A Comparison of Particle Filtering Variants for Speech Feature Enhancement

Reinhold Haeb-Umbach, Joerg Schmalenstroeer

Department of Communications Engineering
University of Paderborn, Germany
{haeb,schmalen}@nt.uni-paderborn.de

Abstract

This paper compares several particle filtering variants for speech feature enhancement in non-stationary noise environments. By analyzing the random processes of clean speech, noise and noisy speech, appropriate proposal densities are derived. The performances of the resulting particle filters, i.e. modified *Sampling-Importance-Resampling* (mod-SIR), auxiliary SIR and likelihood particle filter, are compared in terms of word accuracy achieved by the subsequent speech recognizer on the AURORA 2 database. It turns out that for the noises found in this database, noise compensation techniques that assume stationary noise work equally well.

1. Introduction

Robust speech recognition in noisy environments remains a widely unsolved problem, while at the same time being of great practical importance.

Here we concentrate on the problem of speech feature enhancement, where a speech signal, which is corrupted by additive noise, is processed in the feature extraction stage such that the resulting feature vectors are those of the undistorted, clean speech signal component. The problem can be considered as one of Bayesian parameter estimation and various attempts have been made to compute MMSE estimates of the speech features, e.g. [1, 2]. This results in fairly complex estimators since the noise distorts the speech features in a highly non-linear manner. Simpler estimators can be obtained by linearization, e.g. by Taylor series approximation [3] or statistical linear approximation [4]. While in most approaches noise is considered stationary, some explicitly address non-stationary noise. Among these are sequential EM [13], sequential MAP [2] and particle filtering [5, 12].

Typical noises, e.g. street, bar, subway, that corrupt speech signals have a large component that is relatively slow-varying [5]. This means that, even for non-stationary noise, the noise spectrum of the present frame can well be predicted from previous frames. The use of dynamical systems to represent noise in the context of speech recognition has probably first been proposed by Varga and Moore [6], who represent noise as the output of a hidden Markov Model. While a HMM is a dynamical system, where the state vector can assume a finite number of values, a continuous state space model has been proposed by Raj et al. [5, 7]. Since a dynamical model of the log spectra of clean speech would be very complex, the idea here was to consider the noise as the state variable to be tracked, whose observation is corrupted by speech. This leads to a simple state equation. However, the measurement equation is non-linear and the measurement noise, the clean speech signal, is non-Gaussian. They proposed to use a variant of particle filters for the resulting non-linear estimation problem. Particle filtering was also proposed in [12], where, however, a simple random walk model for the noise trajectory was used.

Particle filters are sequential Monte Carlo methods for non-linear and/or non-Gaussian Bayesian tracking problems. They have been successfully applied to a large variety of applications [8]. In the speech signal processing domain they have, among others, been successfully applied to speaker tracking, e.g. [9, 10]. The most critical issue in particle filter design is the choice of the proposal density, from which samples are drawn to obtain a discrete approximation of the searched after posterior density. In [5] the *Sampling-Importance-Resampling* (SIR) particle filter [8] is applied to speech feature enhancement, where the proposal density is the state transition density. This is a popular choice because of its simplicity.

By studying the properties of the contributing random processes, more suitable proposal densities can be found. In this paper we present the design of an SIR filter with modified resampling, an auxiliary SIR filter, and a likelihood particle filter for speech feature enhancement. We give recognition results on the AURORA 2 database and compare the performance with that obtained by the ETSI advanced frontend (ETSI AFE) and what is obtained if the noise is assumed stationary.

