DYNAMIC NOISE ADAPTATION

Steven Rennie, Trausti Kristjansson, Peder Olsen, Ramesh Gopinath

IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, USA

ABSTRACT

We consider the problem of robust speech recognition in the car environment. We present a new dynamic noise adaptation algorithm, called DNA, for the robust front-end compensation of evolving semi-stationary noise as typically encountered in the car setting. A large dataset of in-car noise was collected for the evaluation of the new algorithm. This dataset was combined with the Aurora II framework to produce a new, publicly available framework, called DNA + AURORA II, for the evaluation of adaptive noise compensation algorithms. We show that DNA consistently outperforms several existing, related state-of-the-art front-end denoising techniques.

1. INTRODUCTION

We consider the problem of speech recognition in the car environment, motivated by the goal of noise-robust, speech-based user interaction with onboard navigation and control systems.

Environmental compensation in the car environment is challenging because the acoustic background is generally non-stationary and hard to characterize. This makes both matched training and model-based noise removal at the front-end difficult to implement effectively in practice.

The acoustic background is generally comprised of a wide variety of noise types, including: a) quasi-stationary evolving disturbances (acceleration and deceleration related road, wind, and engine noise; passing cars), b) abrupt, step changes in the quasi-stationary noise (an abrupt change in the road surface), c) noise transients (shutting doors, car signals, bumps, horns, windshield wipers), as well as d) more complicated disturbances (the radio, secondary speakers). In this paper, we focus on quasi-stationary noise with abrupt changes as encountered in real data.

One general approach to speech denoising has been to utilize trained codebooks of clean speech and noise to separate out the unwanted noise signal. Several results demonstrating substantial reductions in word error rate (WER) under stationary noise conditions have been reported [1, 2]. Static noise models, however, are not well suited to typical car noise, which predominantly consists of semi-stationary evolving noise components.

An alternate approach to using static noise models is to 'track' the background noise process stochastically [3, 4].

In this paper, we bring together and extend upon these approaches and present a new Dynamic Noise Adaptation algorithm (DNA) for simultaneously tracking evolving semi-stationary noise and producing clean speech estimates.

In addition, we introduce a new, publicly available extension of the Aurora II framework, called the DNA + AURORA II evaluation framework. The framework is based upon a large database of challenging in-car noise (single microphone, recordings 5-20 minutes long) and a collection of scripts for embedding and extracting speech utterances in the data according to a specified utterance gap profile. The framework provides a public platform for the systematic development, analysis, and comparison of adaptive speech denoising algorithms on realistic data, in realistic human-computer interaction modes.

Finally, we present recognition results on the new DNA + AURORA II task demonstrating the performance of our DNA algorithm, and show that it outperforms several existing, related state-of-the-art noise compensation algorithms.

2. MODEL OF NOISY SPEECH

The model for noisy speech in the time domain is (omitting the channel for simplicity)

$$y[t] = x[t] + n[t] \tag{1}$$

where $x[t]$ denotes the clean signal, $n[t]$ denotes the noise, and $y[t]$ denotes the noisy signal. In the power spectrum, the relationship becomes:

$$|Y(f)|^2 = |X(f)|^2 + |N(f)|^2 - 2|X(f)||N(f)|cos(\phi) \tag{2}$$

where $\phi = \angle X(f) - \angle N(f)$. After the Mel transform and the logarithm [1] we arrive at the following relationship in the log Mel power spectral domain:

$$\mathbf{y} = \ln(\exp(\mathbf{x}) + \exp(\mathbf{n})) + \boldsymbol{\epsilon} \tag{3}$$

where $\boldsymbol{\epsilon}$ models the phase term in (2), and is assumed to be a zero mean diagonal covariance Gaussian random variable. For a given processing frame t, then, the noisy speech features $\mathbf{y}_t$ are modelled as conditionally Gaussian:

$$p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{n}_t) = N\left(\mathbf{y}_t; \ln\left(\exp(\mathbf{x}_t) + \exp(\mathbf{n}_t)\right), \boldsymbol{\Psi}\right) \tag{4}$$

