Joint Uncertainty Decoding for Noise Robust Speech Recognition

Jasha Droppo, Alex Acero, and Li Deng

Microsoft Research, One Microsoft Way, Redmond, Washington, USA

Presented by Howard

Outline

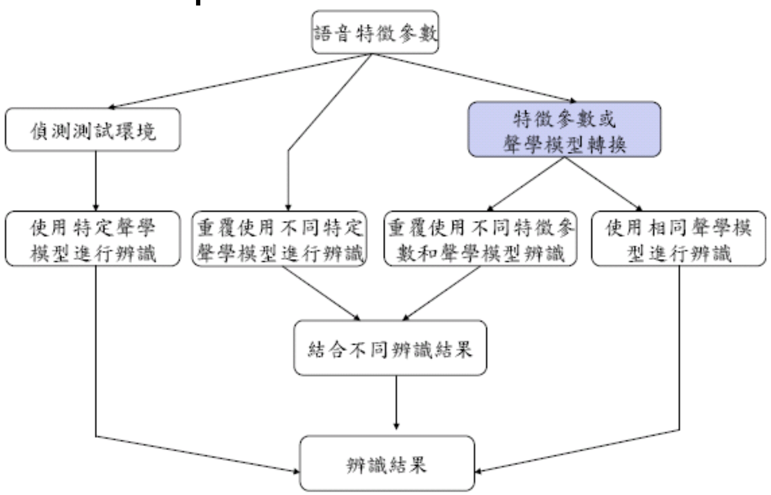
- What is robust?
- Feature-domain & model-domain
- Introduction to SPLICE
- What is uncertainty?
- Concept of uncertainty decoding
- Uncertainty Decoding with SPLICE
- Uncertainty with Joint Uncertainty Decoding

What is robust?

- We can say it is robust if it is hardly affected by extrinsic events.
  - Ex: A waterproof watch in water can still work as usual.
- For ASR
  - Speech recognition performance degrades in the presence of environmental noise, why?

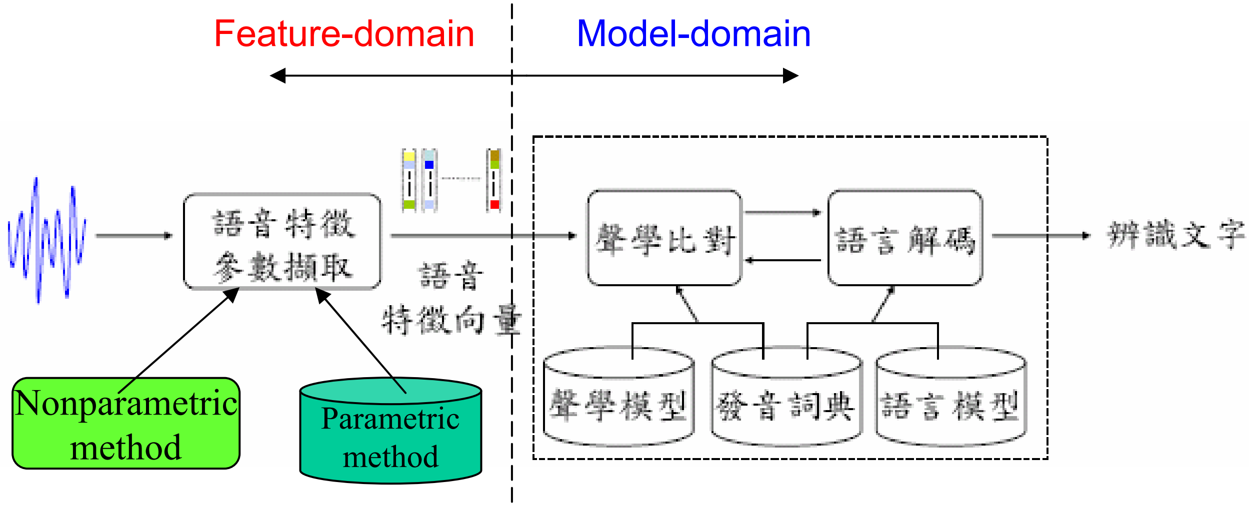
    The answer is the mismatch between training and test condition.
- Solution
  - There are tow main direction using different aspect to cut into this problem.

$$\begin{cases} \text{feature} - \text{domain} \begin{cases} \text{transformation} \\ \text{enhancement} \\ \text{adaptation} \end{cases} \\ \text{model} - \text{domain} \rightarrow \text{using specific model} \end{cases}$$



Feature-domain & Model-domain

- Where is part of feature-domain or model-domain?



