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 well on recordings captured with mid- or far-field microphones, is essential for a natural verbal communication between humans and machines. While a great deal of research effort has addressed one of the two distortions frequently encountered in mid- and far-field sound capture, namely *non-stationary noise* and *reverberation*, much less work has undertaken to jointly combat both kinds of distortions. In our view, however, this joint approach is essential in order to further reduce catastrophic effects of noise and reverberation that are encountered as soon as the microphone is more than a few centimeters from the speaker's mouth. We propose here to integrate an estimate of the reverberation obtained by multi-step linear prediction into a particle filter framework that tracks and removes non-stationary additive distortions. 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 dereveberation, automatic speech recognition

1. INTRODUCTION

While a great deal 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]. Hence, such non-stationary distortions are in fact one of the most significant problems in hands-free automatic speech recognition.

To cope with just such non-stationary distortions, various algorithms based on the *particle filter* (PF) have recently been proposed to *track* distortions in speech features in logarithmic spectral or cepstral domains [4, 5, 6]. While such algorithms cope well with non-stationary additive distortions, they are not able to reduce or remove the effects of convolutive distortions. *Cepstral mean normalization* (CMN), on the other hand, is able to handle convolutive distortions, but only those which are no longer than an observation window, typically 32 ms or less. Tashev *et al.* [7] as well as Petrick *et al.* [8] found that *late reverberation* between 50 ms and the time when the power of the reverberated signal is 40 dB below its peak level has the most damaging effect on the word accuracy of a far-field automatic speech recognition system. Given that their duration is much longer than an analysis window, CMN cannot compensate for these distortions. To obtain an estimate for these harmful late reflections, Kinoshita *et al.* [9] proposed 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 through spectral subtraction without the need to explicitly estimate and invert the room impulse response. This algorithm, however, is effective only against 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 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 able 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 Sections 2.4 through 2.8 are repeated with $k \mapsto (k+1)$ 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

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

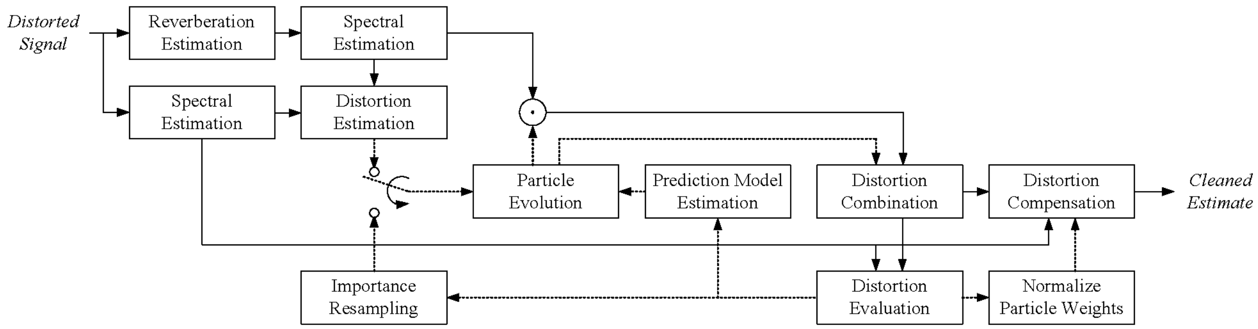

Fig. 1. Schematics of joint particle filter estimation of additive and reverberant distortions. Solid arrows represent the flow of the signal/features. Dotted arrows represent the flow of particle information such as the particle weight and the particle values which represent either estimates of additive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

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

$$\mathbf{c} = \left(E\left\{\mathbf{y}[n-D]\mathbf{y}[n-D]^T\right\}\right)^{-1} E\left\{\mathbf{y}[n-D]\mathbf{y}[n]^T\right\}$$

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

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

where the delay D has been set to 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 $\mathbf{r}_k$. This highly non-stationary distortion estimate can now be treated just like an additive distortion and removed from the distorted sequence $\mathbf{y}_k$ by well known methods, such as spectral subtraction [1].

As the reflection sequence $\mathbf{r}_k$ 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]. We have not observed consistent gains with this technique and thus the pre-whitening filter has not been used for the experiments reported here.

2.2. Spectral Estimation and Working Domain

The reverberant $\mathbf{r}_k$ and distorted $\mathbf{y}_k$ spectra have to be estimated for all frames, $k = 0; \cdots, K$, 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 respons [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 20×20 matrix multiplication. In this domain the relation between the noisy observation $\mathbf{y}$, the clean feature $\mathbf{x}$ and distortion $\mathbf{d}$ can be approximated by

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

The first error term

$$\mathbf{e}_\theta = \ln\left(1 + \frac{2\cos\theta(\Omega)}{\cosh\{\ln|D(\Omega)| - \ln|X(\Omega)|\}}\right)$$

is due to a phase difference θ between x and d. Deng *et al.* [12] have shown that this phase difference is zero-mean and Gaussian distributed, at least in higher mel-scaled frequencies where the central limit theorem holds. The second error term $\mathbf{e}_{\text{envelope}}$ is caused by the applied spectral and cepstral envelope techniques. Both error terms are small enough that the approximation in (1) is sufficient.

