Matched-Condition Robust Dynamic Noise Adaptation

Steven J. Rennie, Pierre L. Dognin, Petr Fousek
IBM T.J. Watson Research Center
Yorktown Heights, N.Y., U.S.A.
{sjrennie, pdognin}@us.ibm.com, petr_fousek@cz.ibm.com

***Abstract*—In this paper we describe how the model-based noise robustness algorithm for previously unseen noise conditions, Dynamic Noise Adaptation (DNA), can be made robust to matched data, without the need to do any system re-training. The approach is to do online model selection and averaging between two DNA models of noise: one that is tracking the evolving state of the background noise, and one clamped to the null mis-match hypothesis. The approach, which we call DNA with (matched) condition detection (DNA-CD), improves the performance of a commerical-grade speech recognizer that utilizes feature-space Maximum Mutual Information (fMMI), boosted MMI (bMMI), and feature-space Maximum Likelihood Linear Regression (fMLLR) compensation by 15% relative at signal-to-noise ratios (SNRs) below 10 dB, and over 8% relative overall.**

Index Terms: Dynamic Noise Adaptation (DNA), Noise Robustness, Model Adaptation, fMLLR, fMMI, Spectral Subtraction, Algonquin.

I. Introduction

Dynamic noise adaptation (DNA) [1], [2] is a model-based technique for improving automatic speech recognition (ASR) performance in noise [1], [2], [3]. DNA is designed to compensate for mis-match between training and testing conditions, and significantly outperforms the ETSI AFE on the Aurora II and DNA+Aurora II tasks. Recently, DNA has been shown to improve the performance of even commercial-grade ASR systems trained on large amounts of data [3]. However, new investigations with yet more data and yet stronger baseline systems have revealed that DNA can sometimes *harm* performance.

DNA as described in previous publications does not degrade performance in mis-match scenarios, but rather when the current noise conditions are well characterized by the acoustic model in the back-end (and the surrogate speech model in DNA). Such issues can be mitagated by applying the model-based approach to the recognizer itself, and training acoustic models of speech that recover a "canonical" representation of speech, together with a noise model, which could be adapted [4], [5], [6]. However, this paradigm is not yet fully mature (even simple dimensionality reduction techniques like LDA complicate matters), and these systems are not yet truly commerical grade.

In this paper we demonstrate that it is possible to automatically and reliably detect when mis-match noise modeling is not beneficial. Our approach is to do online Bayesian model averaging to regularize the influence that mis-match noise modeling has on the output speech feature estimate. Results demonstrate that the technique improves the Sentence Error Rate (SER) of a state-of-the-art embedded speech recognizer that utilizes commerical-grade feature-space Maximum Mutual Information (fMMI), boosted MMI (bMMI), and feature-space Maximum Likelihood Linear Regression (fMLLR) compensation by 15% relative at *signal-to-noise ratios* (SNRs) below 10 dB, and over 8% relative overall.

II. The DNA Model

The DNA model was recently presented in detail in [3]. Here we briefly review DNA so that we can adequately describe DNA with condition detection (DNA-CD).

The DNA model consists of a speech model, noise model, channel model, and *interaction model*, which describes how these acoustic entities combine to generate the observed speech.

A. Interaction Model

The interaction between speech x, noise n and channel effects h is modeled in time domain as

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

where $*$ denotes linear convolution. In the frequency domain we obtain

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

where $|\mathrm{X}|$ and θ_{x} represent the magnitude and phase spectrum of $\mathsf{x}(\mathsf{t})$, and $\theta = \theta_{\mathsf{x}} + \theta_{\mathsf{h}} - \theta_{\mathsf{n}}$. Ignoring the phase term ϵ and assuming that the channel response $|\mathrm{H}|$ is constant over each Mel frequency band, in the log Mel spectral domain we have

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

where y represents the log Mel transform of $|\mathrm{Y}|^2$. As in previous publications, we model the error of this approximation as zero mean and Gaussian distributed:

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

B. Speech Model

As in [3], we model speech by a band-quantized gaussian mixture model (BQ-GMM) [7], which is a constrained, diagonal covariance GMM. BQGMMs have $B \ll S$ shared Gaussians per feature, where S is the number of acoustic components, and so can be evaluated very efficiently.

C. Noise Model

DNA models noise in the Mel spectrum as a Gaussian process. Noise is separated into evolving and transient components, which facilitates robust tracking of the noise level during inference.

