner on tracking or measuring the corrupting
  - The classification problem in general is to assign an observation vectors *x* to a class *C.* But in the missing data case, some components of ***x*** are unreliable or unavailable
    - Likelihood *f(x|c)* cannot be evaluated in the normal manner
    - Two uncertainty conditions are considered
      - Complete ignorance of the unreliable values
      - Knowledge of the interval within which the true data lie

Different approaches of missing feature theory

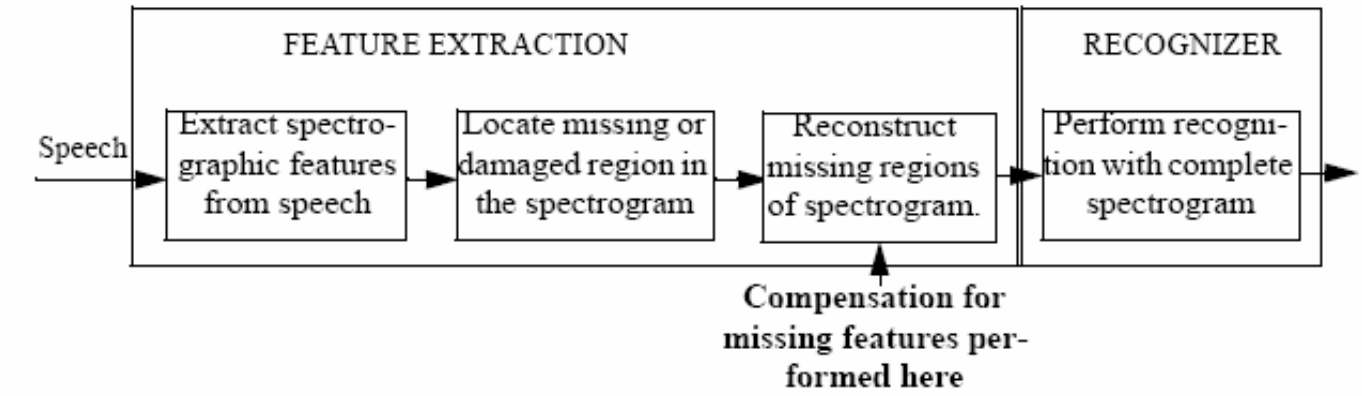
- Classifier-compensation method (Sheffield Group)
  - Data Imputation
    - Estimate unreliable components from the reliable components
  - Marginalization
    - Classify based solely on reliable components
- Feature-compensation method (CMU Group)
  - Correlation-based reconstruction
    - Reconstruct unreliable component from neighborhood reliable components
  - Cluster-based reconstruction
    - Unreliable component are estimated from the distribution of the cluster

Different approaches of missing feature theory (cont.)

- Classifier-compensation method



- Feature-compensation method



Classifier-compensation method

Architecture and assumptions

- Under Continuous Density Hidden Markov Model Speech Recognition (CDHMM)
  - Each chosen speech unit is represented by a trained HMM with a number of states
  - Each state is characterized by a multivariate mixture Gaussian distribution over the components of the acoustic observation vector *x*

$$f(x \mid C_i) = \sum_{k=1}^{M} P(k \mid C_i) f(x \mid k, C_i)$$

M Gaussians with diagonal-only covariance structure

Data Imputation (1)

- Compute the distribution of the unreliable parts of the feature vector using the reliable components and the joint p.d.f.
  - Then some representative value for the unreliable feature can be chosen using this distribution
    - Usually it is the mean of the conditional distribution
- If $f(x \mid C)$ denotes the distribution of the complete vector in a given class C, then unreliable items are replaces by values drawn from $f(x_u \mid x_r, C)$

Data Imputation (2)

$$f(x_u \mid x_r, C_i) = \frac{f(x_u, x_r \mid C_i)}{f(x_r \mid C_i)}$$

M mixture Gaussians

$$= \frac{\sum_{k=1}^{M} P(k \mid C_i) f(x_u, x_r \mid k, C_i)}{f(x_r \mid C_i)}$$

Independent (diagonal covariance)

$$= \frac{\sum_{k=1}^{M} P(k \mid C_i) f(x_r \mid k, C_i) f(x_u \mid k, C_i)}{f(x_r \mid C_i)}$$

$$= \sum_{k=1}^{M} \frac{P(k \mid C_i) f(x_r \mid k, C_i)}{f(x_r \mid C_i)} f(x_u \mid k, C_i)$$

Chain rule

$$= \sum_{k=1}^{M} \frac{f(x_r, k \mid C_i)}{f(x_r \mid C_i)} f(x_u \mid k, C_i)$$

Chain rule

$$= \sum_{k=1}^{M} \frac{P(k \mid x_r, C_i) f(x_r \mid C_i)}{f(x_r \mid C_i)} f(x_u \mid k, C_i)$$

Bayes' rule

$$= \sum_{k=1}^{M} P(k \mid x_r, C_i) f(x_u \mid k, C_i)$$

$$E_{x_u \mid x_r, C_i}\{x_u\} = \int f(x_u \mid x_r, C_i) x_u \, dx_u$$

$$= \sum_{k=1}^{M} P(k \mid x_r, C_i) \int f(x_u \mid k, C_i) x_u \, dx_u$$

$$= \sum_{k=1}^{M} P(k \mid x_r, C_i) \mu_{u \mid k, C_i}$$

Data Imputation (3)

- Hence, data imputation estimates unreliable component of the observation vector using

$$\hat{x}_{u,i} = \sum_{k=1}^{M} P\left(k \mid x_r, C_i\right) \mu_{u|k,C_i}$$

where

$$P\left(k \mid x_r, C_i\right) = \frac{P\left(k \mid C_i\right) f\left(x_r \mid k, C_i\right)}{\sum_{k=1}^{M} P\left(k \mid C_i\right) f\left(x_r \mid k, C_i\right)}$$

Marginalization (1)

- Compute the HMM state output probabilities using a reduced distribution based solely on reliable components
  - Require the marginal determined by integrating over all missing components

$$f(x_r \mid C_i) = \int f(x_r, x_u \mid C_i)\, dx_u$$

$$= \int \sum_{k=1}^{M} P(k \mid C_i) f(x_r, x_u \mid k, C_i)\, dx_u$$

$$= \sum_{k=1}^{M} P(k \mid C_i) f(x_r \mid k, C_i) \int f(x_r \mid k, C_i)\, dx_u$$

without any knowledge about the bound of unreliable data

$$\int f(x_r \mid k, C_i)\, dx_u = 1$$

Has knowledge about the bound of unreliable data

$$\int_{x_{low}}^{x_{high}} f(x_r \mid k, C_i)\, dx_u$$

Marginalization (2)

