PREDICTED WALK WITH CORRELATION IN PARTICLE FILTER SPEECH FEATURE ENHANCEMENT FOR ROBUST AUTOMATIC SPEECH RECOGNITION

Matthias Wölfel

Institut für theoretische Informatik, Universität Karlsruhe (TH)
Am Fasanengarten 5, 76131 Karlsruhe, Germany
wolfel@ira.uka.de

ABSTRACT

Previous particle filter feature enhancement techniques for robust automatic speech recognition have ignored the fact that neighbored spectral bins are correlated. In those cases, the spectral bins have been treated as uncorrelated components in the sampling stage of the particle filter. In this publication we propose to consider the correlation between the individual spectral bins by correlating the random variation after a predicted walk realized by a linear prediction matrix.

Experiments on artificially added dynamic noise at different signal to noise ratios as well as on actual recordings with different speaker to microphone distances show reasonable word error rate reduction before and after acoustic model adaptation of the automatic speech recognition system.

Index Terms— speech feature enhancement, particle filter, correlation between spectral bins, automatic speech recognition

1. INTRODUCTION

Speech feature enhancement can be formulated as a tracking problem where the clean speech features $\mathbf{x}_k$ have to be estimated, for each frame k, given the observation history of the noisy features $\mathbf{y}_{1:k}$. The clean and noisy features are related by the probabilistic relationship $p(\mathbf{x}_k|\mathbf{y}_{1:k})$. As stated in Julier and Uhlmann [1] the minimum mean square error solution to such a tracking problem consists in finding the conditional mean $\mathrm{E}[\mathbf{x}_{1:k}|\mathbf{y}_{1:k}]$. Assuming that $(\mathbf{x}_k)_{k\in\mathbb{N}}$ is a Markov process and that the current observation is only dependent on the current state, facilitates sequential calculation of the conditional mean, the solution is given by

$$\mathrm{E}[\mathbf{x}_k|\mathbf{y}_{1:k}] = \int \mathbf{x}_k p(\mathbf{x}_k|\mathbf{y}_{1:k}) d\mathbf{x}. \tag{1}$$

Introducing the noise $\mathbf{n}_k$ as a hidden variable

$$p(\mathbf{x}_k|\mathbf{y}_{1:k}) = \int p(\mathbf{x}_k, \mathbf{n}_k|\mathbf{y}_{1:k}) d\mathbf{n}_k$$

with the relation $p(\mathbf{x}_k, \mathbf{n}_k|\mathbf{y}_{1:k}) = p(\mathbf{x}_k|\mathbf{y}_{1:k}, \mathbf{n}_k) p(\mathbf{n}_k|\mathbf{y}_{1:k})$ and a changed integration order we obtain

$$\mathrm{E}[\mathbf{x}_k|\mathbf{y}_{1:k}] = \int \underbrace{\int \mathbf{x}_k p(\mathbf{x}_k|\mathbf{y}_{1:k}, \mathbf{n}_k) d\mathbf{x}_k}_{=h_k(\mathbf{n}_k)} p(\mathbf{n}_k|\mathbf{y}_{1:k}) d\mathbf{n} \tag{2}$$

which can be approximated by Monte Carlo integration (details are provided in [2]):

$$\mathrm{E}[\mathbf{x}_k|\mathbf{y}_{1:k}] \approx \sum_{j=1}^{M} \tilde{\omega}_k^{(j)} h_k(\mathbf{n}_k^{(j)}) \tag{3}$$

To solve for (3) requires the evolution of noise modeled by weighted $\tilde{\omega}_k^{(j)}$ random variations $j = 1, 2, \cdots, M$. In the past, the individual bins of the feature vector have been assumed to be independent and identically-distributed (i.i.d.) from each other, ignoring the fact that neighbored spectral bins (in particular after spectral envelope processing or partly overlapping filterbanks) are correlated. Therefore, in this publication we propose to account for the correlation between spectral bins by correlating the random variation after a predicted walk, which in our case is realized by a linear prediction matrix.