In order to avoid any detrimental effects from using a PF of lower dimensionality than that of the reverberation estimate, it is important to use the processing chain shown in Figure 2.

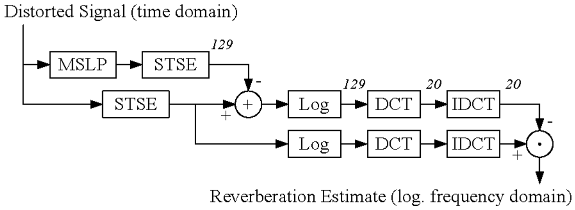
Distorted Signal (time domain)
MSLP
STSE
129
STSE
-
+
+
Log
129
DCT
20
IDCT
20
Log
DCT
IDCT
-
+
·
Reverberation Estimate (log. frequency domain)

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,

$$p(\mathbf{p}_0) = \begin{bmatrix} p(\mathbf{a}_0) \\ \cdots \\ p(\mathbf{s}_0) \end{bmatrix},$$

is a concatenation of the prior additive distortion density $p(\mathbf{a}_0)$ and the prior scale density $p(\mathbf{s}_0)$ of the reverberation estimate. Unfortunately, the prior additive distortion density $p(\mathbf{a}_0)$ can not be estimated directly. It can, however, be decomposed into two densities which can be estimated:

- The *prior overall distortion density* $p(\mathbf{d}_0)$ derived on silence regions of the input signal which contains *additive* and *convolutive* distortions and
- the *prior reverberation density* $p(\mathbf{r}_0)$ which is estimated over all frames derived on the reflection sequence $\mathbf{r}_k$.

With the prior overall distortion density and the prior reverberation density it is now possible to derive the *prior additive distortion density* as

$$p(\mathbf{a}_0) = \ln\left(\exp p(\mathbf{d}_0) - \exp p(\mathbf{r}_0)\right).$$

The *prior scale density* $p(\mathbf{s}_0)$ is given by a Gaussian $\mathcal{N}(1.0, \mathbf{\Sigma}_s)$ with mean 1.0 and a small variance term $\mathbf{\Sigma}_s$, as we assume the reverberation energies in the spectra to be accurately estimated.

2.4. Particle Evolution

The evolution for each particle $\mathbf{p}_k^{(s)}$, $s = 0, ..., S-1$, is estimated by an autoregressive process

$$\mathbf{p}_k^{(s)} = \mathbf{P}_{k-1}\mathbf{p}_{k-1}^{(s)} + \epsilon_k^{(s)};\ \epsilon_k^{(s)} \sim \mathcal{N}(0, \sigma_\epsilon^2).$$

The estimate of the AR matrix $\mathbf{P}_{k-1}$ can be represented as a joint matrix. We obtained better results, however, by considering the additive distortion and the scale terms as independent components, such that,

$$\mathbf{P}_k = \left[\begin{array}{c:c} \mathbf{A}_k & \mathbf{0} \\ \hdashline \mathbf{0} & \mathbf{S}_k \end{array}\right],$$

where the additive distortion matrix $\mathbf{A}_k$ is recalculated for each frame k by the dynamic autoregressive process [13]. The scale terms $s_k[b]$, $b = 0, 1, \cdots, B-1$ can either

- *share a scaling factor*

  $$s[b] = p[B];\ \mathbf{S}_k = 1,$$

  adding one dimension to the PF,
- *share a scaling factor and a tilt factor*

  $$s[b] = p[B] + p[B+1](b - (B+1)/2);\ \mathbf{S}_k = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

  which enables the lower and higher frequency bins of the spectral reverberation estimate to be scaled 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 terms, here $\mathbf{S}_k$ can be either modeled as a random walk $\mathbf{S}_k = \mathbf{I}$ 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 $d_k^{(s)}[b]$ are calculated for each particle $\mathbf{p}_k^{(s)}$, $s = 0, ..., S-1$ and frequency bin $b = 0, ..., B-1$ as

$$d_k^{(s)}[b] = \ln\left\{\exp a_k^{(s)}[b] + s_k^{(s)}[b]\exp r_k[b]\right\}$$

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 $p_{\text{speech}}(\cdot)$ each distortion sample $\mathbf{d}_k^{(s)}$ is evaluated according to the likelihood

$$p(\mathbf{y}_k|\mathbf{d}_k^{(s)}) = \frac{p_{\text{speech}}(\mathbf{y}_k + \ln(\mathbf{1} - e^{\mathbf{d}_k^{(s)} - \mathbf{y}_k}))}{\prod_{b=1}^{B}\left|1 - e^{\mathbf{d}_{k,b}^{(s)} - \mathbf{y}_{k,b}}\right|}, \qquad (2)$$

