ITERATIVE SPEECH ENHANCEMENT USING A NON-LINEAR DYNAMIC STATE MODEL OF SPEECH AND ITS PARAMETERS

Stefan Windmann, Reinhold Haeb-Umbach

University of Paderborn
Dept. of Communications Engineering
33098 Paderborn, Germany
{windmann,haeb}@nt.uni-paderborn.de

ABSTRACT

A marginalized particle filter is proposed for performing single channel speech enhancement with a non-linear dynamic state model. The system consists of a particle filter for tracking line spectral pair (LSP) parameters and a Kalman filter per particle for speech enhancement. The state model for the LSPs has been learnt on clean speech training data. In our approach parameters and speech samples are processed at different time scales by assuming the parameters to be constant for small blocks of data. Further enhancement is obtained by an iteration which can be applied on these small blocks. The experiments show that similar SNR gains are obtained as with the Kalman-EM-iterative algorithm. However better values of the noise level and the log-spectral distance are achieved.

1. INTRODUCTION

Single-channel speech enhancement has been an area of active research for a long time, with an even increased interest in recent years due to new challenging applications in the domain of mobile phones, which often operate in very adverse acoustic environments. While algorithms employing the short-term frequency transform of the incoming signal provide large signal-to-noise ratio gains, time domain algorithms are known for delivering excellent speech quality [1]. The latter approach typically builds on the autoregressive (AR) model of speech production, from which a state space model of the speech signal can easily be derived, making the Kalman filter the preferred choice for time-domain speech enhancement [2]-[5].

Since the AR model parameters are not known in advance and change over time, they have to be estimated alongside the speech enhancement. Gannot has developed the Kalman-EM-iterative (KEMI) algorithm, a batch algorithm to iteratively estimate the parameters of the state space model and enhance the noisy speech signal [3]. He also proposed a sequential algorithm which almost achieves the performance of the KEMI algorithm. Recursive sequential algorithms have also been proposed in [4] and [5] to estimate the AR parameters.

Recently, attempts have been reported to explicitly model the evolvement of the AR parameters over time by a state space model. The state vector then consists of both the recent clean speech samples and the AR parameters. The resulting non-linear estimation problem can be approached by applying a linearisation, e.g. the analytical linearisation of an extended Kalman filter or the statistical linearisation of an unscented Kalman filter [6] with, however, mixed results sofar. Alternatively, particle filters may be used [7, 8]. These sequential Monte Carlo methods have already been used successfully in many non-linear tracking applications [9, 10]. While in [7] the enhancement of individual phoneme transitions is considered by employing a random walk model for the time varying AR parameters, a model based on partial correlation (PARCOR) coefficients is developed in [8] which is applied to short sections of both speech and music data. Particle filters have also been employed for speech feature enhancement for subsequent speech recognition [11], [12].

Our approach is based on the work of [7] and [8] in that the same Rao-Blackwellized particle filtering structure is used: taking advantage of the conditional linearity of parts of the state model sampling in a high-dimensional state space can be avoided. We, however, operated the particle filter on a block of input data, which, alongside computational advantages, resulted in better enhancement.

While the AR model is the parametric representation of speech that is most commonly used in the context of speech enhancement [2] - [7], we employed Line Spectral Pair (LSP) parameters. They showed superior predictive power when used in state space models learnt from clean speech training data.

This paper is organized as follows. In the next Section we describe the Rao-Blackwellized particle filter introduced in [7] and [8]. In Section 3 we modify the particle filter to operate on blocks of input samples. In Section 4 we motivate the use of LSPs, estimate a corresponding dynamic state space model in Section 5 and outline our proposed speech enhancement algorithm in Section 6. Then we present experimental results in Section 7 and finish with some conclusions.

2. NON-LINEAR DYNAMIC STATE MODEL

We are concerned with estimating the clean speech signal $s(k)$ from the noisy observations

$$y(k) = s(k) + \sqrt{g_v(k)}v(k), \quad k = 1, 2, \ldots \tag{1}$$

where $v(k)$ denotes normalized (zero mean, unit variance) additive Gaussian noise, which is assumed to be white for simplicity. An extension to colored noise is possible e.g. by following the approach described in [3]. $\sqrt{g_v(k)}$ is the gain of the noise.

Using the source filter model of speech production, the clean

speech signal $s(k)$ can be written as follows

$$s(k) = \sum_{j=1}^{p} a_j(k)s(k-j) + \sqrt{g_s(k)}w_s(k). \quad (2)$$

Here $a_j(k)$, $j = 1, \ldots, p$ are the time-varying autoregressive (AR) parameters, $w_s(k)$ represents the normalized white Gaussian excitation noise and $\sqrt{g_s(k)}$ is the gain of the noise. Assuming that the AR coefficients and the noise level are known, a linear state space model with state vector $\mathbf{s}(k) = (s(k), \ldots, s(k-p+1))^T$ and a state transition matrix which contains the AR coefficients can easily be derived from eq. (2) [3].