2. BRIEF REVIEW OF SPEECH FEATURE ENHANCEMENT BY PARTICLE FILTERS

Different approaches to speech feature enhancement by particle filters exist. We follow Singh and Raj [3] who have proposed to track the noise frame by frame in the logarithmic spectral domain and later on subtract the noise estimates from the contaminated speech signal. An extended algorithm of the original approach as stated by Singh and Raj can be outlined as follows:

1. *Draw noise samples*
   At the start frame $k = 0$, M particles (noise hypotheses) $\mathbf{n}_0^{(j)}$ $(j = 1, ..., M)$ are drawn from the prior noise density $p_{\mathrm{noise}}(\mathbf{n}_0)$ which in our case is estimated on speech absent regions detected by voice activity detection.

   For frames $k > 0$, M particles $\mathbf{n}_k^{(j)}$ are sampled from the noise transition probability $p(\mathbf{n}_k|\mathbf{n}_{k-1})$ which has been estimated on speech absent regions.

   The evolution of noise spectra and different sampling techniques will be laid out in more detail in Section 3.

2. *Evaluate noise samples*
   The normalized importance weights are calculated as

$$\tilde{\omega}_k^{(j)} = \frac{p(\mathbf{y}_k|\mathbf{n}_k^{(j)})}{\sum_{m=1}^{M} p(\mathbf{y}_k|\mathbf{n}_k^{(m)})}$$

where the importance weight for each particle $\mathbf{n}_k^{(j)}$ is evaluated according to the likelihood

$$p(\mathbf{y}_k|\mathbf{n}_k^{(j)}) = \frac{p_{\text{speech}}(\mathbf{y}_k + \log(\mathbf{1} - e^{\mathbf{n}_k^{(j)} - \mathbf{y}_k})}{\prod_{i=1}^{d} \left|1 - e^{\hat{\mathbf{n}}_{k,i}^{(j)} - \mathbf{y}_{k,i}}\right|} \quad (4)$$

if $n_{k,i}^{(j)} < y_{k,i}\ \forall$ dimensions i. Otherwise $p(\mathbf{y}_k|\mathbf{n}_k^{(j)})$ can't be evaluated nor has a physical meaning and thus has to be rejected by setting the particle weight to zero. This causes a decimation of the particle population which can be remedied by a *fast acceptance test* [4] that virtually boosts the number of particles by redrawing samples in case of rejection.

To account for the dynamics of speech we have proposed to replace the evaluation of the particle weights based on a general speech model p_{speech} by a phoneme-specific model p_{phoneme} where the phoneme alignment is obtained by a previous pass of a speech recognition system [5]. Thus, a coupling between the particle filter and the speech recognition system is established whereas the two components have been treated as independent components in the past.

3. *Compensate for noise estimates*
Clean speech spectra can be estimated by using the discrete Monte Carlo representation of the continuous filtering density $p(\mathbf{n}_k|\mathbf{y}_{1:k})$, the so called *weighted empirical density*

$$\tilde{p}(\mathbf{n}_k|\mathbf{y}_{1:k}) = \sum_{j=1}^{M} \tilde{\omega}_k^{(j)} \delta_{\mathbf{n}_k^{(j)}}(\mathbf{n}_k) \quad (5)$$

The term $\delta_{\mathbf{n}_k^{(j)}}$ denotes a translated Dirac delta function.

Two different methods to compensate for the noise estimates will be presented in Section 4.

4. *Resample noise*
The normalized weights are used to resample among the noise hypotheses $\mathbf{n}_k^{(j)}$ $(j = 1, ..., M)$. This can be regarded as a pruning step where likely hypotheses are multiplied and unlikely ones are removed from the population.

Those steps are repeated with $k \mapsto (k+1)$ until all time-frames are processed.