2. State Space Model

We are given a speech signal which is corrupted by additive noise:

$$\mathbf{Z}_k = \mathbf{S}_k + \mathbf{N}_k. \tag{1}$$

Here, $\mathbf{S}_k$, $\mathbf{N}_k$ and $\mathbf{Z}_k$ denote the vector of DFT coefficients of the k-th frame of the undistorted speech signal, the noise, and the noise-corrupted speech, respectively. Squaring (1) we obtain

$$|\mathbf{Z}_k|^2 = |\mathbf{S}_k|^2 + |\mathbf{N}_k|^2 + 2|\mathbf{S}_k| \cdot |\mathbf{N}_k| \cdot cos(\theta) \tag{2}$$

where θ denotes the angle between the two complex variables $\mathbf{S}_k$ and $\mathbf{N}_k$. In the following the phase term $2|X_k| \cdot |N_k| \cdot cos(\theta)$ is neglected, a commonly used approximation which simplifies density computations later-on. Omission of this term, however, is known to produce artifacts, e.g. in spectral subtraction [2, 11], and we will come back to this issue in section 3.1.

Applying a Mel filterbank and taking the (element-wise) logarithm we obtain

$$\mathbf{z}_k = \mathbf{s}_k + \log(1 + e^{\mathbf{n}_k - \mathbf{s}_k}) \tag{3}$$

where $\mathbf{s}_k$, $\mathbf{n}_k$ and $\mathbf{z}_k$ are D-dimensional log-spectral feature vectors arising from the clean speech component, the noise and the noisy speech, respectively.

A major difficulty in removing the noise from $\mathbf{z}_k$ is the non-linear nature of eq. (3). Various approximations have been proposed to cope with the non-linearity, e.g. Vector Taylor Series (VTS) [3] and statistical linear approximation [4].

The approach taken here is different: Rather than linearizing (3) we are going to devise a non-linear state-space filter to remove the noise from the noisy speech. Since a dynamical model of speech can be very complicated, we identify the noise as the state variable: $\mathbf{x}_k := \mathbf{n}_k$, which is "corrupted" by speech, as was proposed in [5].

In [5] it was shown that a wide variety of noise types can be well modeled by a first-order AR model. We employ the following state equation

$$\mathbf{n}_{k+1} = \mathbf{A} \cdot \mathbf{n}_k + \mathbf{w}_k, \tag{4}$$

where the process noise $\mathbf{w}_k$ is assumed to be a white Gaussian noise vector with mean $\boldsymbol{\mu}_\mathbf{w}$ and diagonal covariance $\boldsymbol{\Sigma}_\mathbf{w}$. Note that the $D \times D$ state matrix $\mathbf{A}$ is not diagonal since there is considerable correlation among the components of a log-spectral feature vector. The state matrix and the process noise parameters are estimated from a segment of training examples of noise. In informal experiments on AURORA 2 we observed that the model (4) achieved noticeably better recognition accuracy than a simple random walk model ($\mathbf{A} = \mathbf{I}$, the identity matrix in (4)).

While the state equation (4) is linear, the measurement equation (3) is not. The "measurement noise" $\mathbf{s}_k$, i.e. the clean speech feature vector, is assumed to be drawn from a Gaussian mixture model (GMM):

$$p(\mathbf{s}_k) = \sum_{m=1}^{M} c_m \cdot \mathcal{N}(\mathbf{s}_k; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \tag{5}$$

where the weights c_m, the mean vectors $\boldsymbol{\mu}_m$ and the covariances $\boldsymbol{\Sigma}_m$, which are assumed to be diagonal with diagonal entries $\sigma^2_{m,d}$, $d = 1, \ldots, D$, can be obtained from noise-free training data.

Since the measurement noise is non-Gaussian and the measurement equation is non-linear, a Kalman filter cannot be applied. Also, local linearization, as is done in an extended Kalman filter, is not very promising due to the multimodal nature of the measurement noise. In the following section we are going to devise different particle filters for this estimation problem.

3. Particle Filter Design

Particle filters perform sequential Monte Carlo estimation based on point mass ("particle") representations of probability densities. Unlike Kalman Filters, the a posteriori density $p(\mathbf{x}_k|\mathbf{z}_{1:k})$, of the state vector $\mathbf{x}_k$, given all previous measurements $\mathbf{z}_{1:k} := \mathbf{z}_1, \ldots, \mathbf{z}_k$, is described by samples $\mathbf{x}_k^i$, $i = 1, \ldots, N$, rather than by moments [8].

