Robust Speech Recognition using Dynamic Noise Adaptation

Steven Rennie, Pierre Dognin, and Petr Fousek

IBM Thomas J. Watson Research Center, NY, USA

May 26, 2011

Outline

Model-based Robust ASR

Model-based Robust ASR

A well established paradigm:

- Explicit models of noise, channel distortion, and their interaction with speech.
- Many interesting modeling/inference techniques , interaction models (refs., see paper)

Model-based Robust ASR

A well established paradigm:

- Explicit models of noise, channel distortion, and their interaction with speech.
- Many interesting modeling/inference techniques , interaction models (refs., see paper)

Relevance to commercial-grade ASR not definitively established:

- Promising WERRs on less sophisticated ASR systems (small training/test sets, simple pipelines, artificially mixed data).
- Little evidence that these techniques can improve truly state-of-the-art ASR systems.

Dynamic Noise Adaptation (DNA)

Dynamic Noise Adaptation (DNA)

Method:

- Model-based approach: GMM for speech, dynamically evolving model of noise (gaussian process).
- Features: mis-match model–no noise model training or system re-training req'd; models uncertainty in noise estimate.

Dynamic Noise Adaptation (DNA)

Method:

- Model-based approach: GMM for speech, dynamically evolving model of noise (gaussian process).
- Features: mis-match model–no noise model training or system re-training req'd; models uncertainty in noise estimate.

Previous Results:

- Significantly outperforms well-known techniques like the ETSI AFE, and fMLLR on the Aurora II and DNA+Aurora II tasks.

Dynamic Noise Adaptation (DNA)

Method:

- Model-based approach: GMM for speech, dynamically evolving model of noise (gaussian process).
- Features: mis-match model–no noise model training or system re-training req'd; models uncertainty in noise estimate.

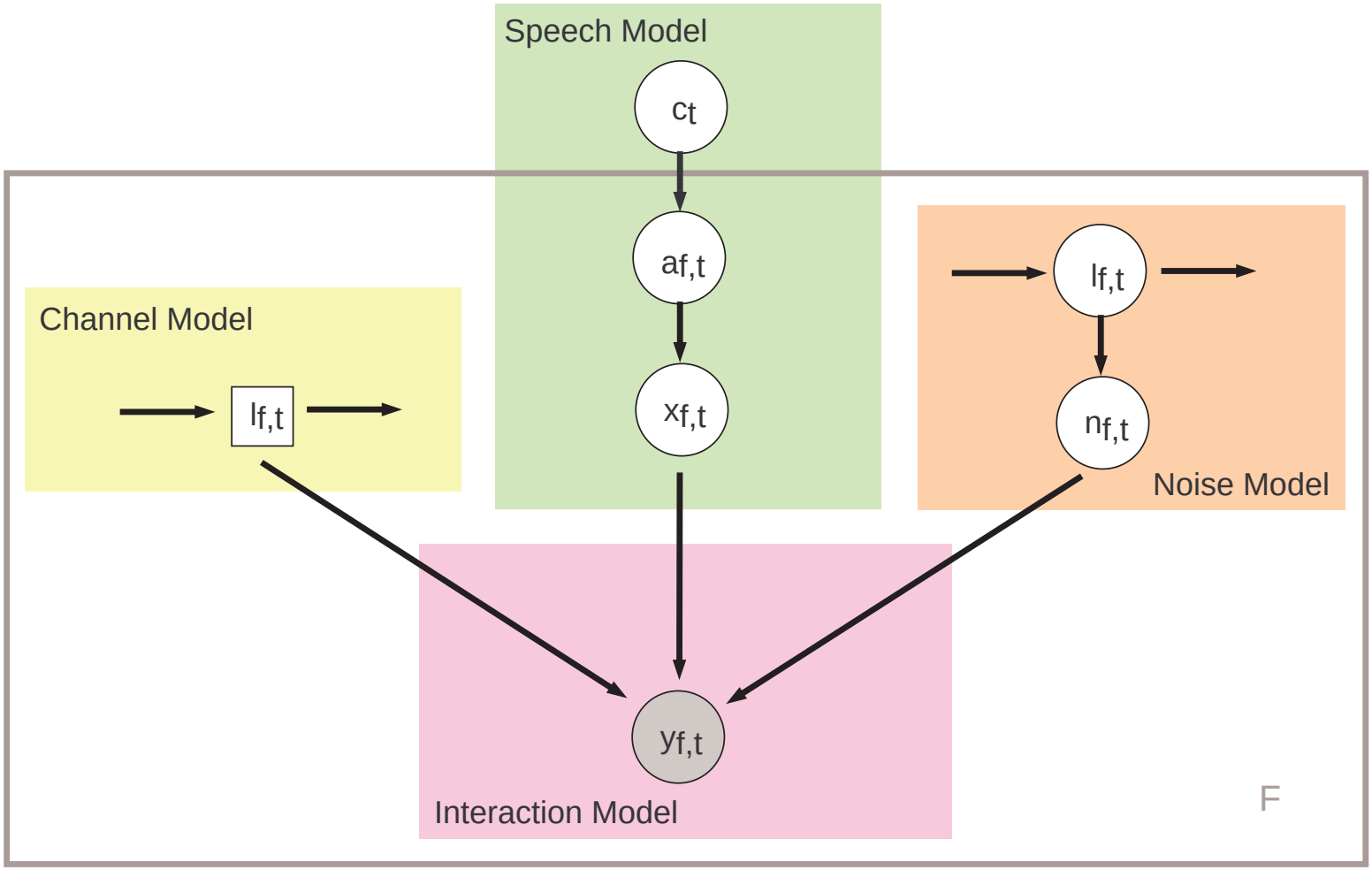
Previous Results:

- Significantly outperforms well-known techniques like the ETSI AFE, and fMLLR on the Aurora II and DNA+Aurora II tasks.