Working Domain

Particle filters for speech feature enhancements are typically applied in the logarithmic spectral domain after dimension reduction by mel-filterbanks. Due to the properties of the used spectral estimation method provided by warped minimum variance distortionless response [6], no filterbank is applied and thus the dimension in the logarithmic spectral domain is not reduced. As the operation of a particle filter with high dimensions (in our case 129) would be infeasible or very slow, we decided to work in the logarithmic spectral domain after cepstral truncation to 20 dimensions by applying an inverse Fourier transformation to the cepstral coefficients. In the *truncated* logarithmic spectral domain the relation between the noisy observation $\mathbf{y}$, the clean feature $\mathbf{x}$ and noise $\mathbf{n}$ can be approximated by

$$\mathbf{x} \approx \log(e^{\mathbf{y}} - e^{\mathbf{n}}) = \mathbf{y} + \log(1 - e^{\mathbf{n}-\mathbf{y}}). \quad (6)$$

3. EVOLUTION OF THE NOISE SPECTRA (SAMPLING)

The used particle filter tracking application requires the prediction of the noise $\hat{\mathbf{n}}_k = p(\mathbf{n}_k|\mathbf{n}_{0:k-1})$ given the trajectory of the noise up to time k. The noise transition probability $p(\mathbf{n}_k|\mathbf{n}_{0:k-1})$ can be modeled by a dynamic system model, which can be classified into *random walk* and *predicted walk*.

3.1. Random Walk

The simplest way to model the evolution of noise features is a random walk

$$\hat{\mathbf{n}_k} = \mathbf{n}_{k-1} + \varepsilon_k$$

where $\mathbf{n}_k$ denotes the noise spectrum at time k, while the ε_k terms are considered to be i.i.d. zero mean Gaussian, i.e. $\varepsilon_k \sim \mathcal{N}(0, \Sigma_{noise})$, where the covariance matrix Σ_{noise} is assumed to be Gaussian.

3.2. Predicted Walk

To consider information about the evolution of the noise, Raj *et al.* [7] proposed and investigated to use a mth-order autoregressive process, $\mathbf{A}_{1:m}$, to predict the evolution of the noise

$$\begin{aligned} \hat{\mathbf{n}}_k &= \mathbf{A}_1\mathbf{n}_{k-1} + \mathbf{A}_2\mathbf{n}_{k-2} + \cdots + \mathbf{A}_m\mathbf{n}_{k-m} + \varepsilon_k \\ &= \mathbf{A}_{1:m}\mathbf{n}_{k-1:k-m} + \varepsilon_k. \end{aligned}$$

Learning the Autoregressive Noise Model

The autoregressive noise model consists of two components that have to be learned for a specific type of noise:

- the *linear prediction transition matrix* $\mathbf{A}_{1:m}$ and
- the *covariance matrix* $\mathbf{\Sigma}_{\text{noise}}$ where once again the ε_k terms are considered to be i.i.d. zero mean Gaussians.

Minimization of the prediction error norm results in the following estimate of the linear prediction matrix:

$$\mathbf{A}_{1:m} = \mathbb{E}[\mathbf{n}_k\mathbf{N}_{k-1:k-m}^T]\mathbb{E}[\mathbf{n}_{k-1:k-m}\mathbf{N}_{k-1:k-m}^T]^{-1} \quad (7)$$

Those matrices can be derived from the noise data $1, 2, \ldots, K$ as

$$\mathbb{E}[\mathbf{n}_k\mathbf{N}_{k-1:k-m}^K] = \frac{1}{K}\sum_{k=l}^{K} \mathbf{n}_k\mathbf{N}_{k-1:k-m}^T$$

and

$$\mathbb{E}[\mathbf{n}_{k-1:k-m}\mathbf{N}_{k-1:k-m}^T] = \frac{1}{K}\sum_{k=l}^{K} \mathbf{n}_{k-1:k-m}\mathbf{N}_{k-1:k-m}^T.$$

Note that it is sufficient to estimate the matrices from pieces of noise as long as the pieces are long enough to contain enough history, e.g. we have used noise only pieces between speech regions found by voice activity detection.

To learn a linear prediction matrix of model order m requires d^2m coefficients to be reliably estimated, which can only be established if a huge amount of training data is available. For a reasonable amount of training data only a small reduction in the mean square error can be reached by using higher order models [7]. Thus, a first model order is sufficient for our investigations.