Feature-domain: The outputs from feature extraction are all restored feature vectors without any other information to adapt acoustic model.

model-domain: The outputs from feature extraction may be not feature vectors but contains other information (ex: likelihood) to adapt acoustic model.

Introduction to SPLICE

- Stereo piecewise linear compensation for environment (SPLICE) takes advantage of seamlessly integrating into existing system, without a complete overhaul of existing code.

Total probability

$$\tilde{y}_t = \tilde{x}_t = \mathrm{E}[x_t \mid y_t] = \sum_k p(k \mid y_t)\mathrm{E}[x_t \mid y_t, k]$$

- Assuming that the difference between clean data and corrupted data can be compensated by each single Gaussian providing a linear compensation.

$$\mathrm{E}[x_t \mid y_t, k] \approx y_t + r_k$$

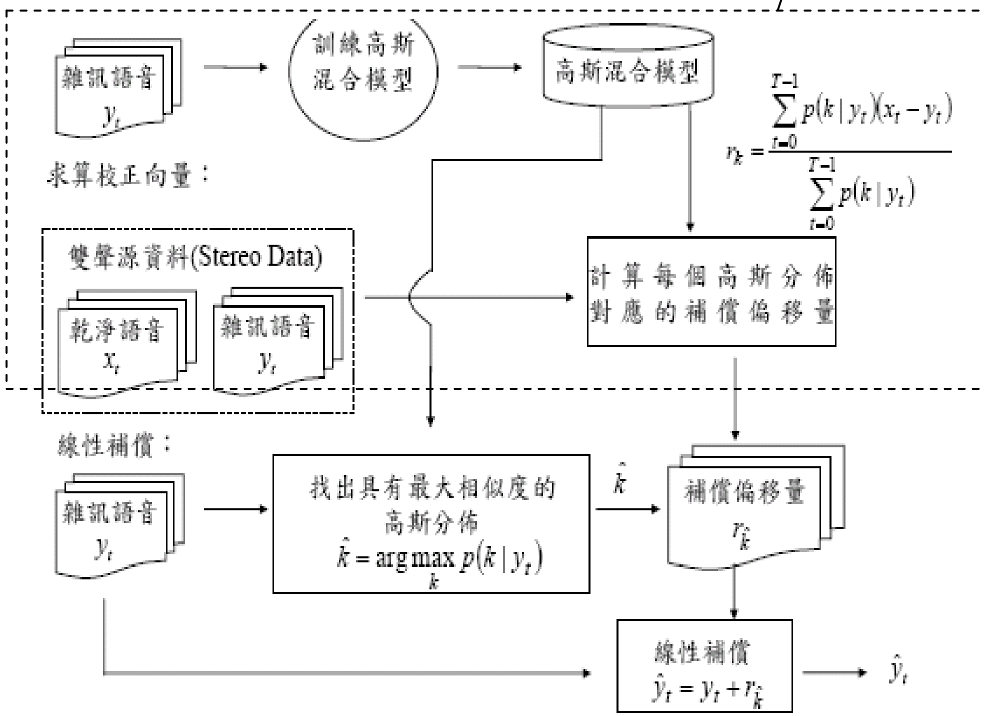
Probability of k are sum up to 1

$$\tilde{y}_t = y_t + \sum_k p(k \mid y_t) \cdot r_k$$

$$\hat{k} = \arg\max_{k} p(k \mid y_t)$$

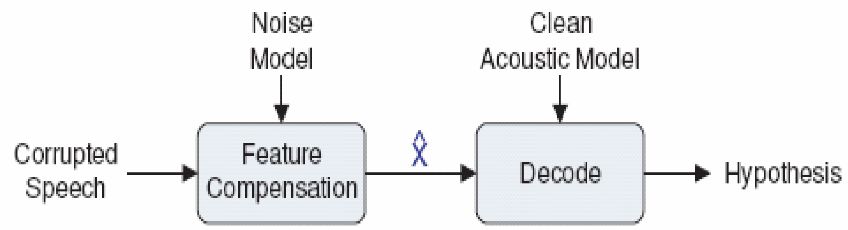
$$p(k \mid y_t) = \frac{p(y_t \mid k)\,p(k)}{\sum_{k'=0}^{k-1} p(y_t \mid k')\,p(k')}$$

During training time



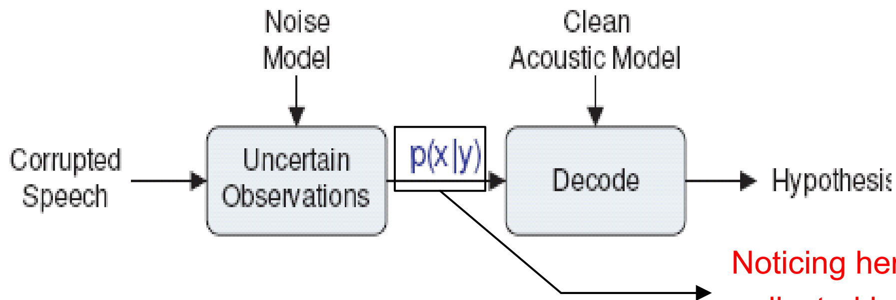
What is uncertainty?

- Feature compensation without uncertainty
  - The corrupted speech is restored by compensation and sent into decoder. The $\hat{x}$ is viewed as the clean feature, is that right?



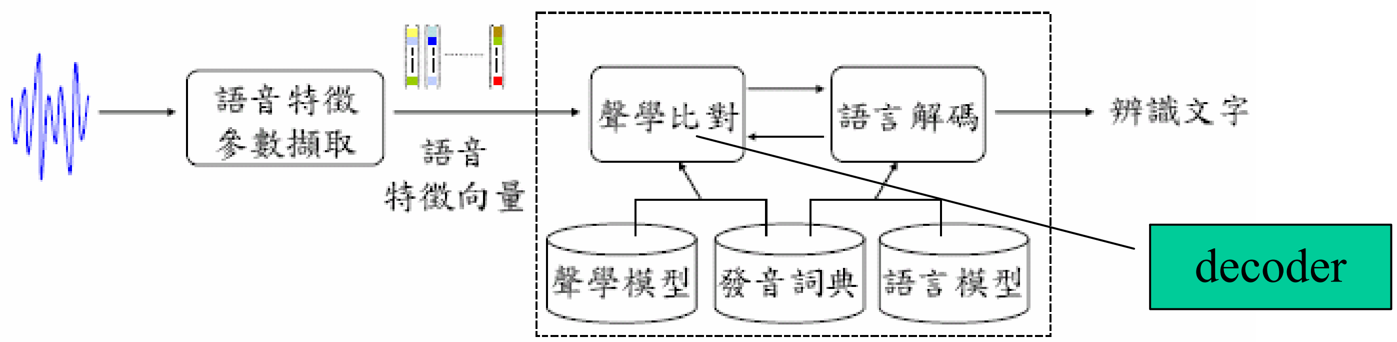
- Feature compensation with uncertainty

  It is intuitively reasonable to incorporate with uncertain observation.