Marginalization

$$f(x_r \mid C_i) = \int f(x_r, x_u \mid C_i) dx_u$$
$$= \int \sum_{k=1}^{M} P(k \mid C_i) f(x_r, x_u \mid k, C_i) dx_u$$
$$= \sum_{k=1}^{M} P(k \mid C_i) f(x_r \mid k, C_i) \int f(x_r \mid k, C_i) dx_u$$
$$= \sum_{k=1}^{M} P(k \mid C_i) f(x_r \mid k, C_i)$$

Bounded Marginalization

$$f(x_r \mid C_i) = \int_{x_{low}}^{x_{high}} f(x_r, x_u \mid C_i) dx_u$$
$$= \int_{x_{low}}^{x_{high}} \sum_{k=1}^{M} P(k \mid C_i) f(x_r, x_u \mid k, C_i) dx_u$$
$$= \sum_{k=1}^{M} P(k \mid C_i) f(x_r \mid k, C_i) \int_{x_{low}}^{x_{high}} f(x_r \mid k, C_i) dx_u$$

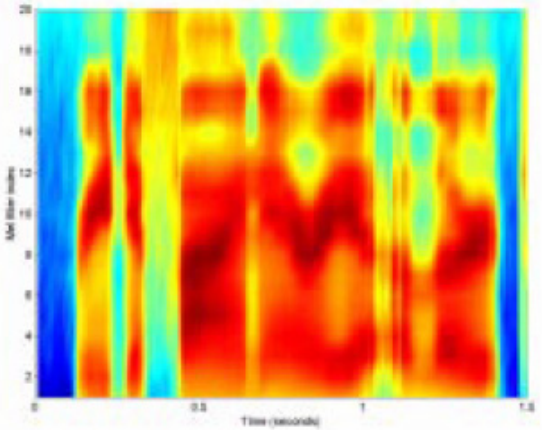
The integral can be evaluated using error function

$$\int_{x_{low}}^{x_{high}} f(x_u \mid k, C_i) dx_u$$
$$= \frac{1}{2} \left[erf\left(\frac{x_{high,u} - \mu_{u,k,i}}{\sqrt{2\sigma_{u,k,i}}} \right) - erf\left(\frac{x_{low,u} - \mu_{u,k,i}}{\sqrt{2\sigma_{u,k,i}}} \right) \right]$$

Feature-compensation method

Background Information

- Reconstruct complete spectrograms from the incomplete ones
  - Estimate the true value of the unreliable spectrographic components from the reliable components
- Notation



$Y(t)$ noisy spectral vector at t-th frame

$Y_r(t)$ reliable component of Y

$Y_u(t)$ unreliable component of Y

$X(t)$ true spectral vector at t-th frame

$X_r(t)$ reliable component of X

$X_u(t)$ unreliable component of X

$$X_r(t) \approx Y_r(t)$$

$$X_u(t) \leq Y_u(t)$$

Background Information

Multivariate Gaussian Distribution

- When X=(X1,…, XL) is a L-dimensional random vector, the multivariate Gaussian pdf has the form

$$f(X = x \mid \mu, \Sigma) = N(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{L}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right)$$

- Conditional distributions
  - If X1 conditional on X2 = a is multivariate normal

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

$$\overline{\mu} = \mu_1 + \underline{\Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2)}$$

mean shift

$$\overline{\Sigma} = \Sigma_{11} - \underline{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}$$

regression coefficients

Background Information

Multivariate Gaussian Distribution (cont.)

observed data

Assume that the element of the complete data $\mathbf{X}=[\mathbf{X_o},\mathbf{X_m}]$

missing data

Let $P(X;\mu,\Theta)$ be a Gaussian distribution with mean vector μ and covariance matrix Θ

The distributions of $P(X_o;\mu,\Theta)$ and $P(X_m;\mu,\Theta)$ also be Gaussian

and also given $\mu=[\mu_o,\mu_m]$ $\Theta=\begin{bmatrix}\Theta_{oo} & \Theta_{om}\\ \Theta_{mo} & \Theta_{mm}\end{bmatrix}$, we can get following equation

Θ_{om} is the cross covariance between X_o and X_m

$$P(X_m \mid X_o,\mu,\Theta)=C\exp(-0.5(X_m-\mu_m-\Theta_{mo}\Theta_{oo}^{-1}(X_o-\mu_o))^T$$
$$(\Theta_{mm}-\Theta_{mo}\Theta_{oo}^{-1}\Theta_{om})^{-1}$$
$$(X_m-\mu_m-\Theta_{mo}\Theta_{oo}^{-1}(X_o-\mu_o))))$$

Background Information

Multivariate Gaussian Distribution (cont.)

- The Conditional Gaussian

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}BQ^{-1}CA^{-1} & -A^{-1}BQ^{-1} \\ -Q^{-1}CA^{-1} & Q^{-1} \end{bmatrix}$$

$$Q = D - CA^{-1}B$$

$$p(y \mid x) = \frac{p(x,y)}{p(x)}$$

$$= \frac{(2\pi)^{k/2}\left|\sum_{(X)}\right|^{1/2}}{(2\pi)^{(k+m)/2}\left|\sum_{(X,Y)}\right|^{1/2}} \times \frac{\exp\left\{-\frac{1}{2}\begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}^T \begin{bmatrix} \sum_{xx} & \sum_{xy} \\ \sum_{yx} & \sum_{yy} \end{bmatrix}^{-1} \begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}\right\}}{\exp\left\{-\frac{1}{2}(x-\mu_x)^T \sum_{(X)}^{-1}(x-\mu_x)\right\}}$$

$$= \frac{1}{\sqrt{(2\pi)^m \left|\sum_{(X)}\right| / \left|\sum_{(X,Y)}\right|}} \times \exp\left\{-\frac{1}{2}\begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}^T \begin{bmatrix} \sum_{xx} & \sum_{xy} \\ \sum_{yx} & \sum_{yy} \end{bmatrix}^{-1} \begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix} + (x-\mu_x)^T \sum_{(X)}^{-1}(x-\mu_x)\right\}$$

$$= \frac{1}{\sqrt{(2\pi)^m \left|\sum_{(X)}\right| / \left|\sum_{(X,Y)}\right|}} \times \exp\left\{\left(y-\mu_y - \sum_{yx}\sum_x^{-1}(x-\mu_x)\right)^T \sum_{Y|X}^{-1}\left(y-\mu_y - \sum_{yx}\sum_x^{-1}(x-\mu_x)\right)\right\}$$