The dynamically evolving component of this noise—the noise level—is assumed to be changing slowly relative to the frame rate, and is modeled as follows:

$$p(l_{f,0}) = \mathcal{N}\left(l_{f,0}; \beta_f, \omega^2_{f,0}\right), \quad (5)$$

$$p(l_{f,\tau}|l_{f,\tau-1}) = \mathcal{N}\left(l_{f,\tau}; l_{f,\tau-1}, \gamma^2_f\right), \quad (6)$$

where $l_{f,\tau}$ is a random variable representing the noise level in frequency band f and frame τ. Note that it is assumed that the noise evolves independently at each frequency band. The transient component of the noise process at each frequency band is modeled as zero-mean and Gaussian:

$$p(n_{f,\tau}|l_{f,\tau}) = \mathcal{N}\left(n_{f,\tau}; l_{f,\tau}, \phi^2_f\right). \quad (7)$$

D. Channel Model

In this work channel distortion, h, as in [3], is modeled as a parameter which is stochastically adapted, as described in [8]:

$$p(h_{f,\tau}) = \delta(h_{f,\tau} - \hat{h}_f(\tau)), \quad (8)$$

where $\hat{h}_f(\tau)$ is the current estimate of the channel in frequency bin f at frame τ.

III. Inference

We evaluate the DNA model in sequential fashion, as described in [3]. For a GMM speech model with $|s| = K$ components, and an utterance with T frames, the exact noise posterior for a given frame τ is a K^T component GMM, so approximations need to be made to make inference tractable. As in previous work, here the noise posterior at each frame, given all previous frames, is approximated as Gaussian:

$$p(l_{f,\tau+1}|y_{0:\tau}) \approx \mathcal{N}(l_{f,\tau+1}; \beta_{f,\tau+1}, \omega^2_{f,\tau+1}), \quad (9)$$

We use a variation of Algonquin [9] to iteratively estimate the conditional posterior of the noise level and speech for each speech Gaussian, as in [3]. Algonquin iteratively linearizes the interaction function (3) in (4) given a *context-dependent* expansion point, usually taken as the current estimates of the speech and noise. For a given Gaussian a:

$$p(y|x,n,h) \approx \mathcal{N}(y; \alpha_a(x+h) + (1-\alpha_a)n + b_a, \psi^2), \quad (10)$$

$$\alpha_a = \left.\frac{\delta f}{\delta x}\right|_{\hat{x}_a, \hat{h}_a, \hat{n}_a} = \frac{|\hat{\mathrm{H}}_a|^2 |\hat{\mathrm{X}}_a|^2}{|\hat{\mathrm{H}}_a|^2 |\hat{\mathrm{X}}_a|^2 + |\hat{\mathrm{N}}_a|^2}, \quad (11)$$

$$b_a = f(\hat{x}_a + \hat{h}_a, \hat{n}_a) - \alpha_a(\hat{x}_a + \hat{h}_a - \hat{n}_a) - \hat{n}_a. \quad (12)$$

Given α_a, the posterior distribution of x and n is Gaussian. Once the final estimate of α_a has been determined, the posterior distribution of l can be determined by integrating out the speech and transient noise to get a Gaussian posterior likelihood for l, and then combining it with the current noise level prior. This is more efficient than unnecessarily computing the joint posterior of x, n, and l.

The approximate Minimum Mean Square Error (MMSE) estimate of the Mel speech features for frame τ under DNA is:

$$\hat{x}_{f,\tau} = \mathrm{E}[x_{f,\tau}|\mathbf{y}_{0:\tau}] = \sum_{s_\tau} p(s_\tau|\mathbf{y}_{0:\tau}) \mathrm{E}[x_{f,\tau}|\mathbf{y}_{0:\tau}, s_\tau]. \quad (13)$$

These features are passed to the backend for recognition.

IV. Matched Condition Detection

To detect matched conditions, we introduce a Null Mis-match (NM) model to compete with the current DNA model. The NM model is a degenerate DNA model that contains no explicit noise model, and shares it's speech and channel model with DNA. Note that noise may still be implicitly represented in the NM model. When the test conditions are well matched, explicit noise modeling is redudant, and can hurt ASR performance.