where $\boldsymbol{\Psi}$ is the covariance of $\boldsymbol{\epsilon}_t$. We use a mixture of diagonal covariance Gaussians to model clean speech:

$$p(\mathbf{x}_t) = \sum_{s_t^x} \pi_{s_t^x} N(\mathbf{x}_t; \boldsymbol{\mu}_{s_t^x}, \boldsymbol{\Sigma}_{s_t^x}) \tag{5}$$

and model the noise at time t, *conditioned* on the previous observations $\mathbf{y}_{0:t-1}$, as a mixture of diagonal covariance Gaussians:

$$p(\mathbf{n}_t|\mathbf{y}_{0:t-1}) = \sum_{s_t^l} \pi_{s_t^l} N(\mathbf{n}_t; \boldsymbol{\mu}_{s_t^l|t-1}, \boldsymbol{\Sigma}_{s_t^l|t-1} + \boldsymbol{\Sigma}_{ln}) \tag{6}$$

where $\boldsymbol{\mu}_{s^l|t-1}$ and $\boldsymbol{\Sigma}_{s^l|t-1}$ are the mean and covariance estimates of the conditional prior of the continuously evolving component

of the noise at time t, which we call the *noise level* and denote by $\mathbf{l}_t$, and $\mathbf{\Sigma}_{ln}$ is the covariance of the random component of the noise, which is modelled as zero mean and Gaussian. A detailed description of our noise model and the associated noise-estimation propagation algorithm will be given in sections 3 and 4.

The conditional posterior distribution of the speech and noise at frame t given $\mathbf{y}_{0:t}$, s_t^x, and s_t^l is given by:

$$p(\mathbf{x}_t, \mathbf{n}_t | \mathbf{y}_{0:t}, s_t^l, s_t^x) = \frac{p(\mathbf{n}_t | s_t^l, \mathbf{y}_{0:t-1}) p(\mathbf{x}_t^f | s_t^x) p(\mathbf{y}_t | \mathbf{x}_t, \mathbf{n}_t)}{\int_{\mathbf{n}_t} \int_{\mathbf{x}_t} p(\mathbf{n}_t | s_t^l, \mathbf{y}_{0:t-1}) p(\mathbf{x}_t^f | s_t^x) p(\mathbf{y}_t | \mathbf{x}_t, \mathbf{n}_t)} \quad (7)$$

Because the relationship (3) is non-linear, (7) cannot be evaluated analytically. An approximation to this conditional posterior can be computed via the Algonquin or Laplace methods, which are described in detail in [1, 2]. The essence of these approaches is to iteratively linearize (3) and re-estimate a (conditionally) Gaussian posterior distribution of the speech and noise using (7), until convergence.

3. DYNAMIC NOISE MODEL

Our dynamic noise model is based upon the observation that continuously evolving noise in the log Mel power spectral domain (LMPSD) at a given frequency can generally be well modelled as consisting of two distinct components: a slowly evolving time-varying component, which we call the noise level; and a zero mean IID random component, which mixes additively with the noise level in the LMPSD at each frequency to generate the corrupting noise signal.

Define $\mathbf{l}_t$ as the underlying noise level at frame t, and define the prior for $\mathbf{l}_0$ as a (diagonal-covariance) mixture of Gaussians, with conditional prior at each frequency given by:

$$p(l_0 | s_0^l) = N(l_0; \mu_{s_0^l}, \sigma_{s_0^l}) \quad (8)$$

where frequency sub-scripts have been omitted to avoid notational clutter. We model the temporal evolution of the noise level at each frequency as a first-order AR process:

$$p(l_{t+1} | l_t) = N(l_{t+1}; l_t, \sigma_d^2) \quad (9)$$

and the conditional probability of the noise n_t given the noise level at each frequency as zero-mean and Gaussian:

$$p(n_t | l_t) = N(n_t; l_t, \sigma_{ln}^2) \quad (10)$$

The key feature of this simple generative model of dynamic noise is that the variability of the random component of the noise (σ_{ln}^2) and the variability of the noise due to the evolution of the noise process (σ_d^2) are explicitly modelled as distinct entities. In noise dominant sections of the data, σ_{ln}^2 and σ_d^2 in tandem under the structure of the model facilitate the robust tracking of the evolution of the noise process, by filtering out the random component of the noise. In sections of the data where speech dominates, on the other hand, the noise process is effectively 'hidden'. By modelling the evolving and randomly varying parts of the noise as distinct components, the rate of growth of the uncertainty in the noise level is made independent of the amount of random noise corruption. The key point is that the overall noise process in the LMPSD is generally *not* well modelled as a first order AR process, because the randomly varying component of the noise is usually quite significant ($\sigma_{ln}^2 \gg \sigma_d^2$). Indeed, we have experimented with modelling n_t directly as a first-order AR process on car noise, and found that the algorithm was highly unstable.

3.1. Dealing with abrupt changes in noise level

In practice, the underlying noise level in the car environment will occasionally change dramatically and abruptly (a window is opened at high speed, the road surface changes), a situation to be distinguished from a short-duration noise transient or strong random fluctuation. In such cases even the noise level certainly not a first order AR process with Gaussian noise, and modelling it strictly as such can result in loss of track (when a large positive change in the noise level occurs, for example, the system can get mistakenly 'stuck' in the 'speech dominant' mode of operation, and the noise level will not be updated as time proceeds). We handle abrupt changes in the noise level by introducing a 're-start' random variable, r_t, at each time step, with low prior probability of activation, that facilitates the automatic detection of loss of track, and the 'switching in' of a re-start model for the noise. The marginal prior for the noise level at time t, then, is given by a convex combination of the propagated noise prior, and the re-start noise prior:

$$\begin{aligned} &p(l_t | y_{0:t-1}) \\ &= p(r_t = 0) p(l_t | y_{0:t-1}, r_t = 0) + p(r_t = 1) p(l_t | r_t = 1) \\ &= \pi_{r_t=0} \sum_{s_p^t} \pi_{s_p^t} N(l_t; \mu_{s_p^t}, \sigma_{s_p^t}^2) + \pi_{r_t=1} \sum_{s_r^t} \pi_{s_r^t} N(l_t; \mu_{s_r^t}, \sigma_{s_r^t}^2) \\ &= \sum_{s_l^t} \pi_{s_l^t} N(l_t; \mu_{s_l^t}, \sigma_{s_l^t}^2) \quad (11) \end{aligned}$$

where both the propagated conditional prior of the noise level and the re-start model are taken as mixtures of Gaussians (the noise level propagation algorithm will be discussed in detail in the next section), and the variable s_l^t represents the mixture index of the marginal prior for the noise level, which is also a mixture of Gaussians [1].

The utilized re-start model can be context-dependent. For example, we have found that propagating a single re-start mode with mean equal to the output of a min filter on the LMPSD input $\mathbf{y}$ over a short (eg. 0.5 second) window, (an intuitive approach for recovering the noise level) efficiently and effectively faciliates the automatic recovery of the noise level when abrupt changes in the noise level lead to loss of track.

Figure 1 depicts a Bayes Net summarizing the dependencies that exist between random variables of the DNA model.

4. SPEECH/NOISE INFERENCE

Given the conditional prior of the noise level at a given frequency for frame t:

$$p(l_t | s_t^l, y_{0:t-1}) = N(l; \mu_{s_t^l | t-1}, \sigma_{s_t^l | t-1}^2) \quad (12)$$

and the (approximate, linearized) likelihood function:

$$p(y_t | n_t, x_t) = N(y_t; c + [a_x \ a_n][x_t \ n_t]^T, \psi^2) \quad (13)$$

[1] Note that when re-start functionality is enabled $p(l_{t+1} | l_t, r_{t+1} = 0) \equiv p(l_{t+1} | l_t)$ as defined previously.