The diagonal covariance can be learned by

$$\sigma_i^2 := \mathbb{E}[(n_{k,i} - \hat{n}_{k,i})^2],$$

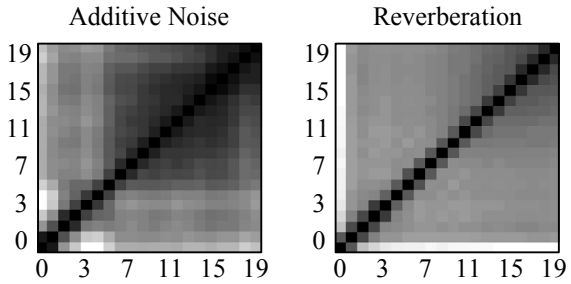

Fig. 1. Correlation matrices of $\Delta\mathbf{n}$ for additive dynamic noise and reverberation. The additive noise shows high correlation over a broad number of bins, while reverberation is less correlated and mostly limited to neighbored bins.

where $n_{k,i}$ denotes the ith vector component of the noise $\mathbf{n}_k$ and $\hat{n}_{k,i}$ denotes the ith vector component of the predicted noise $\hat{\mathbf{n}}_k = \mathbf{A}_{1:m}\mathbf{N}_{k-1:k-m}$. However, in practice, we have yielded better results by manually increasing the variance with identical values over all dimensions. This is probably due to an enlarged search space.

3.3. Predicted Walk with Correlation

In this section we present how to estimate the correlation in the random process and how to apply correlation between the spectral bins into the sampling stage of the particle filter. As the random process represents only the difference between the true noise and predicted noise, we start by

$$\Delta\mathbf{n} = \mathbf{n}_k - \mathbf{A}_{1:m}\mathbf{n}_{k-1:k-m} + \varepsilon_k.$$

The covariance matrix of the random process is then

$$\Sigma_{\Delta\mathbf{n}} = (\Delta\mathbf{n} - \mu_{\Delta\mathbf{n}})(\Delta\mathbf{n} - \mu_{\Delta\mathbf{n}})^T$$

where the mean values are given by

$$\mu_{\Delta\mathbf{n}} = \sum_{m}^{M} \Delta\mathbf{n}_m.$$

Normalizing the single bins of $\Sigma_{\Delta\mathbf{n}}$ by their variances $\sigma^2_{\Delta\mathbf{n}}$ the correlation matrix can now be calculated

$$\mathbf{Corr}(x,y)_{\Delta\mathbf{n}} = \frac{\Sigma(x,y)_{\Delta\mathbf{n}}}{\sigma(x)_{\Delta\mathbf{n}}\sigma(y)_{\Delta\mathbf{n}}}.$$

Given the Cholesky decomposition matrix by solving for

$$\mathbf{Chol}^T_{\Delta\mathbf{n}}\mathbf{Chol}_{\Delta\mathbf{n}} = \mathbf{Corr}_{\Delta\mathbf{n}}$$

correlated noise samples can be drawn from the uncorrelated noise samples, where the vector ε_k is identical to the one used in Section 3.2, by

$$\varepsilon_k^{\text{corr}} = \mathbf{Chol}_{\Delta\mathbf{n}}\varepsilon_k.$$

Figure 1 shows two correlation matrices. The first matrix is calculated on dynamic noise while the second matrix is calculated on reverberant data. On dynamic noise the correlation reaches over a broad number of bins while for reverberation the correlation is mainly limited to neighbored regions.

4. NOISE COMPENSATION

In this section we show two possible ways to evaluate for $h_k(\mathbf{n}_k^{(j)})$, the integral defined in (2).

