interspeech 2008 incorporating SST 08

Beyond Linear Transforms: Efficient Non-linear Dynamic Adaptation for Noise Robust Speech Recognition

Steven J. Rennie, Pierre L. Dognin

IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, USA
{sjrennie, pdognin}@us.ibm.com

Abstract

In this paper, we present new theory and results that combine constrained Maximum Likelihood Linear Regression (MLLR), known as feature space MLLR (fMLLR), a state-of-the-art model adaptation technique, with Dynamic Noise Adaptation (DNA), a state-of-the-art noise adaptation algorithm. We explain how DNA implements a highly non-linear transform on speech model features, and why DNA is better suited for compensating for additive noise than fMLLR. Tests results are presented on the DNA + Aurora II framework, which is based upon a collection of challenging in-car noise recordings, as a function of SNR. The results demonstrate that DNA significantly outperforms block fMLLR on additive noise, and that DNA + fMLLR outperforms the ETSI advanced front-end (AFE) system + fMLLR by a significant margin (over 7% absolute).

Index Terms: robust speech recognition, model adaptation, fMLLR, Dynamic Noise Adaptation (DNA), ETSI AFE, DNA + Aurora II

1. Introduction

Maximum Likelihood Linear Regression (MLLR) is a relatively simple and proven approach to effective speaker and environmental compensation. MLLR, fMLLR, Maximum a Posteriori Linear Regression (MAPLR), feature space MAPLR, etc. [1, 2], are utilized in one form or another by essentially all state-of-the-art automatic speech recognition (ASR) systems today.

In recent years, much research has gone into developing faster and better-constrained forms of MLLR [3, 4]. fMLLR algorithms, for example, rather than adapting the speech model to the features, efficiently map the features to the model under the affine transformation $\hat{\boldsymbol{x}} = [\boldsymbol{A}\ b][\boldsymbol{x}^T\ 1]^T$. Given the demanding operating constraints of most ASR systems in deployment (speed, storage, adaptation rate, amount of adaptation data), and the performance of MLLRs, it is difficult to improve upon the simplicity, convenience, and efficacy of linear transform-based adaptation schemes.

It is well known, however, that additive noise has a highly non-linear effect in the log frequency domain, and therefore on typically deployed ASR features (e.g. Mel frequency cepstral coefficients, MFCCs). A highly active area of research in robust ASR in recent years has been on better modeling and leveraging this relationship, and utilizing explicit models of noise to isolate the target speech or do noise robust feature labeling [5, 6].

Under this paradigm, typically a codebook of noise is trained a priori and utilized at test time in tandem with a speech model. This is a far more efficient way of representing noisy speech than doing matched training, but can similarly be inappropriate if the test conditions are not known a priori. In particular, when the background is highly non-stationary and hard to characterize, as is the often the case with mobile computing devices and highly dynamic acoustic environments such as cars, representing the acoustic background can be prohibitive. The DNA algorithm, first presented in [7], treats this more typical case in an efficient manner by tracking the state of the noise process over time, and allowing both static and dynamic noise hypotheses to compete to explain the data.

Although the aims of robust ASR and model adaptation research overlap significantly, and many techniques actually share similar methodologies [8], relatively little work investigates how techniques in these fields interact, or attempts to combine the state-of-the-art in these areas.

In this paper, we present new theory and results that combine fMLLR with DNA, state-of-the-art techniques in model adaptation and noise adaptation, respectively. The DNA module used in this paper dynamically adapts to the noise conditions, requires no prior information about the noise, and operates at speeds comparable to those of a typical speech detector. DNA + fMLLR with just 32 DNA speech components outperforms the ETSI AFE + fMLLR by over 6% absolute on the DNA + Aurora II database, and outperforms block fMLLR (on generously constructed blocks) by over 23% absolute.