Background Information

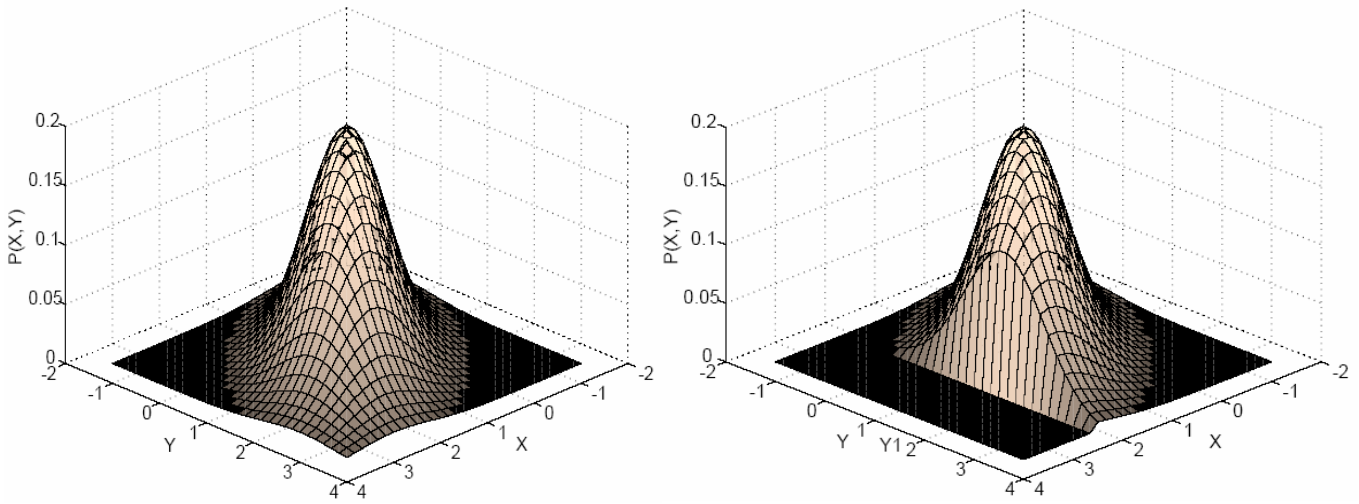
Maximum A-Posteriori (MAP) Estimation

In MAP estimation the missing data are estimated to maximize their Likelihood, conditioned on the value of the observed data

$\hat{X}_m = \arg\max_{X_m} \{P(X_m \mid X_o; \phi)\}$ when $P(X_m \mid X_o; \phi)$ is Gaussian

We get

$$\hat{X}_m = \arg\max_{X_m} \{P(X_m \mid X_o, \mu, \Theta)\} = X_m + \Theta_{mo}\Theta_{oo}^{-1}(X_o - \mu_o)$$



Correlation-based reconstruction (1)

- The sequence of spectral vectors that constitute the spectrogram of a clean speech signal are considered to be the output of a Gaussian wide-sense stationary (WSS)
  - WSS gives us the following properties

$$\mu(t_1,k) = \mu(t_2,k) = \mu(k)$$

$$c(t,t+\tau,k_1,k_2) = c(t_1,t_1+\tau,k_1,k_2) = c(\tau,k_1,k_2)$$

    - Mean is not depend on where it occurs
    - Covariance between the component of two vector depends only on the distance
  - The relative covariance between any two components is given by

$$r(\tau,k_1,k_2) = \frac{c(\tau,k_1,k_2)}{\sqrt{c(0,k_1,k_1)c(0,k_2,k_2)}}$$

Correlation-based reconstruction (2)

- The means of the components of the spectral vectors and the various covariance parameters can be learnt from the training corpus

$$\mu(k) = \frac{1}{\sum_j N_j} \sum_j \sum_{t=1}^{N_j} X^j(t,k)$$

j-th training utterance

# of frames in j-th training utterance

$$c(\tau, k_1, k_2) = \frac{1}{\sum_j (N_j - \tau)} \sum_j \sum_{t=1}^{N_j} \left[\left(X^j(t,k_1) - \mu_{k_1} \right) \times \left(X^j(t+\tau,k_2) - \mu_{k_2} \right) \right]$$

Correlation-based reconstruction (3)

- Now, we can estimate $X_u(t)$ to reconstruct $X(t)$ completely
  - Construct a neighborhood vector $Y_n(t)$ from all reliable components of the spectrogram that have a relative covariance greater than a threshold value with at least one of the component of $X_u(t)$
  - $X_u(t)$ is now estimated as

$$\hat{X}_u(t) = \arg\max_{x_u} \{P(X_u(t), X_u(t) \le Y_u(t) \mid Y_n(t))\}$$

bounded MAP estimate

  - It can be shown that $P(X_u(t) \mid Y_n(t))$, the distribution of $X_u(t)$ conditioned on $X_n(t)$ being equal to $Y_n(t)$, is a Gaussian with

$$\mu(t) + C_{un}(t) C_{nn}^{-1}(t) (Y_n(t) - \mu_n(t))$$

Correlation-based reconstruction (4)

The estimation procedure can be stated as follows

1. Initialize $\overline{X}_u(t,k) = Y_u(t,k), 1 \le k \le K$ where *K* is the total number of components in $X_u(t)$
2. For each of the *K* components

   2a. Compute the MAP estimate

   $$\widetilde{X}_u(t,k) = \arg\max_{x_u(t,k)} \left\{ P\left(X_u(t,k) \mid Y_n(t), \overline{X}_u(t,j) \forall j, j \ne k \right) \right\}$$

   2b. Compute the bounded MAP estimate from the MAP estimate as

   $$\overline{X}_u(t,k) = \min\left(\widetilde{X}_u(t,k), Y_u(t,k) \right)$$

3. If all $\overline{X}_u(t)$ estimate have converged, set $\hat{X}_u(t) = \overline{X}_u(t) \forall k$ to obtain $X_u(t)$, else go back to Step 2