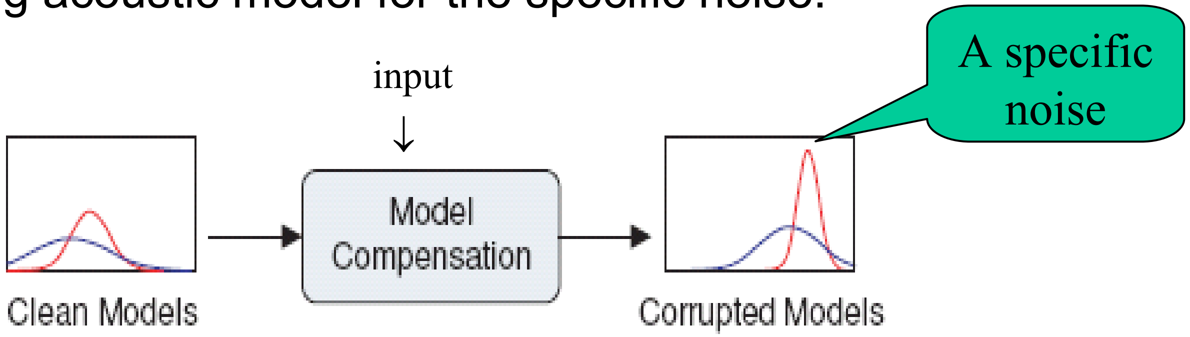


Noticing here, it is the key idea adjusted by SPLICE and JUN uncertainty decoding to make process efficient.

Concept of uncertainty decoding



- Model-compensation
  - Renewing acoustic model for the specific noise.

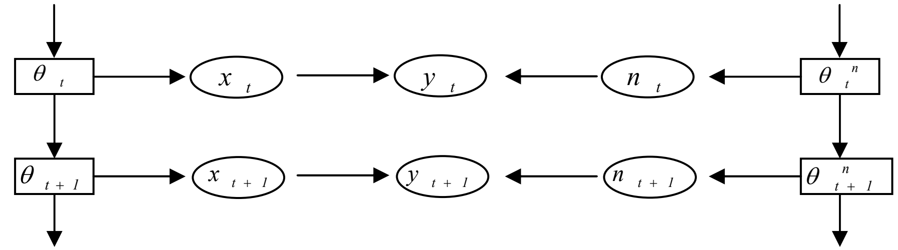


  - The input is either the corrupted speech data or the data combined clean and corrupted speech to achieve this goal.

- Computationally expensive

How to design uncertainty?

- Noise robustness DBN



- Corrupted speech likelihood given by

The key point

$$p(y_t|M,\breve{M},\theta_t) = \int \boxed{p(y_t|x_t,\breve{M})}\, p(x_t|M,\theta_t)dx_t \qquad (1)$$

$$\boxed{p(y_t|x_t,\breve{M})} = \int p(y_t|x_t,n_t)\, p(n_t|\breve{M},\theta_t^n)dn_t \qquad (2)$$

  – Only $p(y_t \mid x_t, \breve{M})$ depend on noise.

- Efficient approximation emerges from above formulation.
  – Independent of clean model complexity.
  – Appropriate form for integration.

Appendix A for (1)

marginalise

$$p(A \mid D) = \int p(A, B \mid D) d_B$$

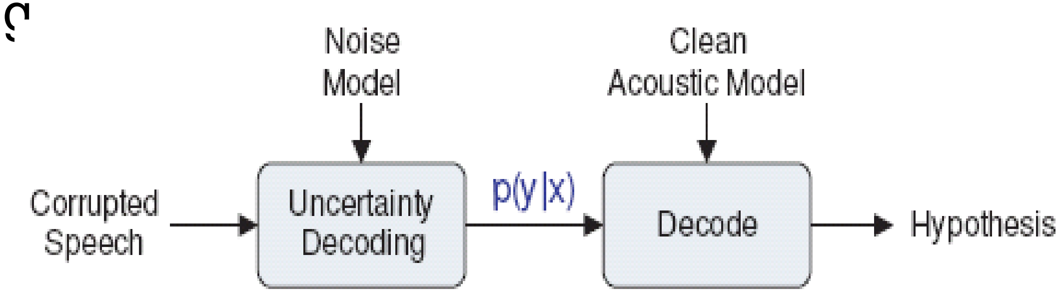
$$\begin{aligned} p(y_t \mid M, \breve{M}, \theta_t) &= \int p(y_t, x_t \mid M, \breve{M}, \theta_t) d_{x_t} \\ &= \int p(y_t | x_t, M, \breve{M}, \theta_t) p(x_t | M, \breve{M}, \theta_t) d_{x_t} \\ &= \int p(y_t \mid x_t, \breve{M}) p(x_t \mid M, \theta_t) d_{x_t} \end{aligned}$$