2. Noisy Speech Model

The model of noisy speech in the time domain is

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

where $*$ denotes linear convolution, $h(t)$ models all channel effects, including propagation distortion and speaker-dependent vocal tract characteristics, and $n(t)$ models all other sources of acoustic interference. In the frequency domain

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

where $|\mathcal{S}|$ and θ_s represent the magnitude and phase spectrum of $s(t)$, and $\theta = \theta_s + \theta_h - \theta_n$. If the distorted speech and noise phases are independent, the expected value of the phase term ϵ is zero, and more generally if the speech dominates the noise or vice versa this term is small.

Ignoring the phase term ϵ, and assuming that the channel response $|\mathcal{H}|$ is constant over each Mel frequency band, in the log Mel spectral domain the relationship becomes

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

for each frequency band, where frequency subscripts are omitted for brevity, and y represents the log Mel transform of $|\mathcal{Y}|^2$.

Accepted after peer review of full paper

Mel binning substantially reduces the error in this approximation. In this work we will model this error as zero mean and gaussian distributed:

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

More accurate models of phase interaction in the log Mel domain have recently been proposed, and improve performance at the expense of increased computation [6]. From (3) we can see that in the log Mel spectral domain (and therefore the Mel cepstral domain) the relationship between the speech and channel is approximately linear, but the relationship between the speech and the noise is highly non-linear.

3. DNA

3.1. Overview

DNA [7, 9] in its more general form models the prior distribution of the noise for a given frame t as the convex sum of a multi-hypothesis dynamic noise model, which is sequentially adapted by treating the noise as a continuously evolving process, and multi-hypothesis static noise model, which can be based on training data, adapted on the fly, or both.

In this work we consider the simplest form of DNA, which utilizes no static models, and propagates a single, dynamic, noise hypothesis at each time step. This instantiation of DNA requires no parameter tuning, no prior information, and can operate at speeds comparable to a standard model-based speech detector when implemented as a front-end module. In this paper we will integrate DNA and fMLLR by applying DNA at the front-end to remove additive noise, and then pass the cleaned features to an fMLLR-enabled ASR system to remove channel effects. DNA could also potentially be integrated directly into the forward pass of Viterbi decoding. In this case DNA and fMLLR could be iterated to further improve performance.

DNA is extremely effective at removing additive noise because:

- the non-linear relationship between speech and noise in the log Mel spectrum is explicitly modeled
- the speech and the noise are jointly estimated for each feature using representative models of speech and noise
- the dynamic noise model distinguishes between transient and evolving fluctuations to do robust tracking

The noise tracking mechanism in DNA implements a *dynamic* forgetting algorithm on the noise and outperforms related tracking-based algorithms that use speech models [7, 9].

In previous work, the focus of explanation and analysis has been on the noise model tracking mechanisms in DNA. In this paper we focus our analysis on how DNA affects the target speech features and show that DNA implements a highly non-linear, dynamic speech feature adaptation algorithm.

3.2. Model

DNA tracks background noise in a robust manner by modeling the noise as consisting of two components: an evolving component, which we call the *noise level*, and a randomly fluctuating component, which is filtered out naturally during inference to facilitate robust noise tracking. We model the noise level at each frequency band as a first-order auto-regressive (AR) process plus additive propagation noise:

$$p(l_t|l_{t-1}) = \mathcal{N}(l_t; l_{t-1}, \omega^2), \quad (5)$$

and the transient component of the noise as gaussian:

$$p(n_t|l_t) = \mathcal{N}(n_t; l_t, \phi^2). \quad (6)$$

Note that noise in the log Mel spectrum is *not* well modeled as a first-order AR process, but is well modeled as a first-order AR process plus noise. These parameters can be estimated from speech free frames at startup, and potentially adapted, but in practice typically $\phi \gg \omega$ because the rate of noise evolution is generally low relative to the frame rate, and the random component of the noise dominates the evolving component. In this case DNA is highly insensitive to these parameters and so parameter adaptation is not necessary to achieve high performance [9]. Once initialized DNA can be run continuously to track background noise, so that when a speech transaction occurs the noise model will be up to date and accurate.