Correlation-based reconstruction (5)

$Y_u(2)$ is constructed as

$$Y_u(2) = [Y(2,1), Y(2,3)]^T$$

The neighborhood vector $Y_n(2)$ is constructed of all the components $Y(t,k)$, such that either $r(t-2,1,k) \geq 0.5$, or $r(t-2,3,k) \geq 0.5$. These are represented by the components with the thick outlines. This gives us

$$Y_n(2) = [Y(1,1), Y(1,3), Y(2,2), Y(3,1), Y(3,2)]^T$$

The mean vectors for $X_n(2)$ and $X_u(2)$, the clean speech counterparts of $Y_n(2)$ and $Y_u(2)$, are constructed as

$$E[X_n(2)] = \mu_n(2) = [\mu(1), \mu(3), \mu(2), \mu(1), \mu(2)]^T$$

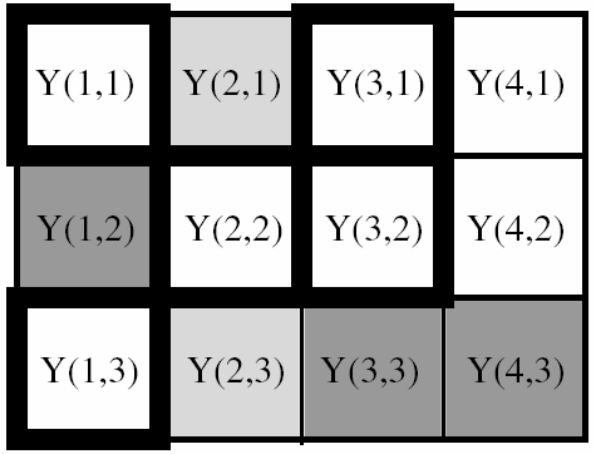
$$E[X_u(2)] = \mu_u(2) = [\mu(1), \mu(3)]$$

The autocovariance matrix of $X_n(2)$ is a 5×5 matrix constructed as

$$C_{nn}(2) = \begin{bmatrix} c(0,1,1) & c(0,1,3) & c(1,1,2) & c(2,1,1) & c(2,1,2) \\ c(0,3,1) & c(0,3,3) & c(1,3,2) & c(2,3,1) & c(2,3,2) \\ c(-1,2,1) & c(-1,2,3) & c(0,2,2) & c(1,2,1) & c(1,2,2) \\ c(-2,1,1) & c(-2,1,3) & c(-1,1,2) & c(0,1,1) & c(0,1,2) \\ c(-2,2,1) & c(-2,2,3) & c(-1,2,2) & c(0,2,1) & c(0,2,2) \end{bmatrix}$$

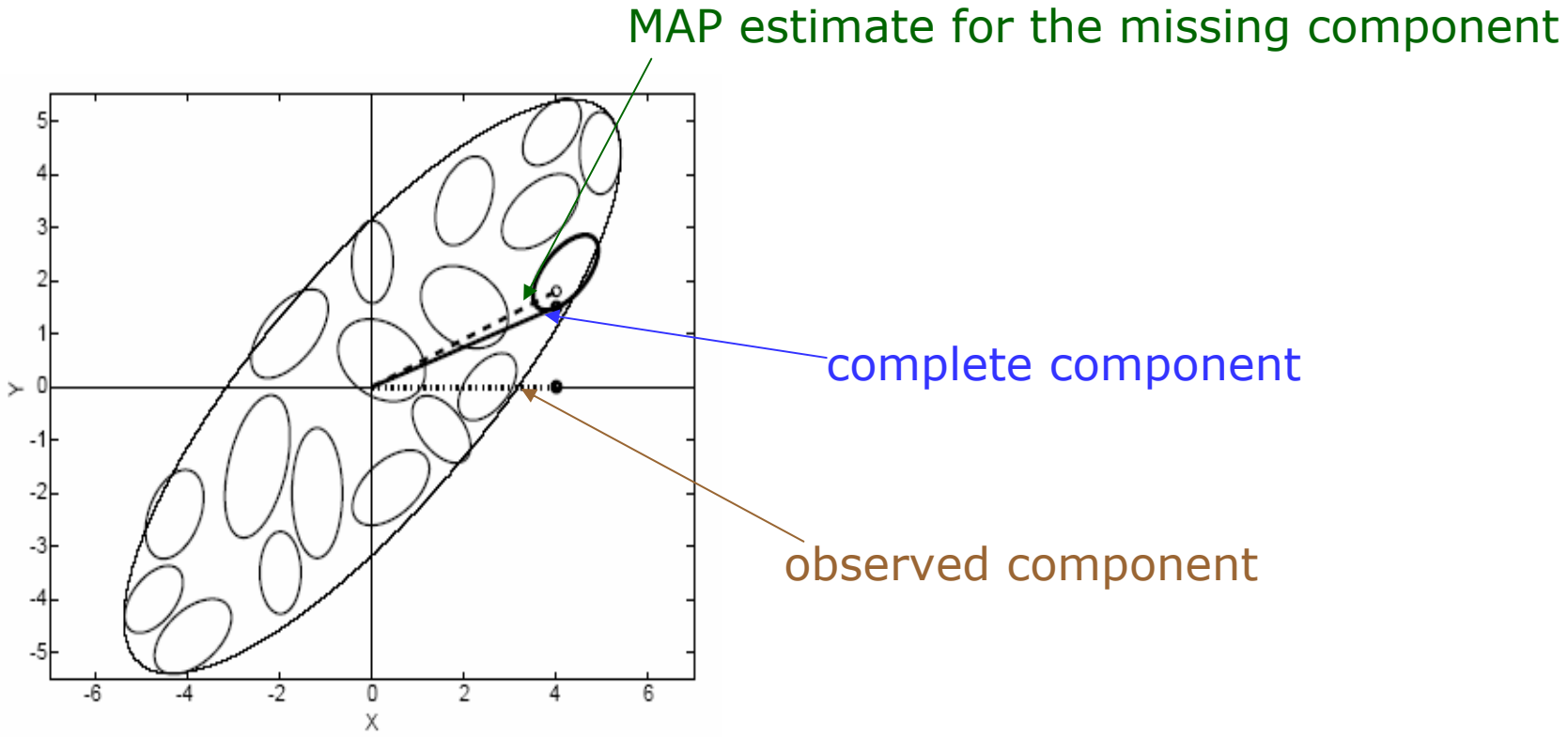
The cross covariance between $X_u(2)$ and $X_n(2)$ is a 2×5 matrix constructed as

$$C_{un}(2) = \begin{bmatrix} c(-1,1,1) & c(-1,1,3) & c(0,1,2) & c(1,1,1) & c(1,1,2) \\ c(-1,3,1) & c(-1,3,3) & c(0,3,2) & c(1,3,1) & c(1,3,2) \end{bmatrix}^T$$