This paper:

- New results: real data, commercial-grade ASR systems.
- Maturation: for low-latency deployments.
- Gains: 22% WERR below 6 dB SNR (fMPE+SS+fMLLR).

DNA: Generative Model



DNA: Interaction Model

DNA: Interaction Model

Time Domain:

$$y(t) = h(t) * x(t) + n(t), \tag{1}$$

DNA: Interaction Model

Time Domain:

$$y(t) = h(t) * x(t) + n(t), \tag{1}$$

Frequency Domain:

$$\begin{aligned} |Y|^2 &= |H|^2|X|^2 + |N|^2 + 2|H||X||N|\cos\theta \\ &= |H|^2|X|^2 + |N|^2 + \epsilon, \end{aligned} \tag{2}$$

DNA: Interaction Model

Time Domain:

$$y(t) = h(t) * x(t) + n(t), \tag{1}$$

Frequency Domain:

$$\begin{aligned} |Y|^2 &= |H|^2|X|^2 + |N|^2 + 2|H||X||N|\cos\theta \\ &= |H|^2|X|^2 + |N|^2 + \epsilon, \end{aligned} \tag{2}$$

Log Mel Domain:

$$y \approx \log(\exp(x+h) + \exp(n)) = f(x+h, n), \tag{3}$$

$$p(y|x+h, n) = \mathcal{N}(y; f(x+h, n), \psi^2). \tag{4}$$

DNA: Speech Model

DNA: Speech Model

A band-quantized GMM:

$$p(s) = \pi_s, \;\; p(x|s) = \prod_f \mathcal{N}\left(x; \mu_{a(s,f)}, \sigma^2_{a(s,f)}\right), \tag{5}$$

- $a(s, f)$ maps acoustic state s to Gaussian a in frequency band f.
- $|a(s, f)| \ll |s|$, so model can be efficiently computed/stored.

DNA: Noise Model

DNA: Noise Model

Noise Level:

$$p(l_{f,0}) = \mathcal{N}\left(l_{f,0}; \beta_f, \omega^2_{f,0}\right), \quad (6)$$

$$p(l_{f,\tau}|l_{f,\tau-1}) = \mathcal{N}\left(l_{f,\tau}; l_{f,\tau-1}, \gamma^2_f\right), \quad (7)$$

DNA: Noise Model

Noise Level:

$$p(l_{f,0}) = \mathcal{N}\left(l_{f,0}; \beta_f, \omega^2_{f,0}\right), \quad (6)$$

$$p(l_{f,\tau}|l_{f,\tau-1}) = \mathcal{N}\left(l_{f,\tau}; l_{f,\tau-1}, \gamma^2_f\right), \quad (7)$$

Transient Noise:

$$p(n_{f,\tau}|l_{f,\tau}) = \mathcal{N}\left(n_{f,\tau}; l_{f,\tau}, \phi^2_f\right). \quad (8)$$

- descriptive of diffuse, slowly evolving noise (rel. to frame rate)
- decomposition facilitates robust noise tracking