Appendix B for (2)

$$
\begin{aligned}
P(y_t \mid x_t, \breve{M}) &= \int p(y_t, n_t \mid x_t, \breve{M}) d_{n_t} \\
&= \int p(y_t \mid n_t, x_t, \breve{M}) \, p(n_t \mid x_t, \breve{M}) d_{n_t} \\
&= \int p(y_t \mid n_t, x_t) \, p(n_t \mid \breve{M}, \theta_t^n) d_{n_t}
\end{aligned}
$$

What's difference of decoding between SPLICE & JUD

- Passing conditional probability to decoding



- Passing conditional probability to decoding
- Tow form of uncertainty decoding
  - Splice with uncertainty $\longrightarrow$ $p(y_t|x_t,\breve{M})$ by Bayes' rule
  - Joint ditribution $\longrightarrow$ $p(y_t|x_t,\breve{M})$ by joint probability
- Both are based on Gaussian mixture model
  - Using different approximation to make process efficient

Uncertainty decoding with SPLICE

- Splice with uncertainty decoding uses Bayes' rule to write GMM as

$$p(y_t \mid x_t, \breve{M}) = \sum_{n=1}^{N} \left(\frac{p(x_t \mid y_t, \breve{s}_n, \breve{M})\, p(y_t \mid \breve{s}_n, \breve{M})\, \breve{c}_n}{p(x_t \mid \breve{M})} \right) \qquad (3)$$

  - $p(x_t \mid y_t, \breve{s}_n, \breve{M})$ related to standard SPLICE estimate
  - Denominator $p(x_t \mid \breve{M})$ is a GMM – simplify using a single Gaussian

Appendix C for (3)

$$
\begin{aligned}
p(y_t \mid x_t, \breve{M}) &= \sum_{n=1}^{N} p(y_t \mid \breve{s}_n, x_t, \breve{M})\, p(\breve{s}_n \mid x_t, \breve{M}) \\
&= \sum_{n=1}^{N} \frac{p(x_t, y_t \mid \breve{s}_n, \breve{M})\, p(\breve{s}_n \mid x_t, \breve{M})}{p(x_t \mid \breve{s}_n, \breve{M})} \\
&= \sum_{n=1}^{N} \frac{p(x_t \mid \breve{s}_n, y_t, \breve{M})\, p(y_t \mid \breve{s}_n, \breve{M})\, p(\breve{s}_n \mid x_t, \breve{M})}{p(x_t \mid \breve{s}_n, \breve{M})} \\
&= \frac{\sum_{n=1}^{N} p(x_t \mid \breve{s}_n, y_t, \breve{M})\, p(y_t \mid \breve{s}_n, \breve{M})\, p(\breve{s}_n \mid x_t, \breve{M})}{p(x_t \mid \breve{M})} \\
&= \frac{\sum_{n=1}^{N} p(x_t \mid \breve{s}_n, y_t, \breve{M})\, p(y_t \mid \breve{s}_n, \breve{M})\, \breve{c}_n}{p(x_t \mid \breve{M})}
\end{aligned}
$$

Total probability

Bayes' rule

Using single Gaussian

Replaced with prior

Uncertainty with SPLICE

- Standard SPLICE uses

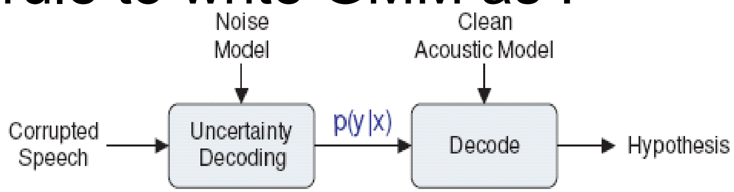
$$\hat{x}_t = \mathrm{E}[x_t \mid y_t] = \sum_k P(k \mid y_t)\mathrm{E}[x_t \mid y_t, k]$$

$$= \sum_{n=1}^{N} P(\breve{s}_n \mid y_t, \breve{M}) \int_x x_t \boxed{P(x_t \mid y_t, \breve{s}_n, \breve{M})} d_{x_t}$$