Let $\mathcal{M}_{\text{DNA}}$ and $\mathcal{M}_{\text{NM}}$ denote the current estimates of the DNA model and NM model, respectively. The posterior probability of the DNA model, for a given frame of data is given by:

$$p(\mathcal{M}_{\text{DNA}}|y_\tau) = \frac{1}{1 + \exp(-\alpha f(y_\tau))}, \quad (14)$$

where

$$f(y_\tau) = g(y_\tau) + c, \quad (15)$$

with

$$g(y_\tau) = \log \frac{p(y_\tau|\mathcal{M}_{\text{DNA}})}{p(y_\tau|\mathcal{M}_{\text{NM}})}, \quad c = \log \frac{p(\mathcal{M}_{\text{DNA}})}{p(\mathcal{M}_{\text{NM}})}, \quad (16)$$

and $\alpha = 1$. This is simply Bayes' rule for a binary random variable, with states $\mathcal{M}_{\text{DNA}}$, and $\mathcal{M}_{\text{NM}}$, respectively. α can be tuned to control how "sharp" the posterior estimate is. $f(y_\tau)$ consists of two terms. The first, $g(y_\tau)$, is simply the log likelihood ratio of the two models. c is a bias term equal to the log of the prior ratio of the models.

A shortcoming of (14) is that the relative complexity of the models that are competing to explain the data is not directly taken into account. When deciding what model best represents the observed test features, it makes sense to penalize model complexity. In this case, one model is actually contained within the other. If the NM model can explain the data as well as the DNA model, clearly the NM model should have higher posterior probability, because it has less parameters. The idea that we should trade simplicity only for explanatory power is a fundamental principle of science: Occam's razor. This suggests that the probability of the DNA model should be zero if the NM model can explain the data as well, and indeed, this strategy was the most effective.

The updated equation (14) estimates a frame-level model-posterior for the DNA model.In this work we approximate the model posterior for frame τ given all previous data frames $y_{0:\tau}$ as:

$$p(\mathcal{M}_{\text{DNA}}|y_{0:\tau}) = \gamma p(\mathcal{M}_{\text{DNA}}|y_{0:\tau-1}) + (1-\gamma) p(\mathcal{M}_{\text{NM}}|y_\tau), \gamma \in (0,1) \quad (17)$$

The speech estimate for frame τ is then given by:

$$E[x_t|y_{0:\tau}] = p(\mathcal{M}_{\text{DNA}}|y_{0:\tau}) E_{\mathcal{M}_{\text{DNA}}}[x|y_{0:\tau}] + (1 - p(\mathcal{M}_{\text{DNA}}|y_{0:\tau})) E_{\mathcal{M}_{\text{NM}}}[x|y_{0:\tau}] \quad (18)$$

Note that in contrast with [1], the state of the DNA noise model is not affected by the current posterior probability of the competing model. In [1], a competing noise GMM was introduced to make DNA more robust to abrupt changes in the noise level. When a "reset" condition was triggered by a high noise GMM probability, the evolving noise model in DNA would be re-initialized. Here, the NM model competes with DNA only for influence in the reconstructed speech estimate.

V. Experiments

In [3] we investigated how DNA interacts with Spectral Subtraction (SS), fMLLR and fMMI, using a commerical-grade speech recognizer, trained on large amounts of data. DNA was able to improve SS and fMLLR on ML systems, but had negligible effect on the fMMI-enabled systems. Here we test DNA using yet more data, using a yet better back-end recognition system. We demonstrate that DNA with condition detection (DNA-CD) *significantly* improves recognition recognition performance in all tested scenarios & conditions.

A. Corpora

Experiments were conducted on real data recorded to characterize in-car recognition scenarios. For this purpose, we utilize two proprietary databases, D1 and D2, recorded in two different in-car domains, which include training and testing sets. Our database D1 is identical to the one used in [3]. Audio data is US English in-car speech recorded in various noise conditions (0, 30 and 60 mph), and sampled at 16kHz. A training subset (train1) is composed of 786 hours of speech, with 10k speakers for a total of 800K utterances. A test subset (test1) contains a total of 206k words in 39k utterances from 128 held-out speakers. There are 47 tasks covering four domains (navigation, command & control, digits & dialing, radio) in 7 various US regional accents.

The D2 database is also US English in-car speech but contains a higher percentage of high SNR data. It is recorded by close-talk and far-talk microphones, sampled at 16kHz. The training data (train2) is 2500 hours of speech uttered by thousands of speakers. Prompts cover command & control, spelling, digits, destinations and continuous speech. The test set (test2) is 167k sentences with digits, spelling, command & control, names and destinations. Training and testing conditions are well matched.