DNA: Channel Model

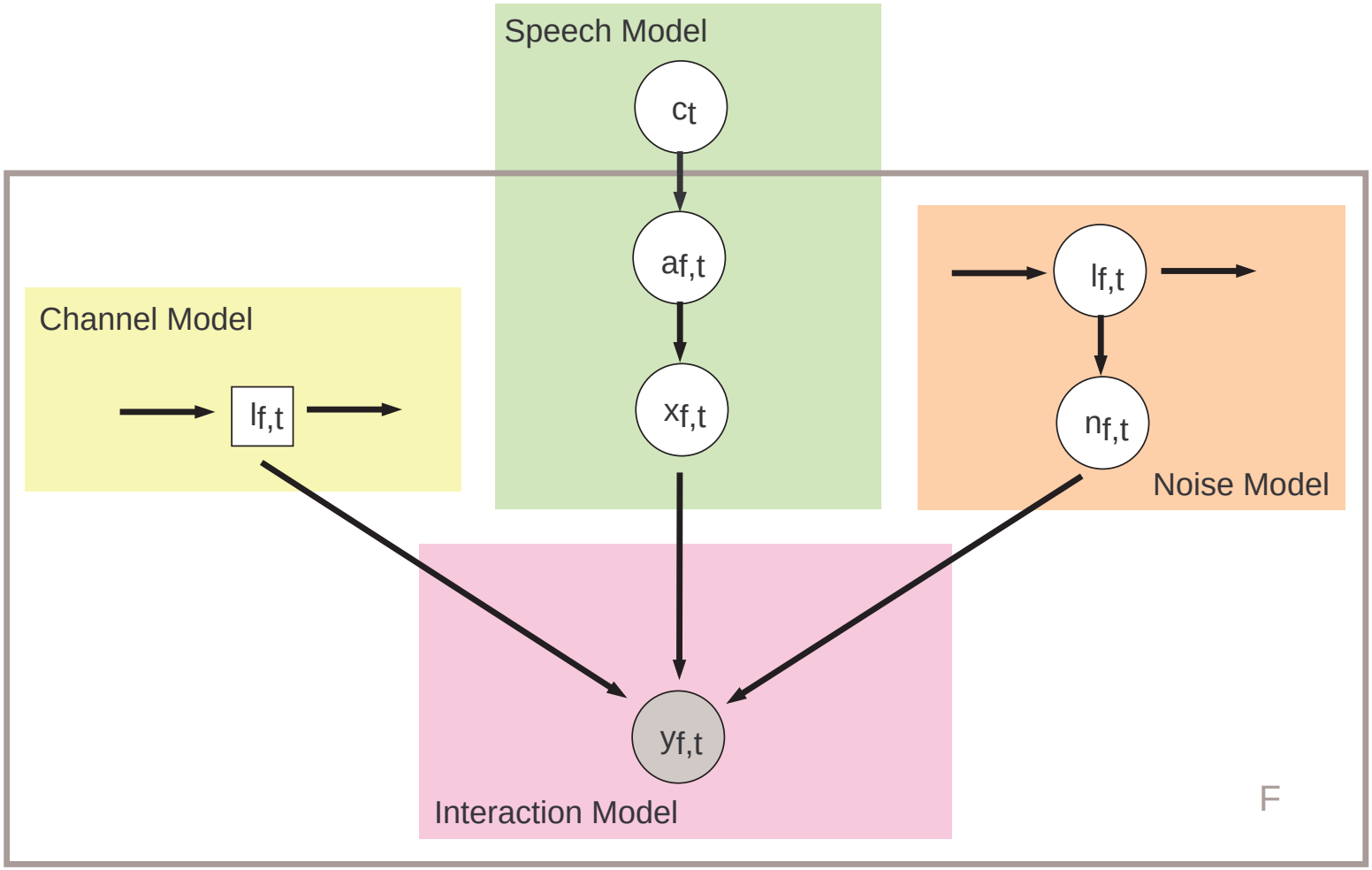
DNA: Channel Model

A stochastically adapted parameter vector:

$$p(h_{f,\tau}) = \delta(h_{f,\tau} - \hat{h}_f(\tau)), \tag{9}$$

- $\hat{h}_f(\tau)$: current estimate in frequency bin f at frame τ.
- $\hat{h}_f(\tau)$ is stochastically adapted (see Deng et al., 2003).

DNA: Generative Model



DNA: Inference

DNA: Inference

Noise Model Update:

- Exact noise posterior at time t has $|s|^T$ components.
- Approximate it as Gaussian:

$$p(l_{f,\tau+1}) \approx \mathcal{N}(l_{f,\tau+1}; \beta_{f,\tau+1}, \omega^2_{f,\tau+1}), \tag{10}$$

DNA: Inference

Noise Model Update:

- Exact noise posterior at time t has $|s|^T$ components.
- Approximate it as Gaussian:

$$p(l_{f,\tau+1}) \approx \mathcal{N}(l_{f,\tau+1}; \beta_{f,\tau+1}, \omega^2_{f,\tau+1}), \tag{10}$$

$$\beta_{f,\tau+1} = \mathrm{E}[l_t|\mathbf{y}_{0:\tau}] = \sum_{s_\tau} p(s_\tau|\mathbf{y}_{0:\tau})\mathrm{E}[l_{f,\tau}|\mathbf{y}_{0:\tau}, s_\tau]\,, \tag{11}$$

$$\begin{aligned}\omega^2_{f,\tau+1} &= \mathrm{Var}[l_{f,\tau}|\mathbf{y}_{0:\tau}] + \gamma^2_f \\ &= \sum_{s_\tau} p(s^x_\tau|\mathbf{y}_{0:\tau})\{\mathrm{Var}[l_{f,\tau}|\mathbf{y}_{0:\tau}, s_\tau] + \\ &\quad (\mathrm{E}[l_{f,\tau}|\mathbf{y}_{0:\tau}] - \mathrm{E}[l_{f,\tau}|\mathbf{y}_{0:\tau}, s_\tau])^2\} + \gamma^2_f.\end{aligned} \tag{12}$$

DNA: Inference (cont.'d)

DNA: Inference (cont.'d)

Likelihood approximation (for a given time and frequency):

- Variant of Algonquin (variant of iterative VTS).
- Iteratively linearize interaction for each Gaussian atom a:

$$p(y|x,n,h) \approx \mathcal{N}(y; \alpha_a(x+h) + (1-\alpha_a)n + b_a, \psi^2), \quad (13)$$

DNA: Inference (cont.'d)

Likelihood approximation (for a given time and frequency):

- Variant of Algonquin (variant of iterative VTS).
- Iteratively linearize interaction for each Gaussian atom a:

$$p(y|x,n,h) \approx \mathcal{N}(y; \alpha_a(x+h) + (1-\alpha_a)n + b_a, \psi^2), \quad (13)$$

$$\alpha_a = \left.\frac{\delta f}{\delta x}\right|_{\hat{x}_a,\hat{l}_a,\hat{n}_a} = \frac{|\hat{\mathrm{H}}_a|^2|\hat{\mathrm{X}}_a|^2}{|\hat{\mathrm{H}}_a|^2|\hat{\mathrm{X}}_a|^2 + |\hat{\mathrm{N}}_a|^2}, \quad (14)$$