Replace k with s

- Uncertainty with SPLICE uses Bayes' rule to write GMM as :

$$P(y_t \mid x_t, \breve{M}) = \sum_{n=1}^{N} \left(\frac{\boxed{P(x_t \mid y_t, \breve{s}_n, \breve{M})} p(y_t \mid \breve{s}_n, \breve{M}) \breve{c}_n}{p(x_t \mid \breve{M})} \right)$$



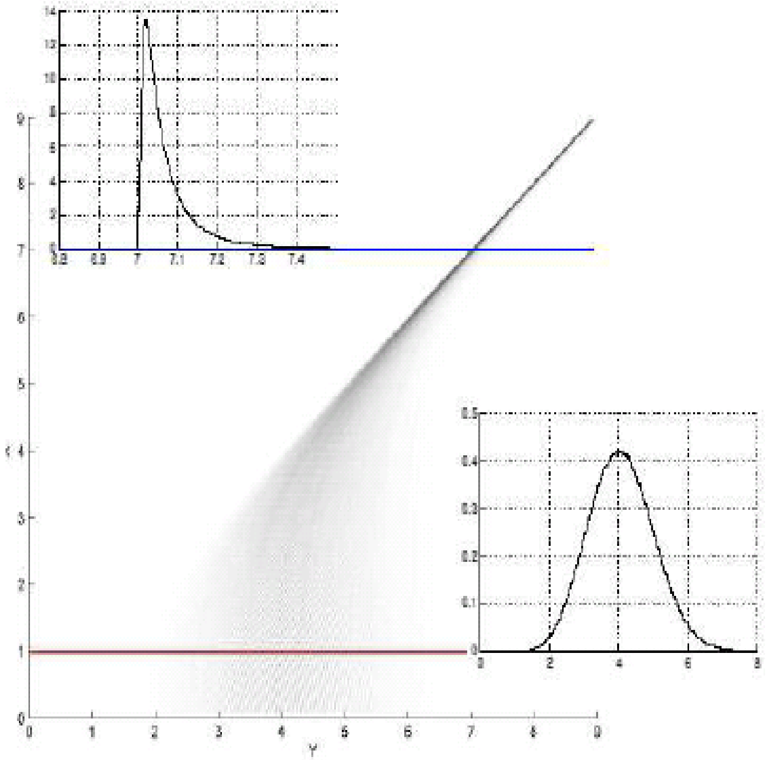
$$\mathrm{p}(\mathrm{y}_t \mid \mathrm{x}_t, \breve{\mathrm{M}}, \breve{\mathrm{s}}_n) = \mathrm{f}(\mathrm{y}_t, \breve{\mathrm{s}}_{n,})\mathrm{N}(\mathrm{A}^{(n)}\mathrm{y}_t + \mathrm{b}^{(n)}; \mathrm{x}_t, \textstyle\sum_b^{(n)})$$

$$\breve{\mathrm{s}}_{n^*} = \arg\max_{\breve{\mathrm{s}}_n} \left(\frac{\breve{\mathrm{c}}_n \mathrm{p}(\mathrm{y}_t \mid \breve{\mathrm{s}}_n, \breve{\mathrm{M}})}{\sum_{i=1}^{N} \breve{\mathrm{c}}_n \mathrm{p}(\mathrm{y}_t \mid \breve{\mathrm{s}}_i, \breve{\mathrm{M}})} \right)$$

$$\mathrm{p}(\mathrm{y}_t \mid \mathrm{M}, \breve{\mathrm{M}}, \theta_t) \propto \sum_{m \in \theta_t} \mathrm{c}_m \mathrm{N}(\mathrm{A}^{(n^*)}\mathrm{y}_t + \mathrm{B}^{(n^*)}; \mu^{(m)}, \textstyle\sum^{(m)} + \sum_b^{(n^*)})$$

Uncertainty decoding with JUD

- Joint distribution $p(x, y)$



When SNR high, the conditional is deterministic.