4.1. The Vector Taylor Series Approach

To solve for the non-linear relation $\mathbf{y} = log(\mathbf{1}+e^{\mathbf{n}_k-\mathbf{x}_k})$ it has been proposed by Moreno *et al.* [8] to use a 0th order *vector Taylor series* (VTS) expansion around the mth Gaussian's mean μ_m.

$$\begin{aligned} h_k^{\text{VTS}}(\mathbf{n}_k) &= \sum_{m=1}^{M} p(m|\mathbf{y}_{1:k},\mathbf{n}_k) \\ &\quad \cdot \int x\delta_{\mathbf{y}_k - log(1+e^{\mathbf{n}_k-\mu_m})}(\mathbf{x}_k)d\mathbf{x}_k \\ &= \sum_{m=1}^{M} p(m|\mathbf{y}_{1:k},\mathbf{n}_k)\left(\mathbf{y}_k - log(1+e^{\mathbf{n}_k-\mu_m})\right) \\ &= \mathbf{y}_k - \sum_{m=1}^{M} p(m|\mathbf{y}_{1:k},\mathbf{n}_k) log(1+e^{\mathbf{n}_k-\mu_m}) \quad (8) \end{aligned}$$

4.2. The Statistical Inference Approach

In Monte Carlo sampling it is only required to consider point observations while the distribution is implicitly contained. Thus, it is possible to use the relationship between $\mathbf{x}_k$, $\mathbf{n}_k$ and $\mathbf{y}_k$ from (6) without the need for approximation [4] and we get the deterministic probability density

$$p(\mathbf{x}_k|\mathbf{y}_{1:k},\mathbf{n}_k) = \delta_{\mathbf{y}_k + log(\mathbf{1}-e^{\mathbf{n}_k-\mathbf{y}_k})}(\mathbf{x}_k).$$

By the substitution of $p(\mathbf{x}_k|\mathbf{y}_{1:k},\mathbf{n}_k)$ in $h_k(\mathbf{n}_k)$ we get the so called *statistical inference approach* (SIA)

$$\begin{aligned} h_k^{\text{SIA}}(\mathbf{n}_k) &= \int \mathbf{x}_k\delta_{\mathbf{y}_k + log(\mathbf{1}-e^{\mathbf{n}_k-\mathbf{y}_k})}(\mathbf{x}_k)d\mathbf{x}_k \\ &= \mathbf{y}_k + log(\mathbf{1} - e^{\mathbf{n}_k-\mathbf{y}_k}) \quad (9) \end{aligned}$$

which can be regarded as spectral subtraction in the logarithmic power domain (for one noise hypothesis).

5. EXPERIMENTS

In order to evaluate the performance of the proposed particle filter enhancements under realistic conditions we have recorded 35 minutes of lecture speech with different microphone types and speaker to microphone distances (similar to RT-06s development and evaluation data [9]) and added dynamic noise with different *signal to noise ratio* (SNR) to the close talk condition. As a speech recognition engine we used the *Janus Recognition Toolkit* (JRTk) with the same setup as described in [10]: The acoustic training material, approximately 100 hours, used for the experiments reported here, was taken from the ICSI, NIST, and CMU meeting corpora, as well as the *Translanguage English Database* (TED) and CHIL lecture corpora resulting in a discriminatively trained semi-continuous quint phone systems that contain 16000 distributions over 4000 codebooks, with a maximum of 64 Gaussians per model. The 3-gram language model contains approximately 23,000 words and has a perplexity of 125 on the test corpora. The used warped minimum variance distortionless response cepstral coefficients [6] have been shown to outperform mel

SNR		20 dB		15 dB		10 dB		5 dB	
Pass		1	2	1	2	1	2	1	2
Enhancement	Sampling	Word Error Rate							
none	—	20.7%	13.3%	29.3%	17.0%	43.4%	26.0%	61.6%	41.6%
VTS	uncorrelated	19.4%	13.2%	26.4%	16.7%	39.5%	23.8%	56.5%	39.4%
SIA	uncorrelated	20.1%	12.9%	27.7%	16.5%	41.4%	24.5%	57.6%	39.0%
VTS	correlated	19.2%	13.2%	26.5%	15.9%	38.6%	23.4%	57.0%	38.0%
SIA	correlated	19.8%	12.6%	27.6%	16.8%	40.3%	24.0%	58.1%	39.5%

Table 1. Word error rates for various particle filter approaches with additive noise at different signal to noise ratios (SNR)s.

