A JOINT PARTICLE FILTER AND MULTI-STEP LINEAR PREDICTION FRAMEWORK TO PROVIDE ENHANCED SPEECH FEATURES PRIOR TO AUTOMATIC RECOGNITION

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

ABSTRACT
Automatic speech recognition, which works reasonable well on recordings captured with mid- or far-field microphones, is essential for a natural verbal communication between humans and machines. While a lot of research and development is devoted to address one of the two distortions frequently encountered in mid- and far-field sound pickup, namely non-stationary noise or reverberation, less effort has been undertaken to jointly combat both kinds of distortions. In our view, however, this is essential to further reduce the demolishing effect by moving the microphone away from the speakers mouth. A first step into this direction is proposed in this paper by integrating an estimate of the reverberation derived by multi-step linear prediction into a framework, which so far tracks and removes non-stationary additive distortions by particle filters. Evaluations on actual recordings with different speaker to microphone distances demonstrate that techniques combating either non-stationary noise or reverberation can be combined for good effect.

Index Terms— speech feature enhancement, particle filter, multi-step linear prediction, joint denoising and dereverberation, automatic speech recognition

1. INTRODUCTION
While a lot of research in speech feature enhancement for automatic speech recognition has focused on compensating either stationary additive noise or reverberation with a stationary room impulse response, most of the observed distortions are non-stationary, additive and convolutive. Those non-stationary distortions, however, can neither be represented well under the stationary assumptions in the feature space by methods such as spectral subtraction [1] or feature space adaptation (FSA) [2] nor in the model space by adaptation techniques such as maximum likelihood linear regression (MLLR) [3] and are in fact one of the most significant problems in hands-free automatic speech recognition.

To cope with the non-stationary behavior of distortions, just recently, various particle filter approaches have appeared to track distortions on speech features in logarithmic spectral or cepstral domain [4, 5, 6]. Those algorithms cope well with non-stationary additive distortions, however, are not able to remedy convolutive distortions. Cepstral mean normalization (CMN), on the other hand, is able to track convolutive distortions, however, only those which are no longer than the observation window, typical ≤ 32 ms. Tashen et al. [7] as well as Petrick et al. [8] found that reverberation between 50 ms and RT 40 has the strongest distortional effect on the word accuracy. Thus, those distortions can not well be represented and compensated by CMN. To estimate this for those harmful late reflections it has been proposed by Kinoshita et al. [9] to use multi-step linear prediction (MSLP). The resulting frame-by-frame distortion estimate behaves more like non-stationary additive distortion in the power frequency domain and thus can be easily removed (without the need to estimate and to invert the impulse response) by spectral subtraction. This algorithm, however, focuses only on reverberation.

To compensate for additive distortions as well as late reflections it is possible to simply concatenate the different, previously described, processing steps. The full potential of speech feature enhancement, however, can only be reached by jointly estimating both kinds of distortions as the individual estimates are not independent to each other.

2. JOINT ESTIMATION AND COMPENSATION FRAMEWORK
In this section a generalized particle filter (PF) framework, which is capable of jointly tracking noise and reverberation on a frame-by-frame basis, is presented. The dimension of the PF which is capable to track additive distortions in the feature space is determined by the number of spectral bins. To jointly consider additive and reverberant distortions the dimensionality of the PF has to be extended. In the proposed framework the new dimensions do not represent the reverberation directly, but scaling terms of the reverberation estimate.

An overview of the joint PF framework is given in Figure 1. A corresponding outlined of the different components is presented in the following sections where the steps described in Section 2.4 until Section 2.9 are repeated with k ↦→ (k + 1) until all frames are processed or the track is lost and has to be reinitialized with the step described in Section 2.3.