In DNA, the utilized speech model is assumed to be conditionally gaussian with diagonal covariance:

$$p(x_t|s_t^x) = \mathcal{N}(x_t; \mu_{s_t^x}, \sigma^2_{s_t^x}),\ p(s_t^x) = \pi_{s_t^x}, \quad (7)$$

where here $x = h + s$ denotes the noise-free speech, including any channel effects (they will be removed by fMLLR).

An important property of DNA is that it explicitly models the non-linear relationship between the speech and noise as in (4), and *jointly* infers posteriors of the speech and noise.

3.3. Algorithm

Because the interaction between the speech and noise (4) is non-linear, the conditional posterior of the speech and noise is non-gaussian. Several analytic approximations to this posterior distribution exist, including PMC, VTS, and the MAX approximation [9]. Here we will use Algonquin to compute the speech/noise posterior, which is the best performing known analytic method for approximating (4).

Algonquin proceeds by iteratively linearizing the interaction function $f(x,n)$ for each speech/noise combination and each frequency band, given a *context-dependent* expansion point, usually taken as the current (e.g. prior or posterior) estimates of the speech and noise for that speech/noise combination:

$$p(y|x,n) \approx \mathcal{N}(y; a_x x + a_n n + b, \psi^2), \quad (8)$$

$$a_x = \left.\frac{\delta f}{\delta x}\right|_{\hat{x},\hat{n}} = \frac{|\hat{\mathcal{X}}|^2}{|\hat{\mathcal{X}}|^2 + |\hat{\mathcal{N}}|^2}, a_n = \left.\frac{\delta f}{\delta n}\right|_{\hat{x},\hat{n}} = 1 - a_x, \quad (9)$$

$$b = f(\hat{x},\hat{n}) - a_x\hat{x} - a_n\hat{n}. \quad (10)$$

The posterior distribution of the speech and noise given a GMM prior is then also a GMM, and so can be readily computed [5]. The linear assumption about the relationship between the speech and noise is a good approximation only *locally*, *given* estimates of *both* the speech and noise configuration.

In DNA, the instantaneous noise $n = l + \varsigma$ is decomposed into an evolving component, l, and a transient component, ς. In this paper we assume that l and ς are gaussian-distributed, as described in equations (5) and (6), and so the joint prior of $[l\ \varsigma\ x]$ is conditionally gaussian at each frequency. Algonquin can therefore be applied to estimate the joint posterior of $[l\ \varsigma\ x]$ at each frequency by linearizing the likelihood w.r.t. these variables. Another option is to estimate marginals of this joint posterior to reduce the amount of required computation. One way to do this is to estimate $p(x,n|y,s^x)$ using Algonquin and then take $p(l|y,s^x) \propto p(l)\int_n p(n|l)p(n|y,s^x)$, as is done in uncertainty

decoding. Another approach is to use Algonquin to estimate $p(x, l|y, s^x)$ to obtain an estimate of the noise level posterior, and then use Algonquin to estimate $p(x, n|y, s^x)$ to obtain an estimate of the speech posterior. Both of these strategies require the inversion of two 2×2 matrices rather than a 3×3 matrix for each dimension at each iteration, and so are more efficient. In this paper we used the latter approach. Note that using the speech posteriors from $p(x, n|y)$ rather than $p(x, l|y)$ leads to much better speech estimates, because the speech and the *instantaneous* noise are being jointly inferred in this case. This step reduces the Word Error Rate (WER) on the DNA + Aurora II data by more than 5% absolute. The use of the other inference approaches outlined above in DNA has not yet been explored.
At each time-step t, the MMSE estimate of the cleaned speech features $\hat{x}_t$ under the estimated speech posterior (a GMM) is:

$$\hat{x}_t = E[x_t|\boldsymbol{y}_{0:t}] = \sum_{s_t^x} p(s_t^x|\boldsymbol{y}_{0:t}) E[x_t|\boldsymbol{y}_{0:t}, s_t^x]. \quad (11)$$