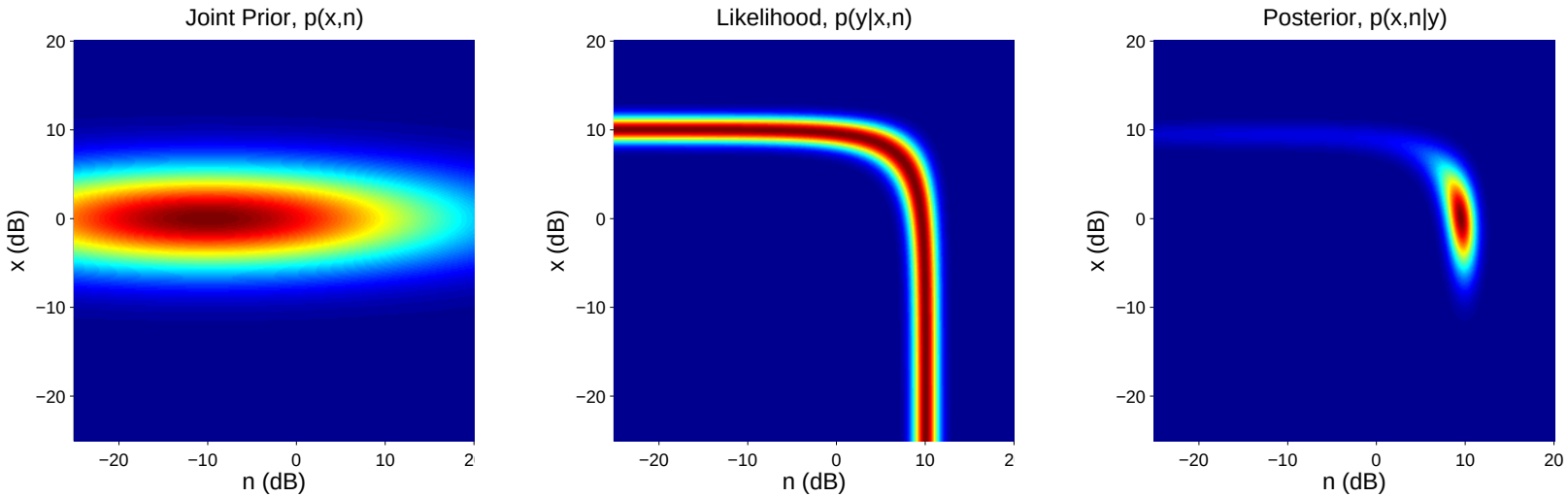
$$b_a = f(\hat{x}_a + \hat{h}_a, \hat{n}_a) - \alpha_a(\hat{x}_a + \hat{h}_a - \hat{n}_a) - \hat{n}_a. \quad (15)$$

- Update speech, noise-level posterior given $p(y|x,n,h)$.

DNA: Inference (cont.'d)

Likelihood approximation (for a given time and frequency):

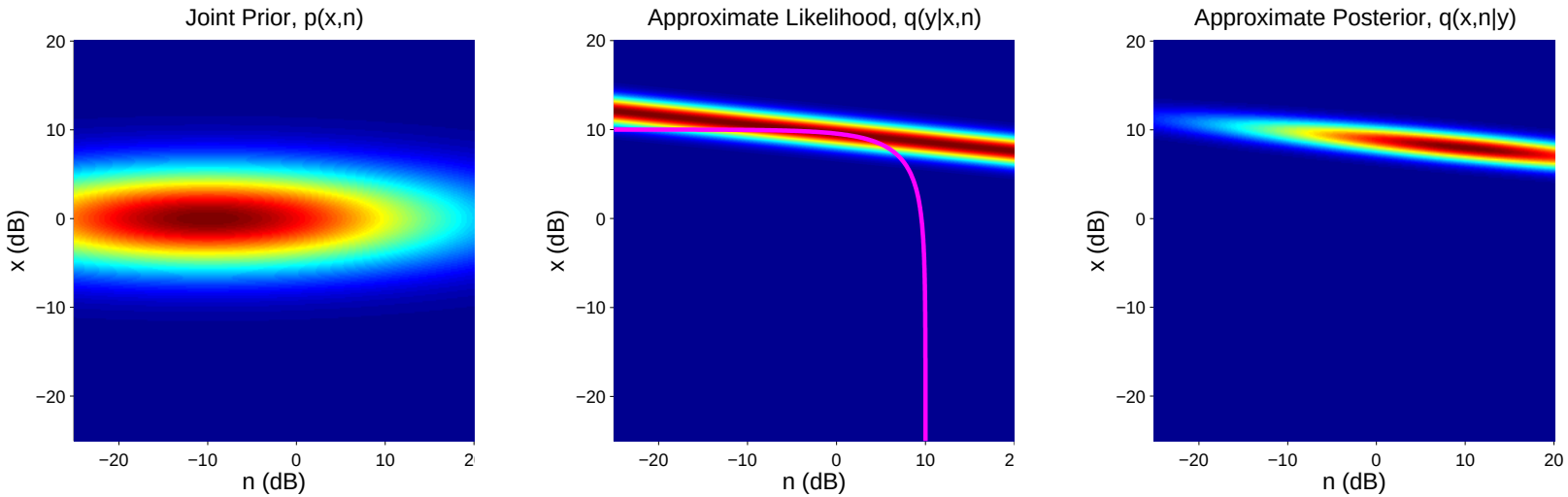
- True posterior (under log-sum model)



DNA: Inference (cont.'d)

Likelihood approximation (for a given time and frequency):