Since drawing samples from the posterior is usually impossible, samples are drawn from a so-called *importance* or *proposal density* $q(\mathbf{x}_k|\mathbf{x}_{k-1}^i, \mathbf{z}_k)$, which must have the same support as the posterior. Then a weight w_k^i can be computed for each particle:

$$w_k^i \propto w_{k-1}^i \frac{p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i)}{q(\mathbf{x}_k|\mathbf{x}_{k-1}^i, \mathbf{z}_k)} \tag{6}$$

(w_{k-1}^i: weight of i-th particle at pervious frame $k-1$) such that the posterior is approximated by a discrete probability mass function:

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) \approx \sum_{i=1}^{N} w_k^i \delta(\mathbf{x}_k - \mathbf{x}_k^i). \tag{7}$$

Probably the most critical issue in the design of a particle filter is the appropriate choice of the importance density. In the following we present different options.

3.1. SIR Filter

The most popular choice of importance density is the state transition density:

$$q(\mathbf{x}_k|\mathbf{x}_{k-1}^i, \mathbf{z}_k) = p(\mathbf{x}_k|\mathbf{x}_{k-1}^i), \tag{8}$$

which results in the so-called *Sampling-Importance-Resampling* (SIR) particle filter. This was also the choice in [5]. Note that in the SIR filter the state space is explored without any knowledge of the observation $\mathbf{z}_k$.

Using (8) in (6) results in

$$w_k^i \propto w_{k-1}^i p(\mathbf{z}_k|\mathbf{x}_k^i). \tag{9}$$

The SIR particle filter

$$[\{\mathbf{x}_k^i,\ i = 1, \ldots, N\}] = \mathit{SIR}[\{\mathbf{x}_{k-1}^i,\ i = 1, \ldots, N\}, \mathbf{z}_k] \tag{10}$$

is described by the following iteration scheme

1. *Sampling*: Draw $\mathbf{x}_k^i \sim p(\mathbf{x}_k|\mathbf{x}_{k-1}^i)$, $i = 1, \ldots, N$.
2. *Weight computation*: Calculate $\tilde{w}_k^i = p(\mathbf{z}_k|\mathbf{x}_k^i)$, $i = 1, \ldots, N$ and normalize: $w_k^i = \tilde{w}_k^i / \sum_j \tilde{w}_k^j$.
3. *Resampling*: Draw N samples with replacement from the approximate discrete representation of the posterior (7).

In light of eq. (5) the likelihood $p(\mathbf{z}_k|\mathbf{x}_k^i)$ can be computed as follows (remember: $\mathbf{x}_k = \mathbf{n}_k$):

$$p(\mathbf{z}_k|\mathbf{n}_k) = \sum_{m=1}^{M} c_m \cdot p(\mathbf{z}_k|\mathbf{n}_k, m) \tag{11}$$

where

$$p(\mathbf{z}_k|\mathbf{n}_k, m) = \prod_{d=1}^{D} p(z_{k,d}|n_{k,d}, m). \tag{12}$$

Here, $z_{k,d}$ and $n_{k,d}$ denote the d-th component of the vector $\mathbf{z}_k$ and $\mathbf{n}_k$, resp. Using (3) we find

$$p(z_{k,d}|n_{k,d}, m) = \frac{e^{-\left(\frac{\log(e^{z_{k,d}} - e^{n_{k,d}}) - \mu_{m,d}}{2\sigma^2_{m,d}}\right)^2}}{\sqrt{2\pi\sigma^2_{m,d}} \cdot |1 - e^{n_{k,d} - z_{k,d}}|} \tag{13}$$

if $n_{k,d} \leq z_{k,d}$, and $p(z_{k,d}|n_{k,d}, m) = 0$ else.

Figure 1 shows the likelihood (13) as a function of $n_{k,d}$ for different values of $z_{k,d}$. Its impulse-like shape causes problems in the SIR recursions: particles $\mathbf{n}_k^i$ which are in the vicinity, but smaller than $\mathbf{z}_k$ (in every component) will be given a large weight according to (9). As a consequence they will be drawn several times during resampling. In the sampling stage they move according to the state equation, and it is likely that at least one component d of a particle will move such that $n_{k,d} > z_{k,d}$. As a result it will be assigned a weight of zero in the next iteration. This massive die out of particles was so severe that the SIR particle filter failed to work properly.