2.1. Reverberation Estimation
In order to estimate the correlation in the speech signal Kinoshita et al. [9] have proposed to use MSLP [10]. In contrast to the well known linear prediction (LP), MSLP aims to predict a signal after a given delay D, the so called step-size. With the prediction error e[n] we can formulate MSLP as

\[ y[n] = \sum_{m=1}^{M} c_m y[n - m - D] + e[n] \]

where \( c_1, \ldots, c_m \) represent the LP coefficients, \( y[n] \) the observed signal and \( M \) the model order. The mean square error solution for the MSLP coefficients \( \mathbf{c} = [c_1, c_2, \ldots, c_M]^T \) is given by

\[ \mathbf{c} = \left( E \{ y[n - D] y[n - D]^T \} \right)^{-1} E \{ y[n - D] y[n]^T \} \]
which can be efficiently solved using the Levinson-Durbin recursion.

An estimate of the reflection sequence $r[n]$ can be obtained by filtering the observation sequence $y[n]$ with the MSLP filter

$$r[n] = \sum_{m=1}^{M} c_m y[n - m - M_{\text{early}} + 1]$$

where the delay $D$ has been set to $M_{\text{early}} = 60$ ms in our experiments. As proposed by Kinoshita et al. [9] the reflection sequence $r[n]$ can now be converted into short-time power spectra $r_k$. This distortion estimate, which is significantly changing for each frame $k$, can now be treated just like additive distortions and thus be easily removed from the distorted sequence $y_k$ by well known methods such as spectral subtraction [1].

As the reflection sequence $r_k$ might still contain some correlation due to the speech production filter, it has been suggested to use pre-whitening prior to the estimation of the MSLP coefficients [9]. In our experiments, however, we have not observed consistent gains and thus the pre-whitening filter has not been used for the experiments reported in this publication.

### 2.2. Spectral Estimation and Working Domain

The reverberant $r_k$ and distorted $y_k$ spectra have to be estimated for all frames, $k = 0; \ldots; K$, from $r[n]$ and $y[n]$ respectively. In order to prevent the PF to work in a very high dimensional space (in our case the spectra, 129 bins, has been estimated by warped minimum variance distortionless responses [11] without a dimension reduction by a filter bank) we decided to work in the logarithmic spectral domain after cepstral truncation to 20 dimensions. The truncated logarithmic spectra was calculated by an inverse discrete cosine transformation established by a simple $20 \times 20$ matrix multiplication. In this domain the relation between the noisy observation $y$, the clean feature $x$ and distortion $d$ can be approximated by

$$x = y + \ln(1 - e^{-d - y}) + e_\theta + e_{\text{envelope}} \approx y + \ln(1 - e^{-d - y}).$$

The first error term $e_\theta$ is due to a phase difference $\theta$ between $x$ and $d$. It is complicated to evaluate, however, has been empirically verified by Deng et al. [12] that the mean is close to zero and to be Gaussian distributed (at least in higher mel-scaled frequencies where the central limit theorem can be applied). The second error term $e_{\text{envelope}}$ is caused by spectral or cepstral envelope techniques. Both error terms are small enough that the approximation in (1) is sufficient.

To not suffer from the feature manipulation and lower dimensionality of the PF working domain with respect to the reverberation estimate, it is important to not simply process the reverberation estimate like the distorted observation but by a processing chain as shown in Figure 2.

![Diagram of the reverberation estimate in the logarithmic frequency domain](image)

**Fig. 1.** Schematics of joint particle filter estimation of additive and reverberant distortions. Solid arrows represent the flow of the signal/features. Dotted arrows represent the flow of particle information such as the particle weight and the particle values which represent either estimates of additive distortions or estimates of the scaling factors of the estimated reverberation.

**Fig. 2.** Diagram of the reverberation estimate in the logarithmic frequency domain. STSE stands for short time spectral analysis, DCT and IDCT for discrete cosine transformation and its inverse respectively and MSLP for multi-step linear prediction. The small numbers give the dimension of the feature stream.