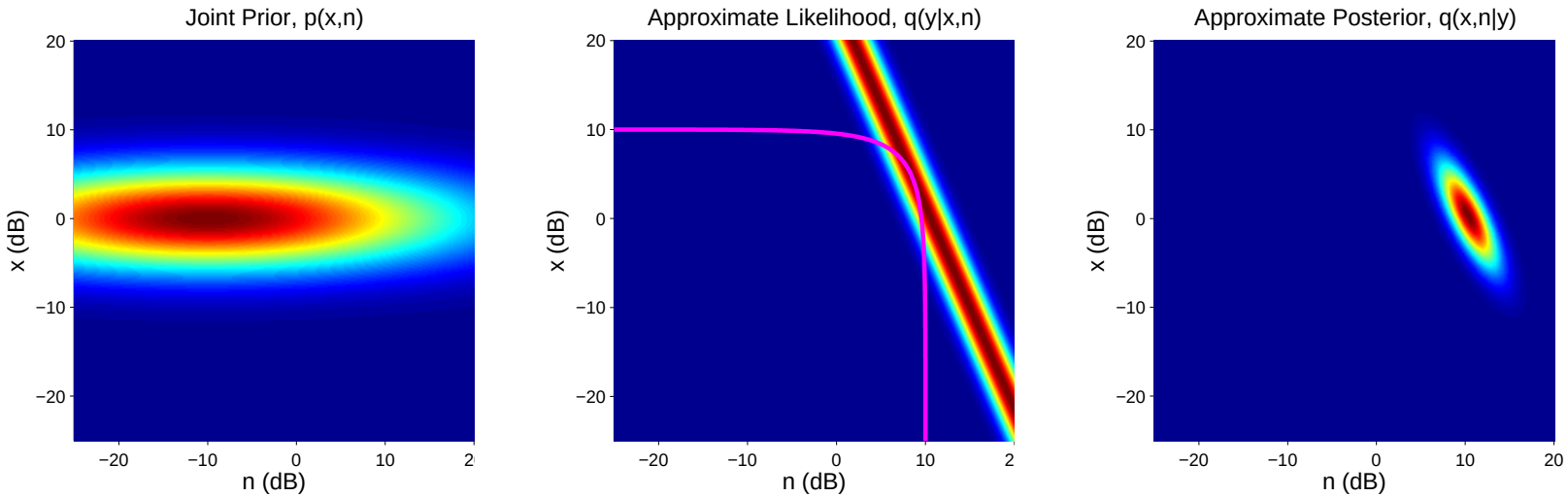
- iteration 0



DNA: Inference (cont.'d)

Likelihood approximation (for a given time and frequency):

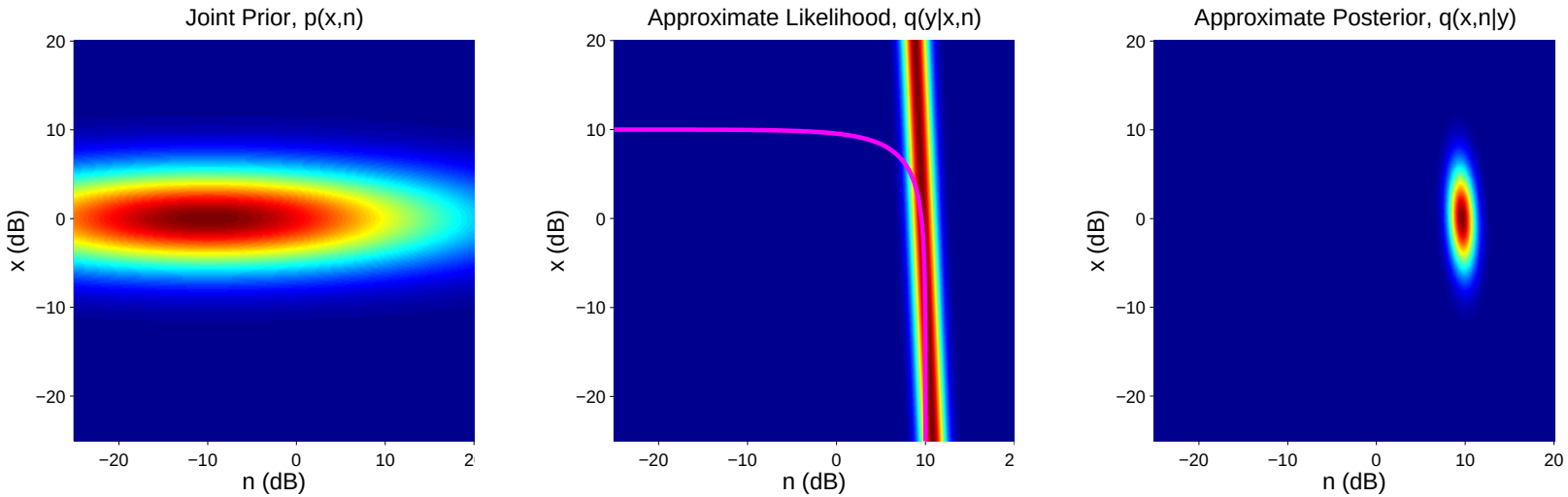
- iteration 1



DNA: Inference (cont.'d)

Likelihood approximation (for a given time and frequency):

- iteration 2



DNA: Inference (cont.'d)

DNA: Inference (cont.'d)

Speech reconstruction:

$$\hat{x}_{f,\tau} = \mathrm{E}[x_{f,\tau}|\mathbf{y}_{0:\tau}] = \sum_{s_\tau} p(s_\tau|\mathbf{y}_{0:\tau})\mathrm{E}[x_{f,\tau}|\mathbf{y}_{0:\tau}, s_\tau]\,. \qquad (16)$$

Data

Data

- US English in-car speech recorded in various noise conditions (0, 30 and 60 mph), and sampled at 16kHz.