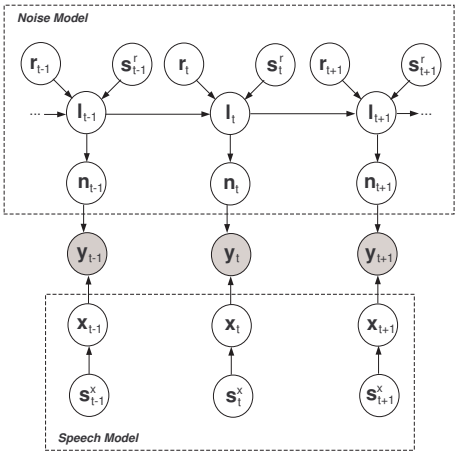

Fig. 1. DNA: Generative Model

the marginal likelihood of the noise level is given by:

$$p(y_t|l_t, s_t^x) = \int\int p(y_t|n_t, x_t)p(n_t|l_t)p(x_t|s_t^x)dx_t dn_t$$
$$= N(l; \frac{1}{a_n}(y - c - a_x\mu_{s_t^x}), \frac{1}{a_n^2}\psi^2 + \frac{a_x^2}{a_n^2}\sigma_{s_t^x}^2 + \sigma_{ln}^2)$$
$$= N(l; \mu_{lik_{l_t,s_t^x}}, \sigma_{lik_{l_t,s_t^x}}^2) \quad (14)$$

The conditional posterior of the noise level is then given by:

$$p(l_t|s_t^l, s_t^x, y_{0:t}) = N(l; \mu_{s_t^l,s_t^x|t}, \sigma_{s_t^l,s_t^x|t}^2) \quad (15)$$

where:

$$\sigma_{s_t^l,s_t^x|t}^2 = \left(\frac{1}{\sigma_{s_t^l|t-1}^2} + \frac{1}{\sigma_{lik_{l_t,s_t^x}}^2}\right)^{-1} \quad (16)$$

$$\mu_{s_t^l,s_t^x|t} = \left(\sigma_{s_t^l,s_t^x|t}^2\right)^{-1}\left(\frac{\mu_{s_t^l|t-1}}{\sigma_{s_t^l|t-1}^2} + \frac{\mu_{lik_{l_t,s_t^x}}}{\sigma_{lik_{l_t,s_t^x}}^2}\right) \quad (17)$$

The update for the conditional posterior mean (mode) of the noise level is therefore given by a convex combination of prior and data influences, where the relative weight assigned to each influence depends on the relative uncertainty (inverse variance) associated with each information source.

The conditional prior for l_{t+1} is computed from the conditional posterior of l_t via:

$$p(l_{t+1}|s_t^l, s_t^x, y_{0:t}, r_{t+1} = 0) =$$
$$\int p(l_{t+1}|l_t, r_{t+1} = 0)p(l_t|, s_t^l, s_t^x, y_{0:t})dl_t$$
$$= N(l_{t+1}; \mu_{s_t^l,s_t^x|t}, \sigma_d^2 + \sigma_{s_t^l,s_t^x|t}^2) \quad (18)$$

If this mode is chosen for propagation, $p(l_{t+1}|s_{t+1}^p, y_{0:t}, r_{t+1} = 0) = p(l_{t+1}|s_t^l, s_t^x, y_{0:t}, r_{t+1} = 0)$, as will be discussed shortly.

Equations (12-18) collectively define how information about the noise level is extracted from the data and propagated during inference. Looking at the expression for $\sigma_{lik_{l_t,s_t^x}}^2$ in (14), we can see that when the speech dominates ($a_x \gg a_n$), $\sigma_{lik_{l_t,s_t^x}}^2$ will be very large, and therefore the influence of the observation on the update of $\mu_{s_t^l,s_t^x|t}$ in (17) will be very small. When the noise dominates, on the other hand, $\sigma_{lik_{l_t,s_t^x}}^2 \approx \sigma_{ln}^2$, and σ_{ln}^2 and σ_d^2 implement a low-pass filter on the noise, with the noise level as output.