Microphone		CTM		Lapel		Table Top		Wall	
Distance		5 cm		20 cm		100–150 cm		300–350 cm	
SNR		24 dB		23 dB		17 dB		10 dB	
Pass		1	2	1	2	1	2	1	2
Compensation	Sampling	Word Error Rate							
none	—	11.6%	09.8%	11.7%	09.9%	19.0%	14.6%	45.6%	29.0%
VTS	uncorrelated	11.3%	09.6%	11.7%	10.0%	18.7%	14.0%	44.6%	27.4%
SIA	uncorrelated	11.6%	09.3%	11.6%	10.2%	19.3%	14.0%	43.5%	26.5%
VTS	correlated	11.0%	09.7%	11.8%	10.0%	19.0%	13.9%	43.5%	26.8%
SIA	correlated	11.3%	09.5%	11.8%	09.9%	19.0%	14.2%	42.4%	25.1%

Table 2. Word error rates for various particle filter approaches at different speaker to microphone distances.

frequency cepstral coefficients [5] in combination with and without speech feature enhancement.

We evaluated on unadapted (first pass) acoustic models and acoustic models (second pass) which have been unsupervised adapted by *maximum likelihood linear regression* (MLLR), constrained MLLR and *vocal track length normalization* (VTLN). The determined VTLN factors have also been used in the second pass of the particle filter where furthermore the general representation of clean speech has been replaced by a phoneme dependent representation which has been aligned on the previous recognition pass [5].

Table 1 presents results on additive noise experiments and Table 2 presents results on actual recodings with different speaker to microphone distances. The results for various SNR and speaker to microphone distances are mixed. Not surprisingly, for SNR above 20 dB the particle filter, correlated or uncorrelated, can not improve the performance significantly as the signal is already clean. However, at SNR below 20 dB, the particle filter is able to show good performance improvements. At reasonable SNR values, around 10 dB, the particle filter with correlation shows good improvements in performance over the particle filter without correlation. Those improvements are, however, not consistent for other SNR values.

6. CONCLUSIONS

Even though the improvements due to correlation are somewhat intermingled, they show in average small improvements which are in particular significant for SNR values around 10 dB, on additive dynamic noise and also on distant recordings. Thus, we feel it is worthwhile using the predicted walk with correlated sampling as the improvements are established with a limited increase of computation by just one additional matrix multiplication per particle.

7. REFERENCES

[1] S. Julier and J.K. Uhlmann, "A general method for approximating nonlinear transformations of probability distributions," *tech. rep., RRG, Dept. of Engineering Science, University of Oxford*, Nov. 1996.

[2] C.P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer Texts in Statistics. Springer, second edition, 2004.

[3] R. Singh and B. Raj, "Tracking noise via dynamical systems with a continuum of states," *Proc. of ICASSP*, 2003.

[4] F. Faubel and M. Wölfel, "Overcoming the vector tailor series approximation in speech feature enhancement – a particle filter approach," *Proc. of ICASSP*, 2007.

[5] F. Faubel and M. Wölfel, "Coupling particle filters with automatic speech recognition for speech feature enhancement," *Proc. of Interspeech*, Sep. 2006.

[6] M. Wölfel and J.W. McDonough, "Minimum variance distortionless response spectral estimation, review and refinements," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 117–126, Sept. 2005.

[7] B. Raj, R. Singh, and R. Stern, "On tracking noise with linear dynamical system models," *Proc. of ICASSP*, 2004.

[8] P.J. Moreno, B. Raj, and R.M. Stern, "A vector taylor series approach for environment-independent speech recognition," *Proc. of ICASSP*, 1996.

[9] J.G. Fiscus, J. Ajot, M. Michel, and J.S. Garofolo, "The rich transcription 2006 spring meeting recognition evaluation," *Machine Learning for Multimodal Interaction, S. Renals, S. Bengio, and J.G. Fiscus (Eds.), LNCS vol. 4299*, pp. pp. 309–322, 2006.

[10] M. Wölfel, S. Stüker, and F. Kraft, "The ISL RT-07 speech-to-text system," *In Proc. of the Rich Transcription 2007 Meeting Recognition Evaluation Workshop (RT-07), Baltimore, USA*, 2007.