Data

- US English in-car speech recorded in various noise conditions (0, 30 and 60 mph), and sampled at 16kHz.
- Training: 803K utterances, 10.3K speakers (786 hours).

Data

- US English in-car speech recorded in various noise conditions (0, 30 and 60 mph), and sampled at 16kHz.
- Training: 803K utterances, 10.3K speakers (786 hours).
- Test: 38.9K utterances (206K words), 128 held-out speakers.

Data

- US English in-car speech recorded in various noise conditions (0, 30 and 60 mph), and sampled at 16kHz.
- Training: 803K utterances, 10.3K speakers (786 hours).
- Test: 38.9K utterances (206K words), 128 held-out speakers.
- 47 tasks covering four domains (navigation, cmd/ctrl, digits/dialing, radio), 7 regional US accents.

Models

Models

- Back-end acoustic models:
  - word-internal, ± 2 phonetic context, 865 context-dependent states, 10K diag-cov GMM.
  - From 40-dim LDA features, built an ML model, a *clean* ML model (20 dB+), and an fMPE model.
  - Once trained, AMs were compressed using hierarchical band-quantization.

Models

- Back-end acoustic models:
  - word-internal, ± 2 phonetic context, 865 context-dependent states, 10K diag-cov GMM.
  - From 40-dim LDA features, built an ML model, a *clean* ML model (20 dB+), and an fMPE model.
  - Once trained, AMs were compressed using hierarchical band-quantization.
- DNA speech models:
  - From 23-dim log Mel features, built (all/clean) diag-cov GMMs with 256 speech, 16 silence components
  - Once trained, compressed to a BQ GMM (8 Gaussians/dim)

Adaptation Algorithms

Adaptation Algorithms

- Spectral subtraction (SS), DNA, CMN, fMLLR, fMPE

Adaptation Algorithms

- Spectral subtraction (SS), DNA, CMN, fMLLR, fMPE
- SS:
  - Model-based speech detector, noise estimated from speech-free frames, adapted with geometric forgetting ($\alpha = 0.9$)
  - $\hat{x} = min(y - \hat{n}, \beta\hat{n})$