Cluster-based reconstruction (1)

- The unreliable components of the true spectral vector can be estimated by determining the cluster to which the vector belongs, and estimating them from the distribution of the cluster



Cluster-based reconstruction (2)

- In cluster-based reconstruction, the spectral vectors of clean speech are assumed to be segregated into a number of cluster
  - Each cluster is assumed to have a Gaussian distribution
- The distribution of the *k*-th cluster is thus given by

$$P(X \mid k) = \frac{\exp\left(-\frac{1}{2}(X-\mu_k)^T \Theta_k^{-1}(X-\mu_k) \right)}{\sqrt{(2\pi)^d |\Theta_k|}}$$

- The overall distribution of spectral vectors is thus a mixture Gaussian given by

$$P(X) = \sum_{k=1}^{K} c_k P(X \mid k) = \sum_{k=1}^{K} \frac{c_k}{\sqrt{(2\pi)^d |\Theta_k|}} \times \exp\left(-\frac{1}{2}(X-\mu_k)^T \Theta_k^{-1}(X-\mu_k) \right)$$

Cluster-based reconstruction (3)

The estimate for $X_u(t)$ obtained from the distribution of the *k*-th cluster, $\hat{X}_u^k(t)$ is given by

$$\hat{X}_u^k(t) = \arg\max_{X_u} \left\{ P\left(X_u(t), X_u(t) \le Y_u(t) \mid k, Y_r(t)\right) \right\}$$

The overall estimate of $X_u(t)$ is given by

$$\hat{X}_u(t) = \sum_{j=1}^{K} P\left(k \mid Y_r(t), X_u(t) \le Y_u(t)\right) \hat{X}_u^k(t)$$

where

$$P\left(k \mid Y_r(t), X_u(t) \le Y_u(t)\right) = \frac{c_k P\left(Y_r(t), X_u(t) \le Y_u(t) \mid k\right)}{\sum_{j=1}^{k} c_j P\left(Y_r(t), X_u(t) \le Y_u(t) \mid j\right)}$$

Cluster-based reconstruction (4)

$$P\left(Y_r(t), X_u(t) \le Y_u(t) \mid k\right)$$

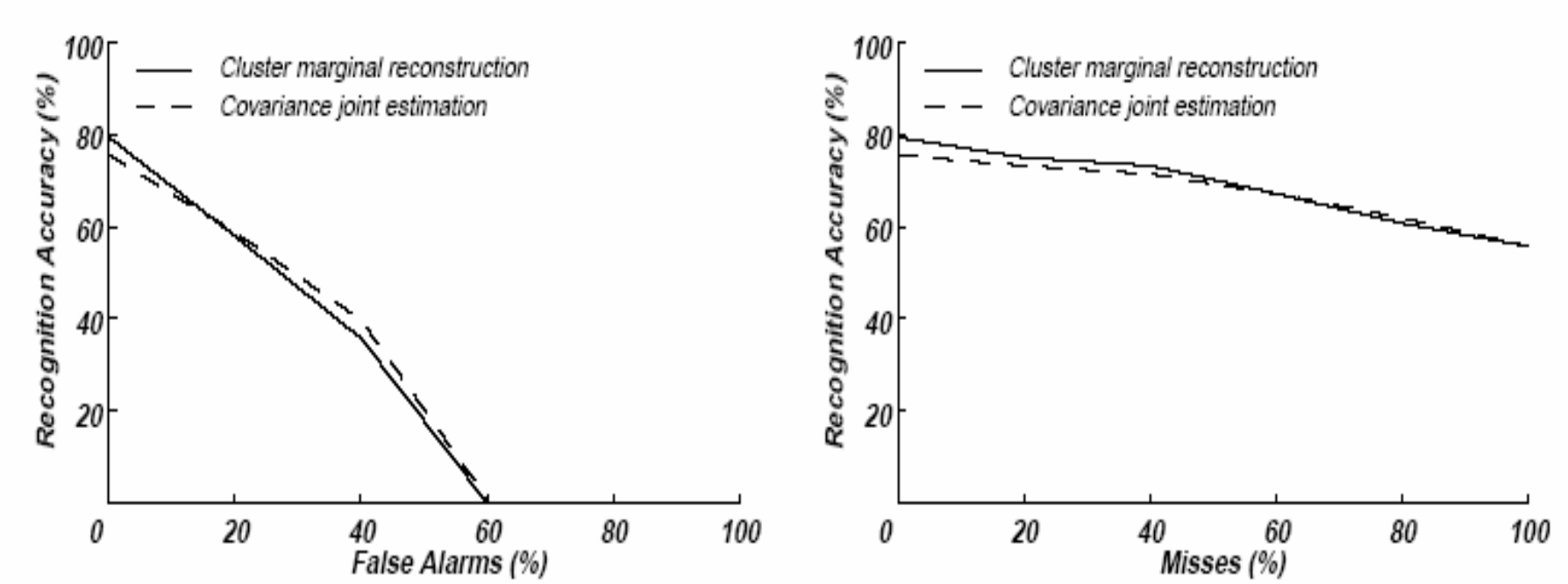
$$= \prod_{i \mid X(t,i) \in X_r(t)} \frac{1}{\sqrt{2\pi\theta_k(i)}} \times \exp\left(- \frac{\left(Y(t,i) - \mu_k(i)\right)^2}{2\theta_k(i)} \right)$$

$$\times \prod_{i \mid X(t,i) \in X_u(t)} \int_{-\infty}^{Y(t,i)} \frac{1}{\sqrt{2\pi\theta_k(i)}} \times \exp\left(- \frac{\left(Y(t,i) - \mu_k(i)\right)^2}{2\theta_k(i)} \right) dX(t,i)$$

Identification of unreliable components (1)