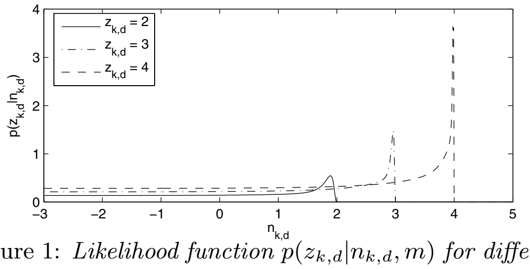

Figure 1: *Likelihood function* $p(z_{k,d}|n_{k,d}, m)$ *for different values of the state variable* $n_{k,d}$.

The problem can be overcome by either of the following means

i) The phase term in (2) should not be neglected. Then, however, a closed form for the likelihood function $p(\mathbf{z}_k|\mathbf{n}_k)$ can no longer be computed, even if the phase term is approximated by a normally distributed random variable. Numerical evaluation of the resulting integrals tend to be very complicated [2]. We therefore did not consider this approach any further.

ii) The resampling stage in the SIR iteration is modified such that each dimension of the particle is drawn independently. A "resampled" particle is obtained from the D draws of scalars. A weight of zero then only means that a component is missing and not the whole particle.

iii) Use an importance density which depends on the measurement $\mathbf{z}_k$. The optimal proposal density (in the sense of minimizing the variance of the weights) is $q(\mathbf{x}_k|\mathbf{x}_{k-1}^i, \mathbf{z}_k) = p(\mathbf{x}_k|\mathbf{x}_{k-1}^i, \mathbf{z}_k)$ [8], which, however, is not computable in our case. Since in our setup the likelihood is peaked, it is far tighter than the state transition density. Therefore it is important that we choose an importance density which depends on the current observation $\mathbf{z}_k$. In the following we present two designs that take this consideration into account.

3.2. Auxiliary SIR Filter

The basic idea of the Auxiliary SIR (ASIR) filter is to perform the resampling step at time $k-1$ using the available measurement at time k, before the particles are propagated to time k. In this way the ASIR filter attempts to mimic the sequence of steps carried out if the optimal importance density were available [8].

Since at the "previous" time step $k-1$ the particle $\mathbf{x}_k^i$ is not yet available, an estimate is required. In our case we used $\hat{\mathbf{x}}_k^i = \mathbf{A}\hat{\mathbf{x}}_{k-1}^i$. The error introduced by this is later compensated in the weight computation.

3.3. Likelihood Particle Filter

The likelihood particle filter uses an importance density which solely depends on the current observation

$$q(\mathbf{x}_k|\mathbf{x}_{k-1}^i, \mathbf{z}_k) = p(\mathbf{x}_k|\mathbf{z}_k). \tag{14}$$

In our setup, drawing from this density amounts to the following operation: first a sample $\mathbf{s}_k^i$ is drawn from the GMM of speech (5). Using this and the observation $\mathbf{z}_k$, the particle $\mathbf{x}_k^i = \mathbf{n}_k^i$ can be computed from the measurement equation (3). Employing (14) in (6) we see that the weights now depend on the state transition density.

4. Noise Compensation

The particle filters estimate the posterior density $p(\mathbf{x}_k|\mathbf{z}_{1:k})$. Point estimates can be derived from this. The approximate MMSE estimate of the system state, i.e. of the noise term, is obtained by:

$$\hat{\mathbf{n}}_k = \sum_{i=1}^{N} w_k^i \mathbf{x}_k^i. \tag{15}$$

From the estimate of the noise and the measurement $\mathbf{z}_k$ we can estimate $\mathbf{s}_k$, the log spectrum of the clean speech [3]:

$$\hat{\mathbf{s}}_k = \mathbf{z}_k - \sum_{m=0}^{M-1} p(m|\mathbf{z}_k, \hat{\mathbf{n}}_k) \log\left(1 + e^{\hat{\mathbf{n}}_k - \boldsymbol{\mu}_m}\right). \tag{16}$$

In the experiments we observed that $\hat{\mathbf{n}}_k$ tends to be biased. We therefore replaced $\hat{\mathbf{n}}_k$ by $\gamma\hat{\mathbf{n}}_k$ which resulted slightly improved performance. Here, γ is an "undersubtraction" factor, which has been determined on training examples of noise to $\gamma = 0.96$.