The number of modes in the noise level posterior at each time step is equal to the product of the number of modes in the speech prior, and the number of modes in the noise level prior, and therefore exact inference (of the speech or noise) under the model is generally intractable, as the number of modes in the exact posterior grows exponentially with time. One simple and very general solution to this problem is to propagate a chosen subset of modes K_t at each time-step as an approximation of the true posterior distribution of the noise level:

$$p(l_{t+1}|\mathbf{y}_{0:t}, r_{t+1} = 0) \simeq$$
$$\frac{\sum_{\{s_t^l,s_t^k\}\in K_t} p(s_t^l, s_t^k|\mathbf{y}_{0:t})p(l_{t+1}|s_t^l, s_t^x, y_{0:t}, r_{t+1} = 0)}{\sum_{\{s_t^l,s_t^k\}\in K_t} p(s_t^l, s_t^x|\mathbf{y}_{0:t})}$$
$$= \sum_{s_{t+1}^l} \pi_{s_{t+1}^p} N(l_t; \mu_{s_{t+1}^p|t}, \sigma_{s_{t+1}^p|t}^2 + \sigma_d^2) \quad (19)$$

In this paper the set K_t is taken as the K most probable modes under the mode posterior $p(s_t^l, s_t^k|\mathbf{y}_{0:t})$.

At each frame an MMSE estimate of the clean speech is computed via:

$$\hat{\mathbf{x}}_t = \sum_{s_t^x, s_t^l} p(s_t^x, s_t^l|\mathbf{y}_{0:t}) \int \mathbf{x}_t p(\mathbf{x}_t|s_t^x, s_t^l, \mathbf{y}_{0:t})d\mathbf{x}_t \quad (20)$$

Note that both the posterior of the noise level, and the MMSE estimate of clean speech at each frame are coupled over frequency by the posterior $p(s_t^l, s_t^k|\mathbf{y}_{0:t})$.

5. EXPERIMENTS

5.1. The DNA + Aurora II evaluation framework

Over two hours of challenging in-car noise was recorded with a microphone attached to the passenger side visor. The data was recorded over the course of 7 complete car trips with naturally varying background noise conditions, comprised of both evolving noise (due to acceleration/deceleration, changing road conditions, passing cars, rain etc.) and transients (bumps, slamming doors, car signals etc.). The collected noise data was then partitioned into 10 files (one trip was broken into three files), each between 5 and 20 minutes long, to define a car noise database.

To generate test data, clean speech from Aurora 2 dataset A (subway) (1001 utterances, consisting of exclusively spoken digits) [5] was artificially embedded into the noise database at various average SNRs [2], and various settings of the utterance gap vector (UTV), which defines (cyclicly) the amount of gap between successive utterances in seconds (e.g. UTV = [2 0] seconds means that the utterance gap cycles between 2 and 0 seconds). The UTV parameter has been defined as such so that the performance of denoising systems as a function of utterance length and/or gap can be analyzed.

[2]The average SNR here was defined as the square root of the ratio of the average speech energy level (output by the ITU software [6, 5]) over all utterances in a given mixed file, and the average noise power, excluding the utterance gaps from the calculation.

METHOD		WA (%)	vs.	SNR (dB)	
	15	10	5	0	-5
DNA1	97.94	**97.14**	94.78	82.50	81.09
DNA1r	**98.15**	97.09	**94.89**	**91.43**	82.62
RSNE+	97.24	95.89	93.58	88.01	80.09
ALQN1FF	96.54	90.87	78.39	67.35	49.50
ALQN1FFr	97.85	96.79	94.19	90.63	**84.02**
NONE	94.34	84.47	69.24	52.25	38.13

Table 1. Word accuracy as a function SNR and applied denoising algorithm on the new DNA + Aurora II dataset. Here UTV = [4] seconds.

METHOD	WA(%)	vs.	UTV
	UTV = [4]	UTV = [2 0]	UTV = [0]
DNA1	82.50	88.97	88.39
DNA1r	**91.43**	**90.37**	**89.78**
RSNE+	88.01	84.79	81.63
ALQN1FFr	90.63	89.78	89.13
NONE	52.25	51.39	51.72

Table 2. Word recognition accuracy as a function of noise observability and applied denoising algorithm on the new DNA + Aurora II dataset. SNR = 0 dB.

The collected car noise and scripts for artificially embedding and then extracting the Aurora II utterances for recognition are publicly available for download at **www.sonsyn.com/DNA**. The overall framework has been designed as an extension of the Aurora II task, created for the evaluation of adaptive speech denoising algorithms on realistic data, in realistic human-computer interaction modes.