- The most difficult aspect of missing feature methods is identifying unreliable spectral components
- The estimation can be performed in multiple ways
  - Estimate the SNR of each spectral component
  - Classify unreliable components directly using some other criteria in place of SNR
  - Or from perceptually-motivated criteria
- The ability of missing feature methods depends critically on the accuracy of the spectrographic masks used
  - False alarm: reliable element declared as unreliable
  - Miss: unreliable element tagged as reliable

Identification of unreliable components (2)



- The recognition performance degrades very quickly with increasing fraction of false alarms
- The sensitivity of missing-feature methods to misses is not so much

Identification of unreliable components (3)

- Based on negative energy

  If the observed magnitude in any frame is denoted by $|s+n|$ and the estimated noise spectrum by $\hat{n}$, then the negative energy criterion drops spectral regions from the mask if

$$|s+n| - \hat{n} < 0$$

- Based on SNR criteria

  The "cleaned" speech $\hat{s}$ is obtained by $|s+n| - \hat{n}$, The SNR criterion treats data as unreliable when the estimated SNR is negative

$$\log\left(\frac{\hat{s}^2}{\hat{n}^2}\right) < 0 \text{ or } \hat{s}^2 < \hat{n}^2$$

Identification of unreliable components (4)

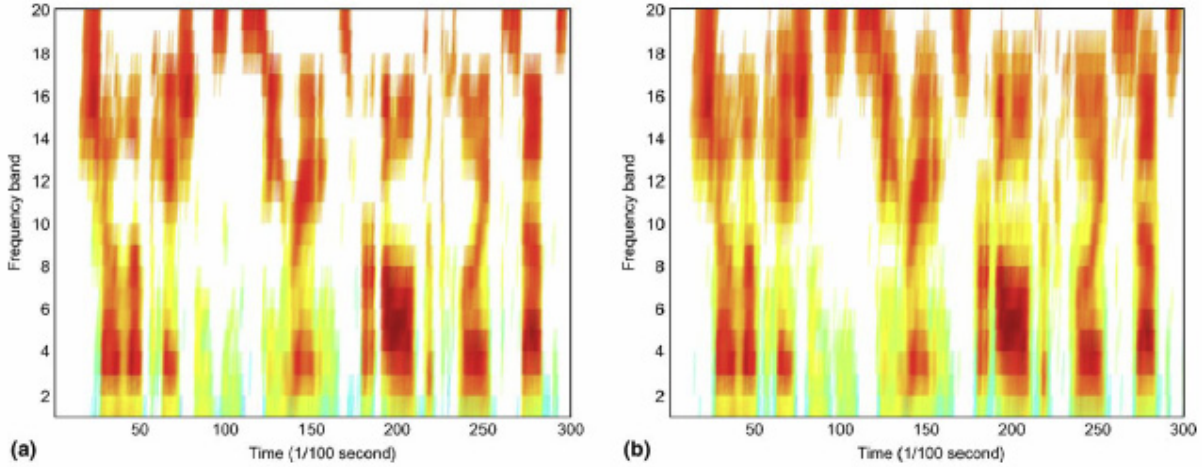
- Based on Bayesian classifier
  - A set of feature is computed for every time-frequency location of the spectrogram
    - Features are designed that exploit the characteristics of the speech signal itself, rather than measurements of the corrupting noise
    - These features are then input to a conventional Bayesian classifier to determine whether a specific time-frequency component is reliable or not
  - The feature for any time-frequency location (*t*,*k*) include
    - The ratio of the first and second autocorrelation peaks
    - The ratio of the total energy in the *k*-th frequency band to the total energy of all frequency bands
    - The kurtosis of the signal samples within the *t*-th frame of speech
    - The variance of the spectrographic components adjoining (*t*,*k*)
    - The ratio of the energy within $Y(t,k)$ to the estimated energy of the noise $N(t,k)$

Identification of unreliable components (5)

– Separate classifiers are trained for voiced and unvoiced speech, and for each frequency band
  - All distribution are modeled as a mixture of Gaussians, the parameters of which are learned from the feature vectors
  - The a priori probabilities of the reliable and unreliable classes for each classifier are also learned from training data
– The (*t*,*k*)-th time-frequency location is classified as reliable if

$$P_{V,k}(reliable)P_{V,k}(F(t,k)\,|\,reliable)$$
$$> P_{V,k}(unreliable)P_{V,k}(F(t,k)\,|\,unreliable)$$

SNR criteria



Bayesian classifier

Experiments (1)

- Experiments were conducted using the Resource Management database
  - with the Sphinx-3 HMM based speech recognition system
  - with 2000 tied state distribution with Gaussian state output densities
  - speech corrupted by white noise and music noise
- Speech recognition performance does not degrade significantly when a randomly dropped 80% of the elements

Experiments (2)