An alternative to (16) would be to first compute clean speech "particles" $\hat{\mathbf{s}}_k^i$, $i = 1, \ldots, N$, by using $\mathbf{n}_k^i$ instead of $\hat{\mathbf{n}}_k$ in (16). Then $\hat{\mathbf{s}}_k$ is obtained as the average of the clean speech particles [5]. Recognition experiments showed that this delivers only marginal improvement at greatly increased computational effort. We therefore used (16) in the experiments reported in the next section.

5. Experimental Results

A series of recognition experiments was run on the AURORA 2 database, which was chosen because it contains a variety of noise types and signal-to-noise ratio (SNR) conditions. Baseline results were obtained by using the ETSI standard front end (ES 201 108) and the ETSI advanced front end (ES 202 050), both with clean condition training.

Further, the Vector Taylor Series (VTS) algorithm [3] was used as a benchmark. In VTS the non-linearity (3) is approximated by a Taylor series, for the results reported here by a 0-th order approximation, and an estimate of the noise $\mathbf{n}_k$ is obtained which is used in (16) to remove the noise from the log-spectral feature vectors. The 0-th order VTS is computationally quite inexpensive and is known to achieve good performance.

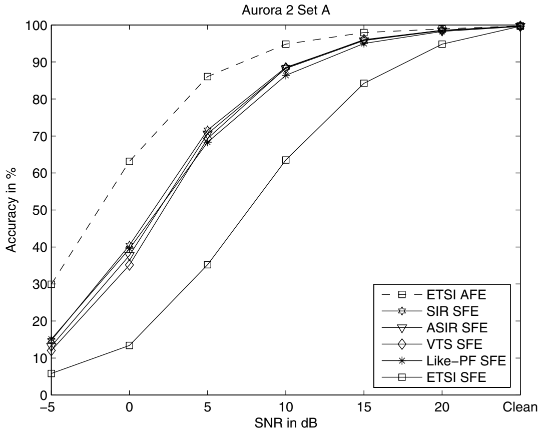

Figure 2: *Word accuracy on testset A of AURORA 2 as a function of SNR.*

Figures 2 and 3 show word accuracies as a function of SNR for testsets A and B of the AURORA 2 database, respectively. The figures display results for speech feature enhancement by particle filtering with an otherwise unmodified ETSI standard frontend (SFE) for three particle filtering variants: the SIR filter with modified resampling (SIR SFE), as explained in Section 3.1, the auxiliary filter (ASIR SFE), and the likelihood particle filter (Like-PF SFE). All particle filters achieved about the same word accuracy. However it was observed that the SIR particle filter was the least sensitive to a reduction of the number of particles. While the performance of the other filters quickly degraded when the number of particles was reduced from $N = 200$, the SIR particle filter maintained roughly its performance even for $N = 100$.

The figures also show that, as an average over all noise types of testsets A and B, the VTS algorithm, which assumes stationary noise over the whole utterance, achieves equally good results, at, however, greatly reduced computational complexity. Actually for no noise type a clear advantage of the particle filter approach was observed. While others reported improved performance on highly non-stationary noise [12], the noise types of AURORA 2 seem to be good-natured enough to be treated by stationary noise compensation techniques.