Similarly, the noise level prior propagated at each time-step is gaussian and has parameters:

$$\mu_{l_{t+1}} = E[l_t|\boldsymbol{y}_{0:t}] = \sum_{s_t^x} p(s_t^x|\boldsymbol{y}_{0:t}) E[l_t|\boldsymbol{y}_{0:t}, s_t^x], \quad (12)$$

$$\begin{aligned} \sigma^2_{l_{t+1}} &= Var[l_t|\boldsymbol{y}_{0:t}] + \omega^2 \\ &= \sum_{s^x} p(s_t^x|\boldsymbol{y}_{0:t}) \{Var[l_t|\boldsymbol{y}_{0:t}, s_t^x] + \\ &\quad (E[l_t|\boldsymbol{y}_{0:t}] - E[l_t|\boldsymbol{y}_{0:t}, s^x])^2\} + \omega^2. \quad (13) \end{aligned}$$

Note that the speech state posterior, $p(s_t^x|\boldsymbol{y}_t)$, couples these updates over frequency.

3.4. Non-linear feature/model adaptation using DNA

The marginal distribution of the noisy speech y given the speech state s^x at each frequency band is given by:

$$p(y|s^x) = \mathcal{N}(y; a_x\mu_x + a_n\mu_n + b, a_x^2\sigma_x^2 + a_n^2\sigma_n^2 + \psi^2), \quad (14)$$

and therefore DNA adapts the speech model as follows:

$$\mu_x \longmapsto a_x\mu_x + a_n\mu_n + b, \quad (15)$$
$$\sigma_x^2 \longmapsto a_x^2\sigma_x^2 + a_n^2\sigma_n^2 + \psi^2. \quad (16)$$

When the speech dominates the frequency band, $a_x \approx 1$, and $\mu_x \mapsto \mu_x$. When the noise dominates, $a_n \approx 1$, and $\mu_x \mapsto \mu_n$. When the noise and speech have similar power μ_x is mapped accordingly under the speech/noise interaction model. Similarly when the speech dominates $\sigma_x^2 \mapsto \sigma_x^2 + \psi^2$, where $\psi^2 \ll \sigma_x^2$ inherently, and $\sigma_x^2 \mapsto \sigma_n^2 + \psi^2$, when the noise dominates. In summary, in frequency bands where the noise dominates the speech model, the features are transformed to the noise features, to reflect more appropriately the likelihood of each speech configuration given the noise conditions. Where the speech dominates, the speech model is not modified.
Note that the DNA noise model encodes with a single gaussian a set of *speech-state-specific transformations* of the speech model. In contrast, fMLLR with a single linear transformation can only compensate effectively for global mismatch between the speech model and the data. Furthermore, while background noise generally has naturally evolving structure in the log Mel spectrum, the appropriate linear transform to apply may change abruptly depending on what speech states are active.
DNA adapts $2F$ variables, the means and variances of each frequency band, over time. Loosely speaking, where the noise dominates the speech the noise estimate is updated and the uncertainty in the noise configuration is very low. Where the speech dominates updating of the noise estimate shuts down and the variance of the noise model increases. The inherently pseudo-continuous nature of the underlying noise level makes the strategy highly effective and stable.