For D1, our reference acoustic model is a state-of-the-art 10k Gaussian with 865 context-dependent (CD) states. We use a set of 91 phonemes modeled by three-state hidden Markov models (HMM). fMMI uses a secondary acoustic model with 512 Gaussians, with an inner and outer context of 17 and 9 frames respectively. Once trained, acoustic models are quantized [10]. For D2, the reference acoustic model is a quantized 30k Gaussian model with 1200 CD states. We use 164 three-state phoneme models; digits are modeled separately. Similarly to models for D1, we have online fMLLR, and fMMI transforms with the same inner and outer contexts. Acoustic models are trained under maximum likelihood and then under boosted MMI (bMMI) criterion. For both D1 and D2, the recognition runs using our IBM embedded recognizer. Decoding is done in a single pass using static graphs pre-compiled from task-specific word grammars.

	D1		D2	
	SER (%)	WER (%)	SER (%)	WER (%)
fMMI+bMMI	3.77	1.34	11.82	5.36
fMMI+bMMI+fMLLR (baseline)	3.00	1.08	11.44	5.17
baseline + SS	2.79	1.00	11.38	5.15
baseline + DNA	2.89	0.99	11.59	5.19
baseline + DNA-CD	2.73	0.93	10.72	4.74

TABLE I
WERs and SERs comparing spectral subtraction, DNA, and DNA with matched condition detection (DNA-CD) on the D1 and D2 test databases.

B. DNA Models

The acoustic models used for DNA are trained using speech/noise segmentation coming from a forced-alignment, which is also used for SNR estimation. For each training set (D1 and D2), a DNA speech model was built on all SNR conditions in the training data. DNA speech models are diagonal covariance GMMs, internally represented as BQ-GMMs for speed and storage efficiency. Both DNA models use 256 Gaussians to model speech and 16 Gaussians to model silence in the training data.

C. Recognition Setup

Experiments were conducted using the IBM embedded speech recognizer. Front-end includes a speech end-pointing module that is used to drive spectral subtraction and also turn off the engine in silence parts. Features are 13-dimensional mel-frequency cepstra with deltas and double deltas, mean-normalized, and transformed using LDA, MLLT, fMMI and fMLLR transforms, finally forming a 40-dimensional vector. A hierarchical acoustic scorer produces likelihoods for a decoder using static graphs pre-compiled from constrained task-specific word grammars. The recognition run continuously in one pass, all adaptive transforms work in on-line adaptation mode. DNA (or DNA-CD) is placed in the frontend, after the production of the log Mel feature (and after spectral subtraction if present), and before the conversion to cepstra. DNA/DNA-CD is reset at each utterance. For more details, consult [3].

VI. Experimental Results

We provide the following experimental results for both the D1 and D2 domains:

- Baseline vs. SS, DNA, DNA-CD.
- SNR-dependent results for SS, DNA and DNA-CD.
- DNA-CD results in close and far talk conditions.

A. Overall evaluation

Table I gives overall SERs and WERs for SS, DNA and DNA-CD on databases D1 and D2, using our best embedded ASR system configuration, which utilizes fMMI, a bMMI-trained back-end acoustic model, and stochastic fMLLR. Examining the results, we can see that for dataset D1, both SS and DNA

outperform the baseline, and perform comparably. On dataset D2, SS sightly outperforms the baseline, but DNA degrades overall performance. For both dataset D1 and D2, DNA-CD outperforms both SS and DNA significantly.

B. Performance as a function of SNR

Figure 1 and Figure 2 show the WER and SER as a function of SNR for test datasets D1 and D2 respectively. DNA-CD improves performance in all conditions, and especially at low SNRs (below 15dB).

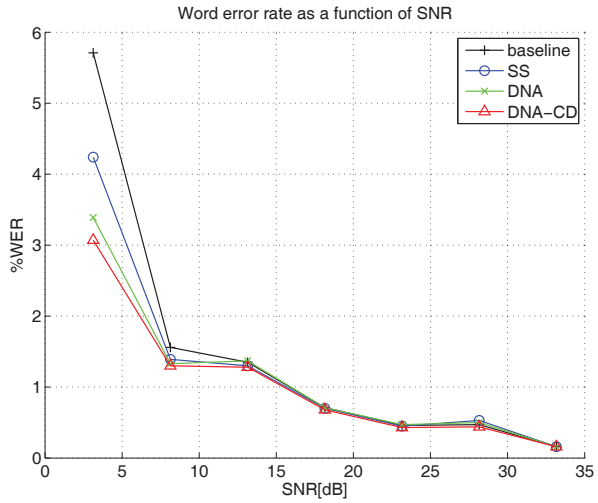

Fig. 1. WER as a function of SNR on test dataset D1. All results were generated using a commercial-grade ASR system that utilizes fMMI, bMMI, and fMLLR.