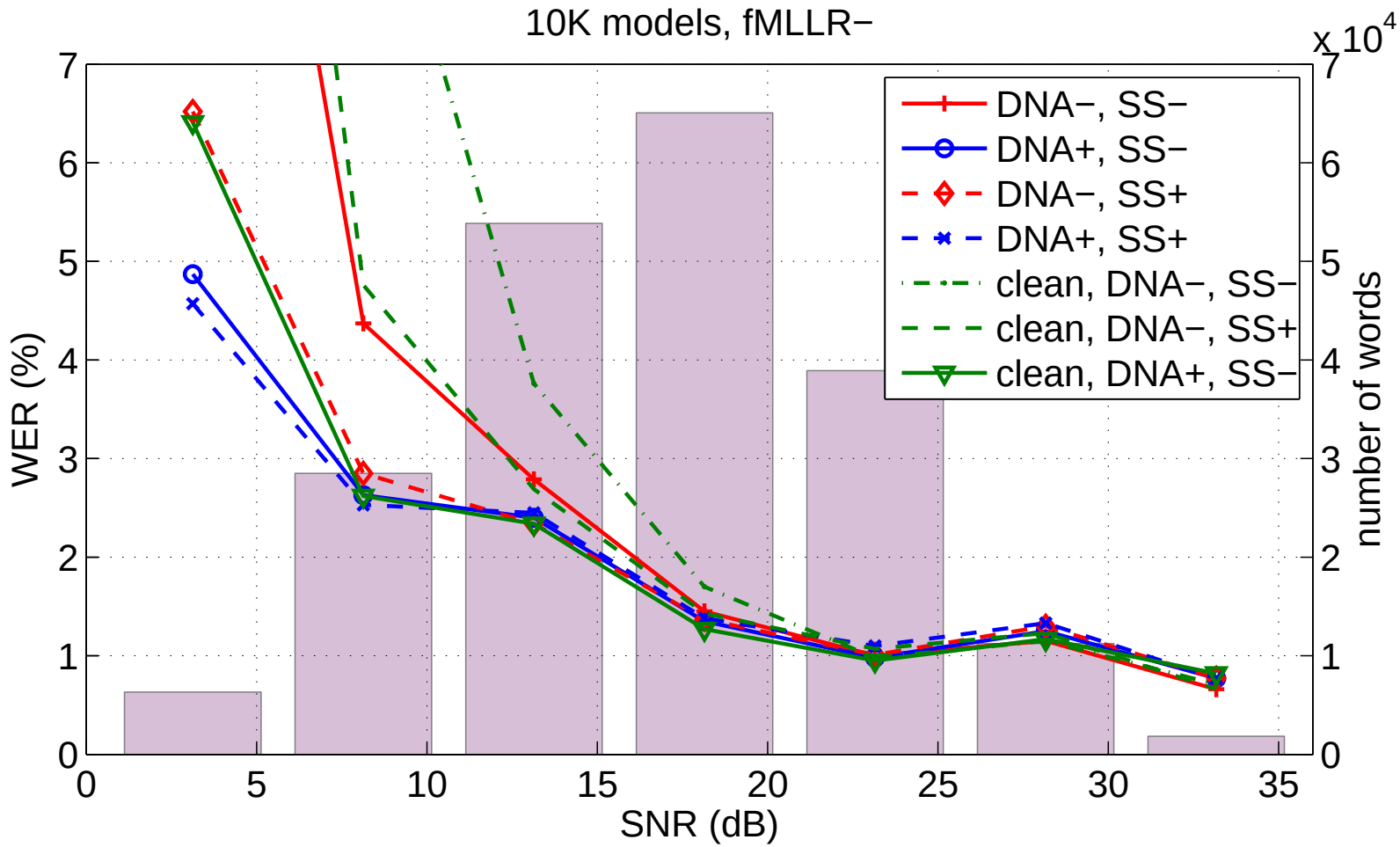
Adaptation Algorithms

- Spectral subtraction (SS), DNA, CMN, fMLLR, fMPE
- SS:
  - Model-based speech detector, noise estimated from speech-free frames, adapted with geometric forgetting ($\alpha = 0.9$)
  - $\hat{x} = min(y - \hat{n}, \beta\hat{n})$
- fMLLR:
  - online adaptation, every 5 frames, using stochastic gradient

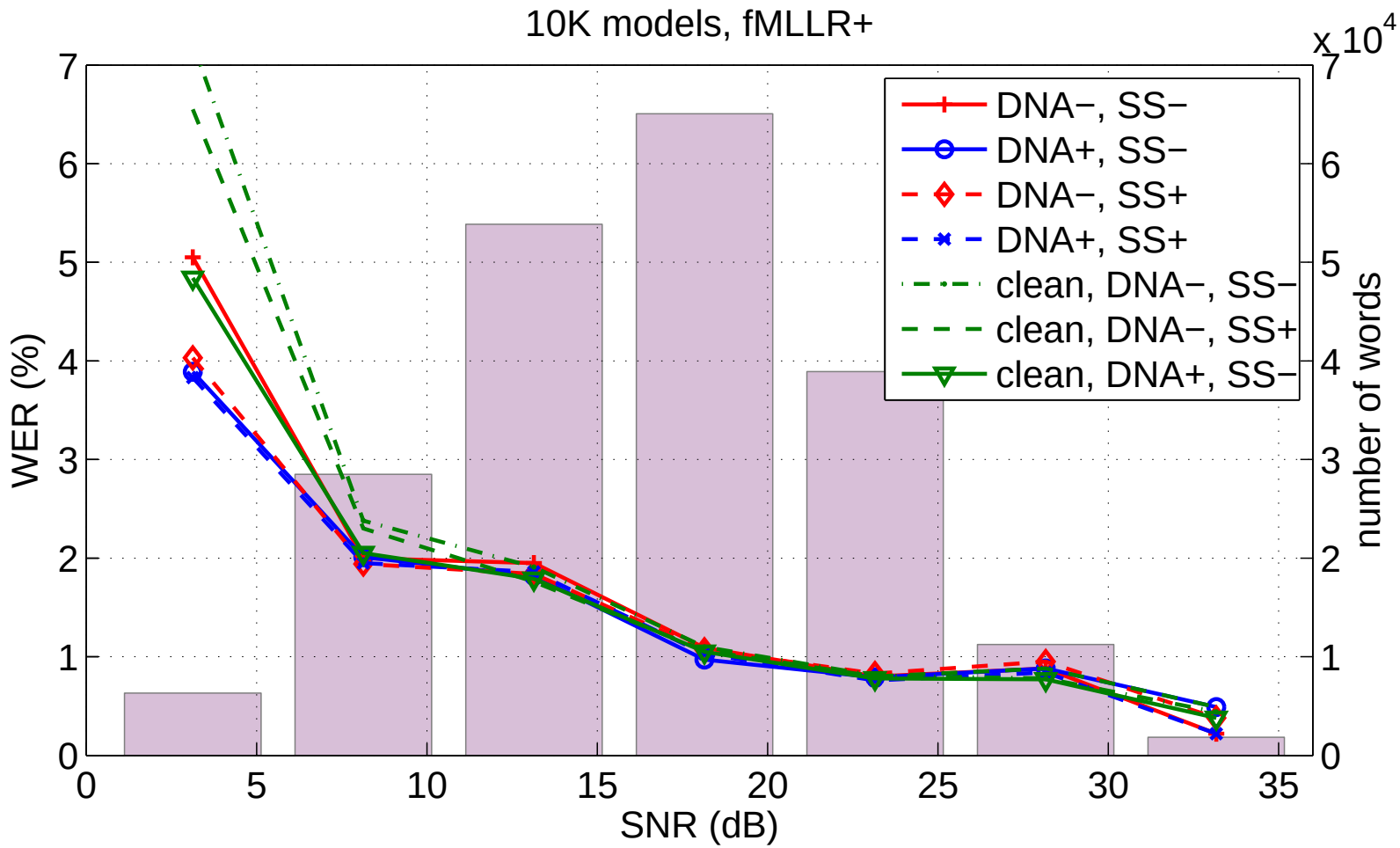
Adaptation Algorithms

- Spectral subtraction (SS), DNA, CMN, fMLLR, fMPE
- SS:
  - Model-based speech detector, noise estimated from speech-free frames, adapted with geometric forgetting ($\alpha = 0.9$)
  - $\hat{x} = min(y - \hat{n}, \beta\hat{n})$
- fMLLR:
  - online adaptation, every 5 frames, using stochastic gradient
- fMPE:
  - 512 Gaussians, 17 frame inner context, 9 frame outer context