4. Experimental Results

Test data was generated by embedding clean TIDIGITS (Aurora II set A) utterances into an approximately 2.5 hour database of non-stationary, diverse car noise, at various SNRs, using the publicly available DNA + Aurora II framework/database [7] (104 speakers, 11443 words/SNR, 0 utterance gap).
Acoustic models were built using the Aurora II clean training set (6 hours of TIDIGITS). All 11 spoken digit words ('1-9', '0' and 'oh') were modeled as a sequence of tri-state phonemes. The shortest digit ('oh') has one phoneme while the longest ('seven') has five. Long word models could be used to improve performance at the expense of task specificity.
A flat-start, context independent (CI) model was trained on speaker-dependent Cepstral Mean Normalized (CMN) MFCCs (c0-c12) using 30 iterations of Expectation-Maximization, and used to generate alignments and create a context-dependent (CD) model. Linear Discriminant Analysis (LDA) was used to project the MFCCs (9 frames at a time) to 40-dimensional features. A context-dependency tree with 400 CD states and 5000 diagonal covariance gaussians with two phoneme left and right context was learned using a predefined question set.
Decoding was performed using the CD acoustic model and a static decoding graph compiled from a digit loop grammar. At test time, cleaned MFCC features derived from 20-30 second, contiguous, speaker-specific blocks of the test data were fed into the decoder, normalized, and projected into LDA space. An fMLLR transformation was then estimated independently for each block by iterating between decoding each (1-5 second) utterance within the block, and updating the features based upon the obtained alignments two times. Each utterance was then decoded using the adapted features. Blocks of utterances were used in an attempt to boost the performance and stability of fMLLR, which are known issues with fMLLR when adaptation is done on limited data.
Features were denoised at the front-end either by DNA, or by the ETSI AFE system [10], which is the most well known and highly regarded real-time denoising system that we are aware of. Both systems are real-time, sequential, low-latency algorithms with computational requirements comparable to that of a conventional model-based speech detector. In contrast to how fMLLR was applied, both DNA and the AFE were run in continuous mode on each DNA + Aurora II database file (5 - 15 minutes long) to stress test their robustness, and evaluate the merits of maintaining a continuously evolving background noise model. Note that the DNA + Aurora II database is much better suited for evaluating the robustness of adaptive noise compensation algorithms than the more standard Aurora II task, whose test files are unordered, and only 2-5 seconds long. For results comparing the performance of DNA to the AFE and other algorithms on the Aurora II-M task, created by embedding utterances into *raw* Aurora II noise files, see [9].
Table 1 depicts WER results as a function of SNR on the DNA + Aurora II dataset. Looking at the results, we can see that DNA + fMLLR with 32 speech prototypes in the speech model

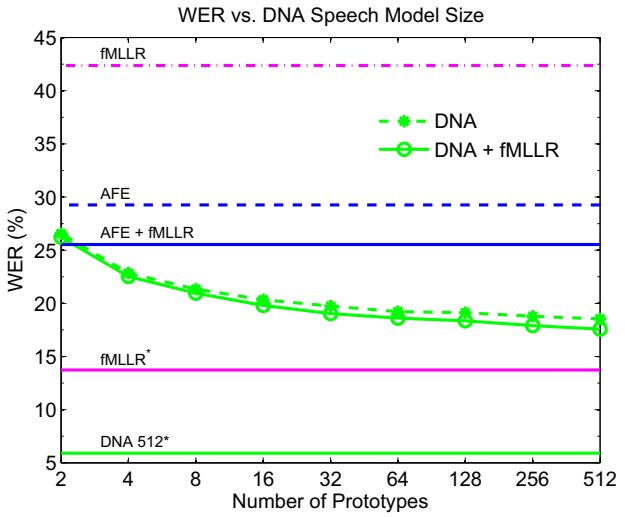

Figure 1: WER performance as as a function of the number of DNA speech components on the DNA + Aurora II dataset.

Method	WER (%) vs. SNR (dB)						
	20	15	10	5	0	-5	All
RAW	9.0	22.1	41.2	60.9	76.6	86.4	49.3
AFE	**3.1**	7.6	16.9	31.4	49.7	66.9	29.3
DNA 32	**2.2**	**4.1**	8.2	16.7	32.9	54.5	19.7
DNA 512	**2.2**	**3.9**	**7.6**	15.7	30.2	51.6	18.5
DNA 32*	**2.1**	**2.8**	**3.7**	**4.8**	**9.9**	**25.8**	**8.2**
DNA 512*	**1.9**	**2.2**	**2.7**	**3.0**	**6.0**	**19.8**	**5.9**
w/ fMLLR							
RAW	3.3	12.1	29.6	52.4	72.4	84.5	42.4
AFE	**1.8**	**4.4**	11.8	25.7	45.5	64.0	25.5
DNA 32	**1.7**	**3.5**	**7.5**	15.5	31.9	54.2	19.2
DNA 512	**1.8**	**3.1**	**6.7**	14.4	29.0	50.5	17.6
fMLLR*	3.3	5.5	7.7	11.9	20.5	33.5	13.7