### 2.3. Distortion Estimation & Particles Initialization

The first step in any PF framework is its initialization by drawing samples from the prior particle density. In our framework the prior particle density

$$p(p_0) = \begin{bmatrix} p(a_0) \\ p(s_0) \end{bmatrix}$$

is a concatenation of the prior additive distortion density $p(a_0)$ and the prior scale density $p(s_0)$ of the reverberation estimate. Unfortunately, the prior additive distortion density $p(a_0)$ can not be estimated directly. However, it can be decomposed into two densities which can be estimated:

- The **prior overall distortion density** $p(d)$ derived on silence regions of the input signal which contains additive and convolutive distortions and
- the **prior reverberation density** $p(r_0)$ which is estimated over all frames derived on the reflection sequence $r_k$. 
With the prior overall distortion density and the prior reverberation density it is now possible to derive the prior additive distortion density as
\[ p(a_0) = \ln (\exp p(d_0) - \exp p(r_0)) \].

The prior scale density \( p(s_0) \) is given by a Gaussian \( \mathcal{N}(1.0, \Sigma_s) \) with mean 1.0 and a small variance term \( \Sigma_s \), as we assume the reverberation energies in the spectra to be accurately estimated.

2.4. Particle Evolution

The evolution for each particle \( p_k^{(s)} \), \( s = 0, \ldots, S - 1 \), is estimated by an autoregressive process
\[ p_k^{(s)} = p_{k-1}^{(s)} + e_k^{(s)} ; e_k^{(s)} \sim \mathcal{N}(0, \sigma_e). \]

The estimate of the AR matrix \( P_{k-1} \) can be represented by a joint matrix. We yielded better results, however, by considering the additive distortion and the scale terms as independent components
\[ P_k = \begin{bmatrix} A_k & 0 \\ 0 & S_k \end{bmatrix} \]
where the additive distortion matrix \( A_k \) is recalculated for each frame \( k \) by the dynamic autoregressive process [13]. The scale terms \( s_k[b], b = 0, 1, \ldots, B-1 \) can either
- share a scaling factor
  \[ s[b] = p[B]; S_k = 1, \]
  adding one dimension to the PF,
- share a scaling factor and a tilt factor
  \[ s[b] = p[B] + p[B+1](b - (B + 1)/2); S_k = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \]
  which allows to scale lower and higher frequency bins of the spectral reverberation estimate differently, adding two dimensions to the PF, or
- scale individually for each bin
  \[ s[b] = p[B + b] \]
  which doubles the dimensions of the PF. While in the previous approaches a random walk is used to model the evolution of the scale term, here \( S_k \) can be either modeled as a random walk \( S_k = I \) or by a dynamic autoregressive process.

As an individual scaling of each bin significantly increases the search space and execution time and furthermore has not been able to outperform the alternative approaches with lower dimensionality, it will not be further investigated, however has been presented here for the sake of completeness.

2.5. Distortion Combination

The distortion samples \( n[b]^{(s)} \) are calculated for each particle \( p_k^{(s)} \), \( s = 0, \ldots, S - 1 \) and frequency bin \( b = 0, \ldots, B-1 \) as
\[ n[b]^{(s)} = \ln \left\{ \exp a[b]^{(s)} + s[b]^{(s)} \exp r[b]^{(s)} \right\} \]
where \( a[b] = p[b] \) represents additive distortions, \( s[b] \) represents the scale terms and \( r[b] \) represents the spectral distortion due to reverberation.

2.6. Distortion Evaluation

With the prior speech density \( p_{speech}(\cdot) \) each distortion sample \( d_k^{(s)} \) is evaluated according to the likelihood
\[ p(y_k | d_k^{(s)}) = \frac{p_{speech}(y_k + \ln(1 - e^{d_k^{(s)} - y_k}))}{\prod_{b=1}^{B} 1 - e^{d_k^{(s)} - y_{k,b}}} \]  
and its normalized weight
\[ w_k^{(s)} = \frac{p(y_k | d_k^{(s)})}{\sum_{m=1}^{S} p(y_k | d_m^{(m)})} \]
is calculated. Note that the likelihood can only be evaluated if
\[ d_k^{(s)} < y_{k,b} \forall b \in B, \]
otherwise the particle weight is set to zero. This causes a decimation of the particle population which we attenuate by the fast acceptance test [14]. It accepts a drawn sample only if the likelihood can be eval-uated and propagates a new sample from a randomly chosen particle otherwise.