and its normalized weight

$$w_k^{(s)} = \frac{p(\mathbf{y}_k|\mathbf{d}_k^{(s)})}{\sum_{m=1}^{S} p(\mathbf{y}_k|\mathbf{d}_k^{(m)})} \qquad (3)$$

is calculated. Note that the likelihood can only be evaluated if

$$d_{k,b}^{(s)} < y_{k,b}\ \forall\, b\ \in\ B,$$

otherwise the particle weight is set to zero. This causes a decimation of the particle population which we prevent by the *fast acceptance test* [14]. In this procedure, a drawn sample is only accepted if its likelihood can be evaluated. Otherwise, a new sample is drawn from another randomly chosen particle.

2.7. Distortion Compensation

The clean feature is estimated with the distortion samples $\mathbf{d}_k^{(s)}$ and their corresponding importance weights $w_k^{(s)}$ over all particle samples S using the non-linear relationship between $\mathbf{x}_k$, $\mathbf{d}_k$ and $\mathbf{y}_k$ as in [14],

$$\mathrm{E}\{\mathbf{x}_k|\mathbf{y}_{1:k}\} = \sum_{s}^{S} w_k^{(s)}\left(\mathbf{y}_k + \ln(\mathbf{1} - e^{\mathbf{d}_k^{(s)} - \mathbf{y}_k})\right). \qquad (4)$$

2.8. Importance Resampling & Prediction Model Estimation

After every time step, the particles are *resampled*, in order to avoid the concentration of the vast majority of probability mass in very few particles, and the prediction model,

$$\mathbf{A}_k = \mathrm{E}\{\mathbf{a}_k\mathbf{a}_{k-1}^T\}\mathrm{E}\{\mathbf{a}_{k-1}\mathbf{a}_{k-1}^T\}^{-1}, \qquad (5)$$

is estimated by the dynamic autoregressive process [13]. Within this framework the expectation of the required matrices

$$\mathrm{E}\{\mathbf{a}_k\mathbf{a}_{k-1}^T\} \ = \ \frac{1}{S}\sum_{s=1}^{S} p(\mathbf{y}_k|\mathbf{a}_k^{(s)})\ \mathbf{a}_k\mathbf{a}_{k-1}^{(s)}{}^T$$

and

$$\mathrm{E}\{\mathbf{a}_{k-1}\mathbf{a}_{k-1}^T\} \ = \ \frac{1}{S}\sum_{s=1}^{S} p(\mathbf{y}_k|\mathbf{a}_k^{(s)})\ \mathbf{a}_{k-1}\mathbf{a}_{k-1}^{(s)}{}^T$$

are calculated by a weighed summation over all additive distortions $\mathbf{a}^{(s)}$, $s = 1, 2, \ldots, S$ due to their corresponding likelihoods (2).

Microphone			CTM		Lapel		Table Top		Wall	
Distance			1 cm		20 cm		150-200 cm		300-400 cm	
SNR			24 dB		23 dB		17 dB		10 dB	
Pass			1	2	1	2	1	2	1	2
Front-End	Compensation		Word Error Rate %							
	Additive	Reverberation								
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 tract 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 approaches are again superior to the independent treatment of the components and demonstrates significant gains for unadapted as well as unsupervised adapted acoustic models. The introduction of the tilt factor (joint 2) in addition to the scaling factor (joint 1) further improves the accuracy. 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

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, Apr. 1979.

[2] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[3] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[4] M. Fujimoto and S. Nakamura, "Particle filter based non-stationary noise tracking for robust speech feature enhancement," *Proc. of ICASSP*, 2005.

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

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

[7] I. Tashev and D. Allred, "Revereberation reduction for improved speech recognition," in *Proc. of HSCMA*, 2005.

[8] R. Petrick, K. Lohde, M. Wolff, and R. Hoffmann, "The harming part of room acoustics in automatic speech recognition," *Proc. of Interspeech*, pp. 1094–1097, 2007.

[9] K. Kinoshita, T. Nakatani, and Miyoshi M., "Efficient dereverberation framework for automatic speech recognition," *Proc. of Interspeech*, pp. 3145–3148, 2005.

[10] D. Gespert and P. Duhamel, "Robust blind identification and equalization based on multi-step predictors," *Proc. of ICASSP*, vol. 26, no. 5, pp. 3621–3624, 1997.

[11] 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.

[12] L. Deng, J. Droppo, and A. Acero, "A bayesian approach to speech feature enhancement unsing the dynamic cepstral prior," in *Proc. of ICASSP*, 2002.

[13] M. Wölfel, "Integration of the predicted walk model estimate into the particle filter framework," in *Proc. of ICASSP*, 2008.

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

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

[16] 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.