* cheating experiments: 20 dB state alignments used to adapt feature transforms

Table 1: WER as a function of denoising algorithm and SNR. The results obtained by our system on RAW features, features denoised by the ETSI AFE, and features denoised by DNA using 32 and 512 prototype speech models are depicted, without and with fMLLR enabled.

outperforms AFE + fMLLR by over 6% absolute overall, and fMLLR by over 23% absolute. The efficacy of fMLLR is highly dependent upon the accuracy of the initial alignments used to adapt the fMLLR matrix. Running DNA (or the AFE) at the front-end leads to much better initial alignments, and therefore much better performance.

The fMLLR* entry in table 1 depicts the results obtained by cheating and using the initial alignments obtained on the 20 dB SNR data to adapt the fMLLR transform for all SNRs. This gives us a rough upper-bound on what is possible using block fMLLR. Entries in the table where the performance of fMLLR* is matched or exceeded are bolded. Impressively, DNA + fMLLR with just 32 prototypes outperforms fMLLR* when the SNR is at or above 10 dB. Also included in table 1 are corresponding results for DNA 32* and DNA 512*, where we use the DNA speech state alignments determined on 20 dB data to transform the speech features. DNA 32* and DNA 512* outperform fMLLR* by 5.5% and 7.8% absolute, respectively.

Figure 4 depicts a plot of overall WER versus the number of prototypes utilized by the DNA model. From the plot we can see that the performance of DNA + fMLLR (and DNA alone) surpasses that of AFE + fMLLR with only $K = 4$ components in the DNA speech model. The results collectively illustrate the importance of the initial alignment used to adapt fMLLR, and how dramatically performance can be improved when fMLLR is bootstrapped by DNA. The results also show that DNA is an extremely effective additive denoising system, which outperforms block fMLLR significantly.

5. Discussion

In this paper we have presented DNA: a non-linear, low-latency feature adaptation algorithm for noise-robust feature labeling and denoising. DNA operates on log Mel features and so can be seamlessly be integrated into the front-end of conventional speech recognition systems, and has computational requirements comparable to that of a conventional model-based speech detector. DNA + fMLLR with 32 DNA speech components outperforms the ETSI AFE + fMLLR by over 6% absolute on the DNA + Aurora II database, and block fMLLR (on generously constructed blocks) by over 23% absolute on the task.

When we cheat and use 20 dB alignments to bootstrap fMLLR, the overall WER is 13.7%. When DNA is run with 20 dB DNA speech model alignments, 512 prototypes, and fMLLR disabled, in contrast, the overall WER on the task is just 5.9%. Interesting directions of future work include integrating some form of DNA directly into the labeler of the decoder, and comparing the stability and performance of DNA to online, highly constrained forms of fMLLR.

6. Acknowledgements

Many thanks to John Hershey and Peder Olsen for suggestions that inspired and strengthened this work.

7. References

[1] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *CSL*, vol. 12, 1998.

[2] C. Chesta, O. Siohan, and C-H. Lee, "Maximum a posteriori linear regression for HMM adaptation," *Eurospeech*, 1999.

[3] B. Varadarajan and D. Povey, "Quick fmllr for speaker adaptation in speech recognition," *ICASSP*, 2008.

[4] K. Visweswariah, V. Goel, and R. A. Gopinath, "Structuring linear transformations for adaptation using training time information.," *ICASSP*, 2002.

[5] 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.

[6] L. Deng, J. Droppo, and A. Acero, "Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," *ISAP*, vol. 12:2, pp. 133–143, 2004.

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

[8] L. Deng, J. Wu, J. Droppo, and A. Acero, "Analysis and comparison of two speech feature extraction/compensation algorithms," *ISPL*, vol. 12:6, 2005.

[9] S. Rennie, *Graphical models for robust speech recognition in adverse environments*, Ph.D. thesis, Univ. of Toronto, 2008.

[10] ETSI, "Advanced front-end feature extraction algorithm," *ETSI TR ES 202 050 V 1.1.3*, 2003.