When SNR low, the conditional is Gaussian

Uncertainty decoding with JUN

- GMM is a standard approach to handle complex distribution
  - It's simple to marginalise tow Gaussians
- Using approximation front-end compensation model $\breve{M}$

$$p(y_t \mid x_t, \breve{M}) \approx \sum_{n=1}^{N} p(\breve{s}_n \mid x_t, \breve{M}) p(y_t \mid x_t, \breve{s}_n, \breve{M})$$

- Only $\breve{M}$ is a function of noise.

- Some issues need to be handled with

  - Component posterior $p(\breve{s}_n \mid x_t, \breve{M})$ is a function of clean speech
  - Component compensation parameters $p(y_t \mid x_t, \breve{s}_n, \breve{M})$
  - Direct use increases number of components

Uncertainty decoding for JUD

- Joint uncertainty decoding uses the GMM directly,

$$p(y_t \mid x_t, \breve{M}) = \sum_{n=1}^{N} p(y_t \mid \breve{s}_n, x_t, \breve{M})\, p(\breve{s}_n \mid x_t, \breve{M})$$

$$= \sum_{n=1}^{N} \frac{p(x_t, y_t \mid \breve{s}_n, \breve{M})\, p(\breve{s}_n \mid x_t, \breve{M})}{p(x_t \mid \breve{s}_n, \breve{M})}$$

but

- Approximates the component posterior of clean speech, using the corrupted speech:

$$p(\breve{s}_n \mid x_t, \breve{M}) \approx p(\breve{s}_n \mid y_t, \breve{M})$$

- This decouples the front-end distribution from being dependent on the acoustic model through the clean speech variable
- conditional probability derived from the joint distribution

$$p(x_t, y_t \mid \breve{s}_n, \breve{M}) = N\left(\begin{bmatrix} x_t \\ y_t \end{bmatrix}; \begin{bmatrix} \breve{\mu}_x^{(n)} \\ \breve{\mu}_y^{(n)} \end{bmatrix}, \begin{bmatrix} \Sigma_{xx}^{(n)} & \Sigma_{xy}^{(n)} \\ \Sigma_{xy}^{(n)} & \Sigma_{yy}^{(n)} \end{bmatrix}\right)$$

covariance matrix is usually made diagonal ofr efficiency

Uncertainty decoding for JUD

- Both uncertainty decoding schemes yield same decoding form:

$$p(y_y \mid M, \breve{M}, \theta_t) \approx \sum_{m=1}^{M} \sum_{n=1}^{N} \alpha^{(mn)} \mathrm{N}(\mathbf{A}^{(n)} y_t + \mathbf{b}^{(n)}; \mu^{(m)}, \Sigma^{(m)} + \Sigma^{(n)})$$

  – Form of $\mathbf{A}^{(n)}, \mathbf{b}^{(n)}$ and $\Sigma^{(n)}$ differ in the two cases

  – For JUN that would be

$$\mathrm{A}^{(\mathrm{n})} = \sum\nolimits_{\mathrm{x}}^{(\mathrm{n})} \sum\nolimits_{\mathrm{yx}}^{(\mathrm{n})\text{-}1}$$

$$\mathrm{b}^{(\mathrm{n})} = \mu_{\mathrm{x}}^{(\mathrm{n})} - \mathrm{A}^{(\mathrm{n})} \mu_{\mathrm{y}}^{(\mathrm{n})}$$

$$\sum\nolimits_{\mathrm{b}}^{(\mathrm{n})} = \mathrm{A}^{(\mathrm{n})} \sum\nolimits_{\mathrm{y}}^{(\mathrm{n})} \mathrm{A}^{(\mathrm{n})\mathrm{T}} - \sum\nolimits_{\mathrm{x}}^{(\mathrm{n})}$$

- To improve efficiency only a single front-end component selected, for Joint based on $p(\breve{s}_n \mid y_t, \breve{M})$

- Compared to model-based compensation computational cost is:
  – only a function of the N,
  – Not the number of components in clean speech model through variance bias must be applied