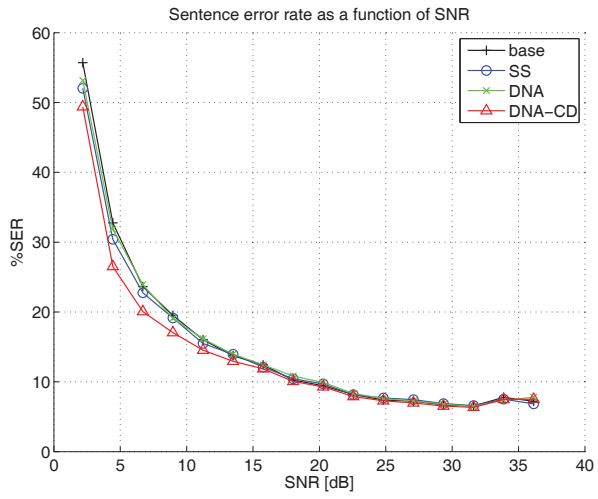

Fig. 2. SER as a function of SNR on test dataset D2. All results were generated using a commercial-grade ASR system that utilizes fMMI, bMMI, and fMLLR.

C. Condition Detection and Channel Noise

DNA-CD significantly improves DNA performance for low SNR conditions. Table II reveals how DNA-CD handles far-talk and close-talk microphones on test dataset D2. In far-talk conditions, DNA-CD improves DNA and SS significantly. For close-talk data, a more moderate improvement is observed, but importantly, DNA-CD does not degrade performance in these less noisy conditions, which are better matched to the acoustic back-end by virtue of the availability of training data in this SNR range.

	Close-talk		Far-talk	
	SER (%)	WER (%)	SER (%)	WER (%)
fMMI+bMMI+fMLLR (baseline)	6.02	2.57	13.99	6.36
baseline + SS	6.27	2.69	13.79	6.27
baseline + DNA	6.12	2.57	14.16	6.38
baseline + DNA-CD	5.90	2.49	12.98	5.76

TABLE II
RECOGNITION PERFORMANCE AS A FUNCTION OF DENOISING ALGORITHM ON CLOSE-TALKING AND FAR-TALKING MICROPHONE DATA (DATABASE D2).

VII. CONCLUSIONS

In this paper, a simple model averaging/selection approach for making online mis-match noise models robust to matched conditions was presented, and applied to DNA. The resulting DNA-CD algorithm improves the performance of our best commercial-grade recognition engines *substantially*, and can be applied to *any* ASR pipeline based on Mel Features, regardless of what noise conditions the back-end system has been exposed to, without causing degradation in matched conditions.

REFERENCES

[1] S. Rennie, T. Kristjansson, P. Olsen, and R. Gopinath, "Dynamic noise adaptation," *ICASSP*, 2006.

[2] S. Rennie and P. Dognin, "Beyond linear transforms: Efficient non-linear dynamic adaptation for noise robust speech recognition," in *Interspeech*, September 2008.

[3] S. Rennie, P. Dognin, and P. Fousek, "Robust speech recognition using dynamic noise adaptation," in *ICASSP*, May 2011.

[4] H. Liao and M. J. F. Gales, "Adaptive training with joint uncertainty decoding for robust recognition of noisy data," in *International Conference on Acoustics, Speech, and Signal Processing*, 2007.

[5] Donglai Zhu and Qiang Huo, "Irrelevant variability normalization based hmm training using map estimation of feature transforms for robust speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, 2008, pp. 4717–4720.

[6] Ozlem Kalinli, Michael L. Seltzer, and Alex Acero, "Noise adaptive training using a vector taylor series approach for noise robust automatic speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 3825–3828.

[7] E. Bocchieri, "Vector quantization for the efficient computation of continuous density likelihoods," in *ICASSP*, 1993, pp. 692–695.

[8] L. Deng, J. Droppo, and A. Acero, "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 11:6, pp. 568–580, 2003.

[9] B.J. Frey, L. Deng, A. Acero, and T. Kristjansson, "Algonquin: Iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition," *Eurospeech*, 2001.

[10] R. Bakis, D. Nahamoo, M. A. Picheny, and J. Sedivy, "Hierarchical labeler in a speech recognition system," U.S. Patent 6023673.