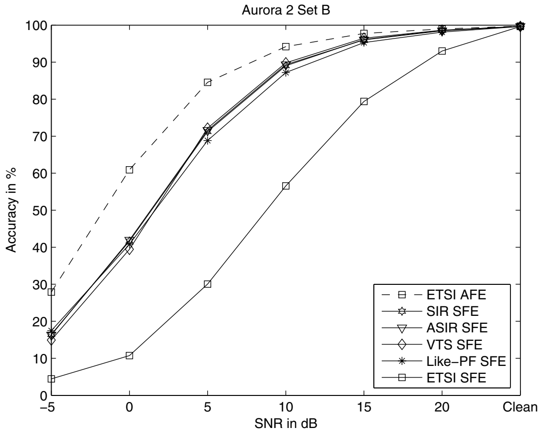

Figure 3: *Word accuracy on testset B of AURORA 2 as a function of SNR.*

6. Conclusions

By investigating the random process of noise corrupted speech features we were able to find appropriate proposal densities to be used in particle filter based speech feature enhancement. While significant improvements over the ETSI standard frontend were achieved, we were not able to reach the performance of the ETSI advanced frontend. Further, speech feature enhancement by Vector Taylor Series approximation, which assumes stationary noise, achieved comparable results. It can be concluded that for the noise types present in AURORA 2 the assumption of stationarity over the duration of an utterance holds and particle filtering, which explicitly addresses non-stationary noise, can also cope with stationary noise, but yields not additional performance advantage.

7. Acknowledgments

This work has in part been supported by the European union project ”Amigo - Ambient intelligence for the networked home environment”. More informations on ”www.amigo-project.org”

8. References

[1] Deng, L., Droppo, J. and Acero, A., ”A Bayesian Approach to Speech Feature Enhancement Using the Dynamic Cepstral Prior”, in *Proc. ICASSP*, Salt Lake City, 2002.

[2] Deng. L., Droppo, J., and Acero, A., ”Enhancement of Log Mel Power Spectra of Speech Using a Phase-Sensitive Model of the Acoustic Environment and Sequential Estimation of the Corrupting Noise”, *IEEE T-SAP*, vol. 12, no. 2, pp. 133-143, March 2004.

[3] Moreno, P., Raj, B. and Stern, R., ”A Vector Taylor Series Approach for Environment-Independent Speech Recognition”, in *Proc. ICASSP*, Atlanta, 1996.

[4] Kim, N., ”Statistical Linear Approximation for Environment Compensation”, *IEEE Signal Processing Letters*, vol. 5, no. 1, pp. 8-10, Jan. 1998.

[5] Raj, B., Singh, R. and Stern, R., ”On Tracking Noise with Linear Dynamical System Models”, in *Proc. ICASSP*, Montreal, 2004.

[6] Varga, A.P. and Moore, R.K., ”Hidden markov Model Decomposition of Speech and Noise”, in *Proc. ICASSP*, 1990.

[7] Singh, R. and Raj, B., ”Tracking Noise via Dynamical Systems with a Continuum of States”, in *Proc. ICASSP*, Miami, 2003.

[8] Ristic, B., Arulampalam, S. and Gordon, N., *Beyond the Kalman Filter*, Artech House, 2004.

[9] Ward, D.B., Lehmann, E.H. and Williamson, R.C. ”Particle Filtering Algorithms for Tracking an Acoustic Source in a Reverberant Environment”, *IEEE Trans. on Speech and Audio Proc.*, vol. 11, pp. 826-836, 2003.

[10] Warsitz, E., Haeb-Umbach, R. and Peschke, S., ”Adaptive Beamforming Combined with Particle Filtering for Acoustic Source Localization”, in *Proc. ICSLP*, Korea, 2004.

[11] Zhu, Q. and Alween, A. ”The Effect of Additive Noise on Speech Amplitude Spectra: A Quantitative Analysis”, *IEEE Signal Processing Letters*, vol. 9, no. 9, pp. 275-277, Sept. 2002.

[12] Fujimoto M. and Nakamura S., ”Particle Filter Based Non-Stationary Tracking for Robust Speech Recognition”, in *Proc. ICASSP*, USA, 2005.

[13] Afify M. and Siohan O., ”Sequential Estimation with Optimal Forgeting for Robust Speech Recognition”, IEEE Trans. SAP, VOL 12, No.1, pp.19-26,2004