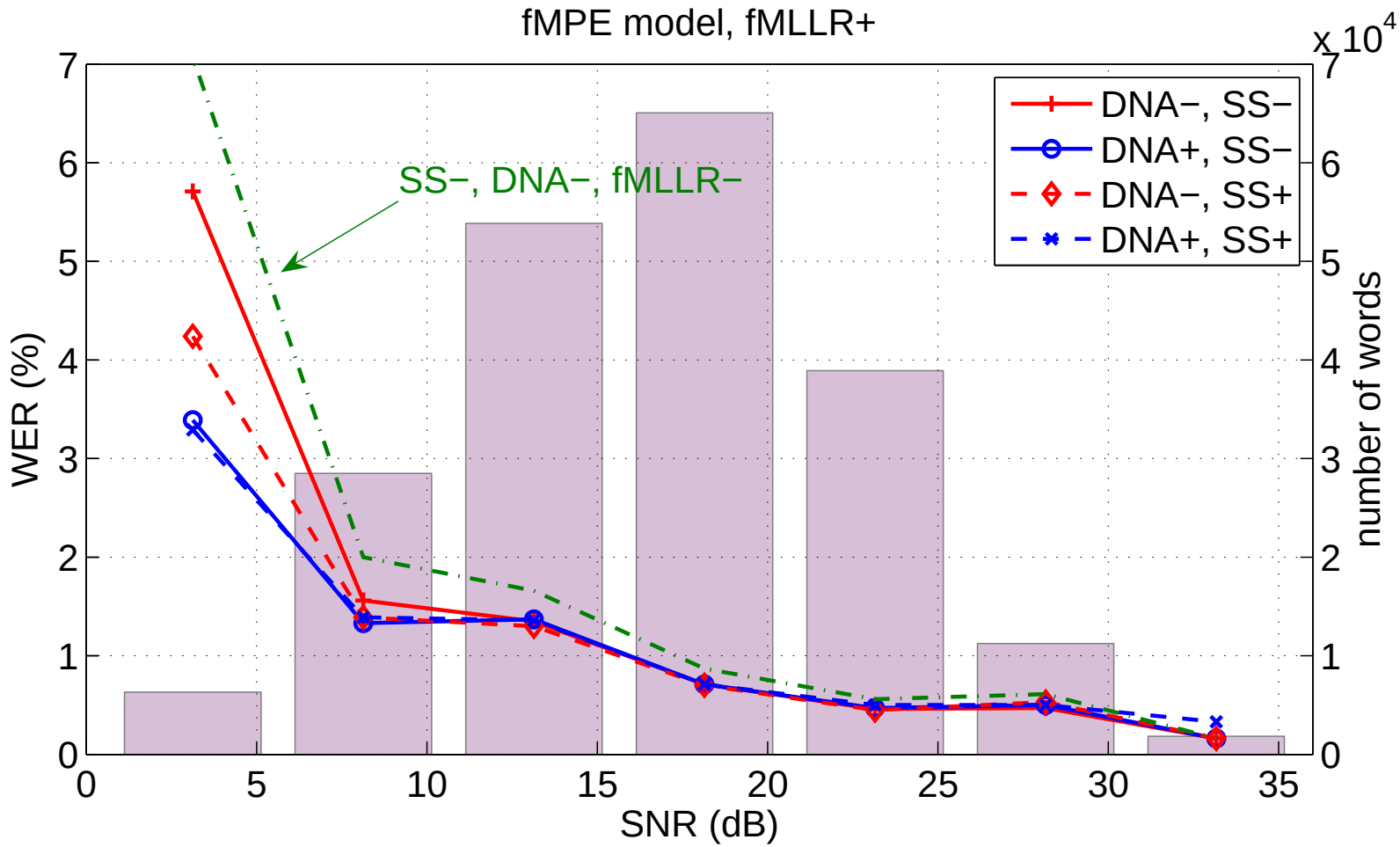
ML Results, fMLLR off



ML Results, fMLLR on



ML Results, fMPE on



Overall results

	WER (%) / SER (%)			
Model	SS- fMLLR-	SS+ fMLLR-	SS- fMLLR+	SS+ fMLLR+
10K reference				
no-DNA	2.49 / 6.15	1.91 / 5.12	1.49 / 3.99	1.43 / 3.86
DNA	1.83 / 5.11	1.86 / 5.21	1.38 / 3.91	1.39 / 3.79
10K clean				
no-DNA	4.10 / 8.60	2.66 / 6.28	1.60 / 4.38	1.52 / 4.21
DNA	1.82 / 5.16	1.88 / 5.17	1.42 / 4.06	1.41 / 3.95
$\text{DNA}^{20\text{dB}+}$	1.88 / 5.27	1.92 / 5.36	1.42 / 3.94	1.41 / 3.97
fMPE				
no-DNA	1.34 / 3.77	1.18 / 3.41	1.08 / 3.00	1.00 / 2.79
DNA	1.21 / 3.77	1.25 / 3.80	0.99 / 2.89	1.00 / 2.92

Summary and Future Work

- Summary: DNA works, noise adaptation is important.
- Future work: stronger models, inference algorithms, tighter back-end integration,...