Recognition with log spectra

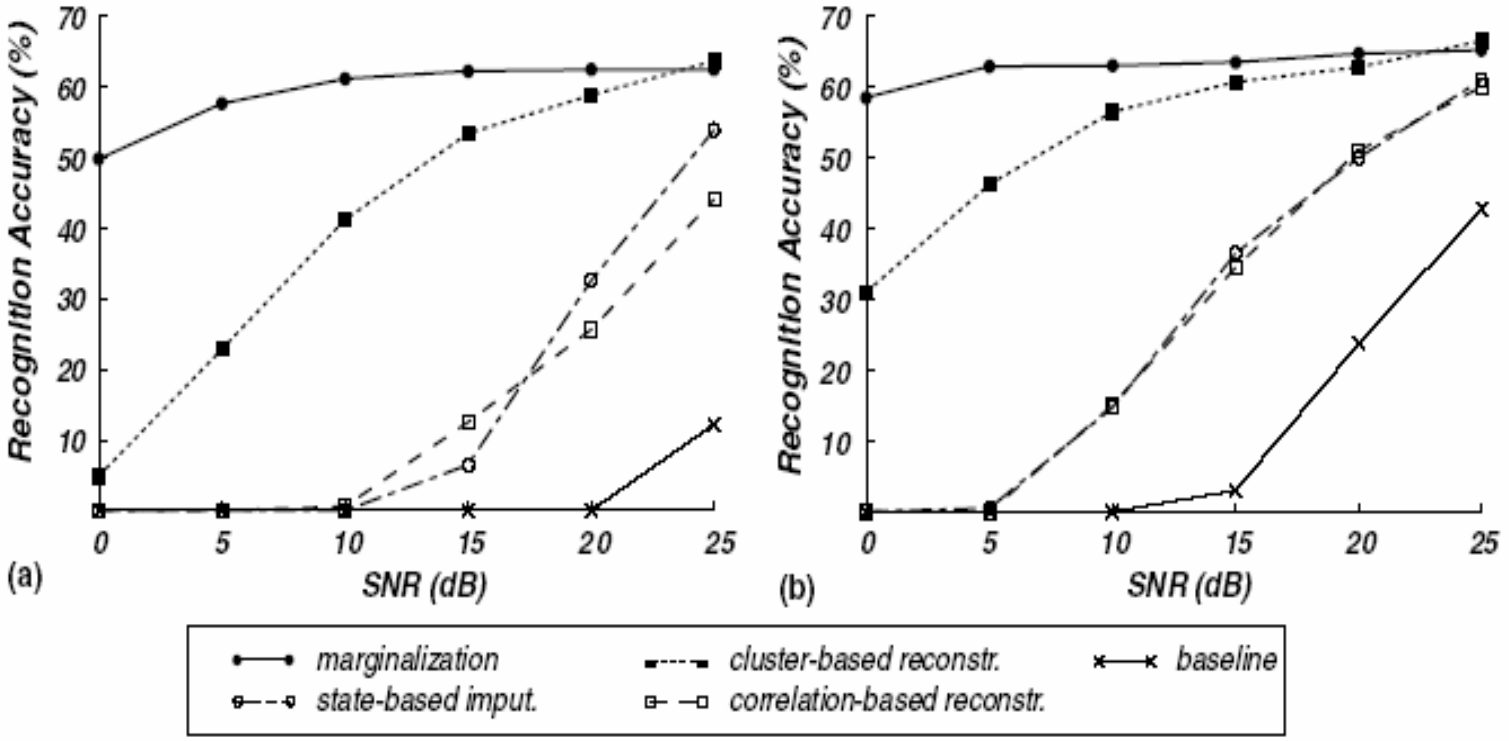


Fig. 6. Recognition performance of various missing feature methods on noisy speech, when unreliable components are located on the basis of their SNR values: (a) speech corrupted by white noise; (b) speech corrupted by music. In both figures the baseline recognition performance with the uncompensated noisy speech is also shown.

Experiments (3)

Recognition with cepstra

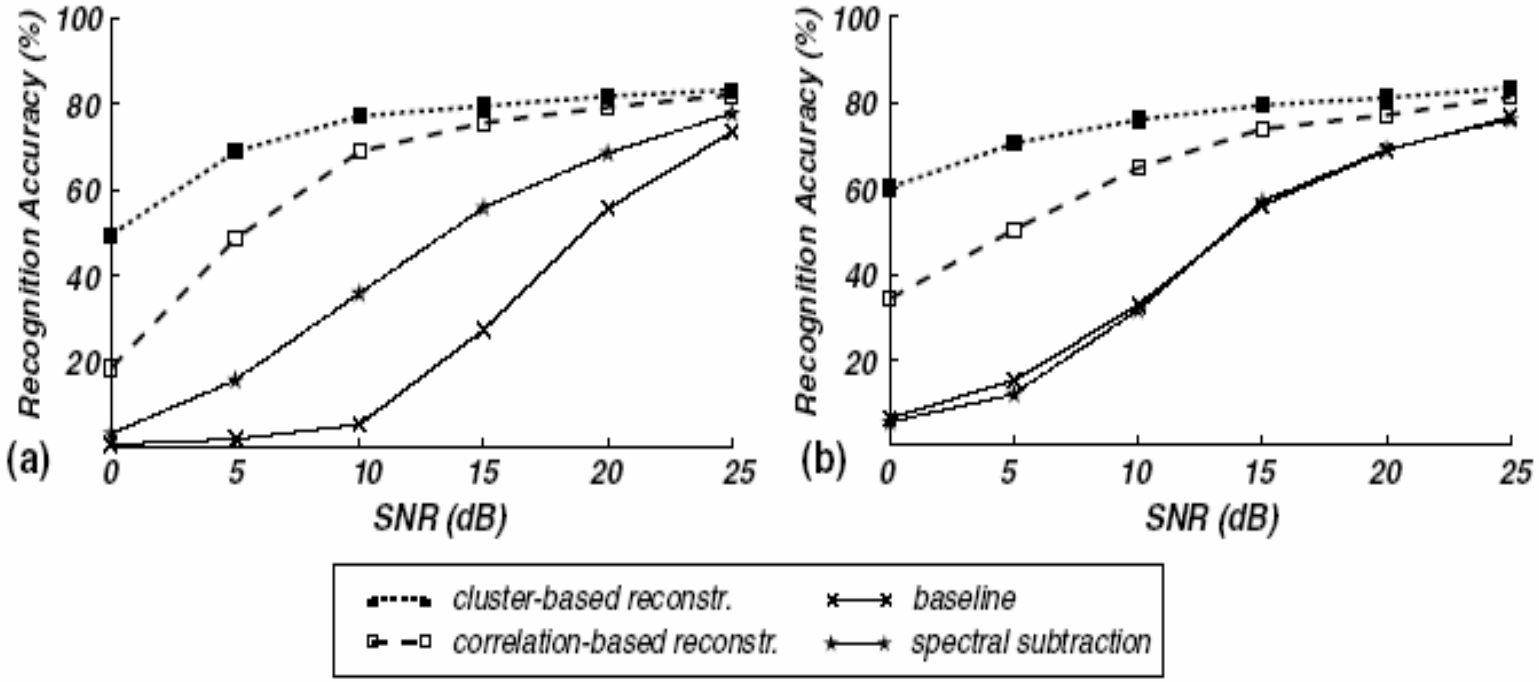


Fig. 8. Recognition performance obtained with cepstra derived from spectrograms reconstructed with prior knowledge of the identity of unreliable components. (a) Recognition on speech corrupted by white noise to various SNRs. (b) Recognition on speech corrupted by music to various SNRs. In both cases the baseline performance with uncompensated noisy speech, and the performance with a typical noise compensation algorithm, spectral subtraction, are shown for contrast.