5.2. Results

Tables 1 and 2 depict word accuracy (WA) results on the DNA + Aurora II dataset, for various SNRs and UTV = [4], and varying utterance gap vector UTV at 0 dB, respectively, for the following front-end speech de-noising systems: 1) DNA1: DNA with one noise mode propagated at each time step, and re-start detection disabled. 2) DNA1r: DNA with one noise mode propagated, and automatic re-start detection enabled (re-start mode defined by a 0.5 second long causal min filter on the input, with prior probability 0.001) 3) RSNE+: a) The recursive stochastic noise estimation algorithm (Deng et. al.) [3] to estimate the noise level (in the LMPSD), followed by b) the application of Algonquin [1] to estimate the speech and noise given the noise level estimate [3]. 4) ALQN1FF: (single gaussian noise prior) Algonquin [1] , with an evolving, (fixed forgetting factor on the noise posteriors) estimate of the noise mean. 5) ALQN1FFr: ALQN1FF with the same re-start functionality as DNA1r added. 6) NONE: no pre-processing front end. For all of the algorithms, a 128 component diagonal covariance GMM speech model trained on clean speech features in the LMPSD domain (from the Aurora II training set) was used. In all cases, the first 20 frames of each test file (each 5 to 20 minutes long) were used to initialize the noise model of the algorithm.

Ignoring the algorithms with reset functionality for a moment, we can see that DNA1 outperforms both the ALQN1FF and RSNE+ algorithms in all tested scenarios, with the exception of the result for UTV=[4] and SNR= 0 dB, where RSNE+ significantly outperforms DNA1. Investigation revealed that both the DNA1 and ALQN1FF algorithms would occasionally and often, respectively, loose track of the noise process when very abrupt changes in the noise level occur, and not always recover. DNA1 is presumably more robust to loss of track than ALQN1FF because the noise level prior covariance in DNA1 is dynamic, making DNA more robust to changing conditions. RSNE+ was found to never loose track: a positive consequence of this algorithm implementing a fixed forgetting rate on the noise level likelihoods. The negative consequence of this property that was observed is that this makes the RSNE+ algorithm prone to significant speech leakage because the algorithm utilizes a fixed 'window of influence' to update the noise estimate. The negative effects of this property of the algorithm become more pronounced when the observability of the noise process goes down, as the results in Table 2 demonstrate. Both DNA1 and ALQN1FF implement dynamic forgetting rate algorithms on the noise level likelihood, and so are robust to speech leakage, but unfortunately it is the same dynamic mechanism that can cause them to loose track when abrupt changes in the noise level occur.

Looking now at the results with re-start functionality enabled, we can see that the performance improvement of the DNA1r and ALQN1FFr over the DNA1 and ALQN1FF algorithms is substantial. Inspection of the results revealed that the re-start functionality had made the DNA1r and ALQN1FFr algorithms robust to abrupt changes in the noise level, while retaining the desirable noise tracking properties of DNA1 and ALQN1FF. The performance of ALQN1FFr (an algorithm presented for the first time, here) surprised us. We can see that although DNA1r performs better essentially always, ALQN1FFr is always very close behind. The re-start functionality has almost fully compensated for the fact that the covariance of the noise prior in ALQN1FF(r) is not dynamic, but not fully presumably because min filter-based re-starts are generally associated with substantial recovery lags. DNA1r on the other hand avoids having to re-start in most situations, automatically adapting to changing noise conditions by virtue of its dynamic prior covariance on the noise level.

[3]The post-processing step b) was used rather than Splice [3], since the Splice algorithm must be trained on corresponding clean and noisy speech data.

6. REFERENCES

[1] 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.

[2] T. Kristjansson, J. Hershey, and H. Attias, "Single microphone source separation using high resolution signal reconstruction," *ICASSP*, 2004.

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

[4] M. Afify and O. Siohan, "Sequential noise esitmation with optimal forgetting for robust speech recognition," *ICASSP*, 2001.

[5] Hirsch H.G. and Pearce D., "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *ASR*, 2000.

[6] ITU recommendation P.56, "Objective measurement of active speech level," 1993.