2.7. Distortion Compensation

The clean feature is estimated with the distortion samples \( d_k^{(s)} \) and their corresponding important weights \( w_k^{(s)} \) over all particle samples \( S \) using the non-linear relationship between \( x_k, d_k \) and \( y_k \) as [14].
\[ E\{x_k|y_{1:k}\} = \sum_{s} w_k^{(s)} \left( y_k + \ln(1 - e^{d_k^{(s)} - y_k}) \right) \]

2.8. Importance Resampling

To prevent that the vast majority of the probability mass is concentrated in a few particles they are resampled after every step.

2.9. Prediction Model Estimation

The prediction model
\[ A_k = E\{a_k a_k^{T}\}E\{a_{k-1} a_{k-1}^{T}\}^{-1} \]
is estimated by the dynamic autoregressive process [13] where the expectation of the required matrices
\[ E\{a_k a_k^{T}\} = \frac{1}{S} \sum_{s=1}^{S} p(y_k | a_k^{(s)}) a_k a_k^{(s)\ T} \]
and
\[ E\{a_{k-1} a_{k-1}^{T}\} = \frac{1}{S} \sum_{s=1}^{S} p(y_k | a_k^{(s)}) a_{k-1} a_{k-1}^{(s)\ T} \]
are calculated by a weighed summation over all additive distortions \( a^{(s)}, s = 1, 2, \ldots, S \) due to their corresponding likelihoods (2).
can be applied without constraints to all microphone conditions. and lapel microphones due to multi-step linear prediction and thus joint approach is able to limit the performance reduction on close unsupervised adapted acoustic models. Furthermore, the proposed components and demonstrates significant gains for unadapted as well as approach is again superior to the independent treatment of the compen-

Compensating for both kinds of distortions leads to further improve-

reverberant distortions improves the accuracy, except for the com-

pass 1 2 1 2 1 2 1 2

Front-End Compensation Word Error Rate %

Additive Reverbation

power spectrum no no 11.3 9.5 12.3 10.3 18.0 14.2 45.9 30.0

warped MVDR no no 11.2 9.1 10.9 9.2 18.6 14.0 45.4 28.6

warped MVDR yes no 10.6 9.0 10.7 9.0 17.8 13.2 42.8 25.4

warped MVDR no yes 14.4 9.5 15.1 9.6 17.7 13.4 39.2 23.9

warped MVDR yes yes 12.1 9.3 13.4 9.5 17.7 13.3 38.3 23.3

warped MVDR joint 1 11.7 9.3 11.8 9.3 17.4 12.8 37.9 22.7

warped MVDR joint 2 11.5 8.6 11.9 9.0 16.9 12.6 38.4 22.2

Table 1. Speech recognition experiments on single channel recordings with different speaker to microphone distances.

3. EXPERIMENTS AND CONCLUSION

In order to evaluate the performance of the proposed algorithm under realistic conditions we have recorded and transcribed 35 minutes of lecture speech with different microphone types and speaker to microphone distances (similar to NIST’s RT-06s development and evaluation data [15]). As a speech recognition engine we used the Janus Recognition Toolkit (JRTk) with a configuration identical to the one used by our lab at NIST’s RT-07 evaluation campaign [16]. The three-gram language model consists of 25k words and has a perplexity of 125 on the test set.

We evaluated on unadapted (first pass) acoustic models and acoustic models (second pass) which have been unsupervised adapted by MLLR, FSA and vocal track length normalization (VTLN). The determined VTLN factors have also been used in the second pass of the particle filter.

Comparing the word error rates of the experiments presented in Table 1 demonstrates that individually compensating additive or reverberant distortions improves the accuracy, except for the compensation of reverberation on the CTM and the Lapel microphones. Compensating for both kinds of distortions leads to further improvements over a single compensation technique. The proposed joint approach is again superior to the independent treatment of the components and demonstrates significant gains for unadapted as well as unsupervised adapted acoustic models. Furthermore, the proposed joint approach is able to limit the performance reduction on close and lapel microphones due to multi-step linear prediction and thus can be applied without constraints to all microphone conditions.

4. REFERENCES