Experiments (4)

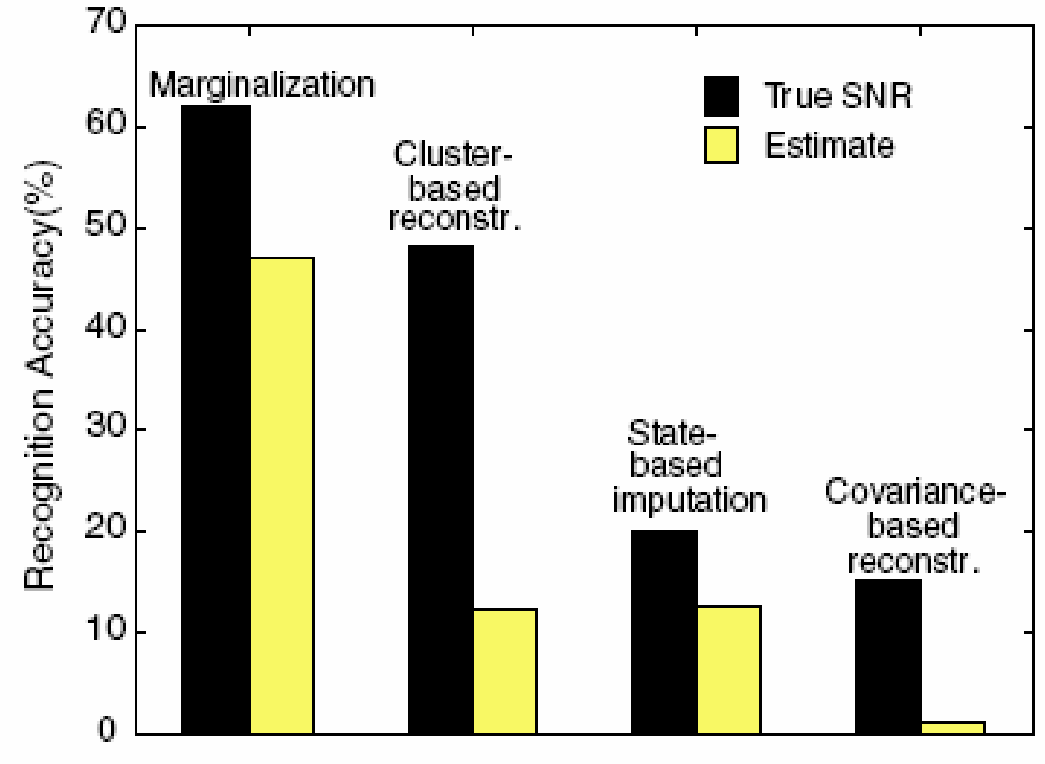


Fig. 9. Comparisons of recognition accuracy obtained when unreliable components are identified based on knowledge of their true SNR with accuracy obtained when the positions of unreliable components are estimated.

Experiments (5)

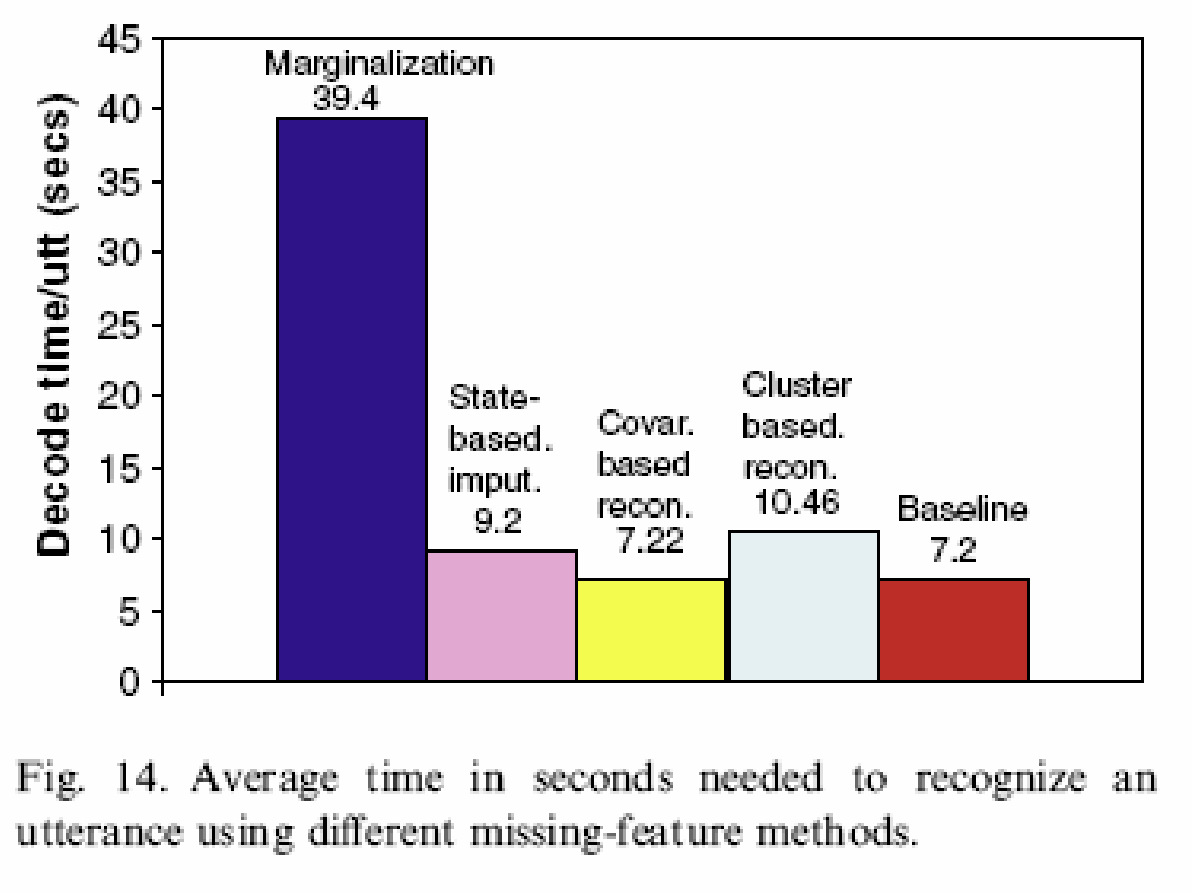


Fig. 14. Average time in seconds needed to recognize an utterance using different missing-feature methods.

Conclusions

- We reviewed the mathematics of four major missing feature techniques
  - Feature imputation
  - Cluster based reconstruction
  - Covariance based reconstruction
  - Classifier modification
- Missing feature approach seems a possible solution to improve the recognition performance
- It also makes no assumption about the noise and there is no requirement to retrain models for each noise condition