In [7] the state-space model was extended so that the evolvement of the AR parameters over time was also considered. Now the state vector

$$\mathbf{x}(k) = (\mathbf{s}^T(k),\ \boldsymbol{\theta}^T(k))^T \quad (3)$$

consists of the signal state, i.e. the recent clean speech samples $\mathbf{s}(k)$, *and* the parameter state $\boldsymbol{\theta}(k) = (\mathbf{a}^T(k), g_s(k), g_v(k))^T$, i.e. the AR coefficients $\mathbf{a}(k) = (a_1(k), \ldots, a_p(k))^T$ and the time-varying gain factors. This state space model is in fact nonlinear, and sequential Monte Carlo methods can be used to estimate the posterior probability of the state vector, given the observations (1). Using the aforementioned conditional linearity of the model of $\mathbf{s}(k)$ given $\boldsymbol{\theta}(k)$, a more efficient combined particle and Kalman filter can be obtained by Rao-Blackwellization [10]. The resulting system consists of a particle filter for the subvector $\boldsymbol{\theta}(k)$ and a Kalman filter for estimating the corresponding signal states $\mathbf{s}^{(i)}(k)$ for each particle $\boldsymbol{\theta}^{(i)}(k)$, $i = 1, \ldots, N_p$ (N_p: number of particles); see [7] and [8] for details.

3. BATCH PROCESSING

In this paper we consider a slightly more general model, which will reveal an interesting interpretation of the resulting enhancement system. Since the AR coefficients may change on a different, typically slower time scale than the speech samples and since they may be easier estimated on blocks of samples than on a single sample, we use a piecewise constant model for the AR coefficients: $\mathbf{a}(mM + l)$ is assumed to be constant during a block of M samples, i.e. during $l = 0, 1, \ldots, M-1$, and to change between subsequent blocks:

$$\mathbf{a}(m+1) = \mathbf{F}_a\, \mathbf{a}(m) + \mathbf{G}_a\, \mathbf{w}_a(m), \quad (4)$$

where m denotes the block counter, and $\mathbf{w}_a(m)$ is white Gaussian noise with $E[\mathbf{w}_a\mathbf{w}_a^T] = \mathbf{I}_p$, $\mathbf{I}_p$ being the $(p \times p)$ identity matrix. The matrices $\mathbf{F}_a$ and $\mathbf{G}_a$ are the state transition and input matrix, respectively, which we learnt from clean speech training data, see Section 4.

The measurement vector for the state estimation of the AR coefficients consists of the M noisy speech samples $y(mM + l)$, $l = 0, \ldots, M-1$. Since all noise terms are assumed to be Gaussian, the likelihood function of the observations $y(mM + l)$, $l = 0, \ldots, M-1$, given a hypothesized AR vector, i.e. given a particle $\mathbf{a}^{(i)}(m)$, and given the estimates $\hat{\mathbf{s}}^{(i)}(mM+l)$, $l = 0, \ldots, M-1$, of the clean speech provided by the i-th Kalman filter, is Gaussian. The exponent of this Gaussian can be approximately written as

$$\frac{1}{2}\sum_{l=0}^{M-1} \Big(y(mM+l) - (\mathbf{a}^{(i)})^T(m)\hat{\mathbf{s}}^{(i)}(mM+l-1)\Big)^2 \quad (5)$$
$$(\boldsymbol{\Sigma}_{y|Y}^{(i)}(mM+l-1) + g_v)^{-1}$$

where subsequent noisy speech samples are considered conditionally independent. $\boldsymbol{\Sigma}_{y|Y}^{(i)}$ denotes the covariance matrix of the predicted speech sample which is provided by the i-th Kalman filter.

Finding the value of $\mathbf{a}(m)$, which minimizes this term (and thus maximizes the likelihood), greatly resembles block-oriented LPC analysis [14], where the next sample (here: $y(mM + l)$) is to be predicted from previous ones (here: $\hat{\mathbf{s}}^{(i)}(mM + l - 1)$). The optimal predictor for this task is exactly the value of $\mathbf{a}(m)$ which minimizes eq. (5). A key advantage of the particle filter is that the speech samples need not be reduced to a single estimate of $\mathbf{a}(m)$, before tracking is done. Hypothesized values of the parameter state variable, i.e. the particles $\mathbf{a}^{(i)}(m)$, are individually assessed by the likelihood function, whose exponent is given by eq. (5), and tracked over time. An estimate of the state variable is then obtained as a weighted sum of particles. This could be called a "track before detect" strategy.

4. LSP VS. AR COEFFICIENTS

The well-known all-pole source-filter model of speech production given in eq. (2) makes a parametric representation of speech by AR parameters the natural choice in the context of Kalman filter based speech enhancement. However, the estimation of these linear prediction coefficients is known to be susceptible to noise. In this section we therefore explore the use of other parametric descriptions. In informal experiments we found that among the various parametric descriptions of speech line spectral pairs (LSP) seemed to be the most promising candidate.

The LSP coefficients $\mathbf{a}^{(lsp)}(m)$ are an equivalent representation of the AR coefficients $\mathbf{a}^{(ar)}(m)$ where the poles of the transfer function $1/A(z)$ in the model of speech generation are mapped onto the unit circle using two auxiliary polynomials $P(z)$ and $Q(z)$. The auxiliary polynomials are calculated via the relations $P(z) = A(z) + z^{-(p+1)}A(z^{-1})$ and $Q(z) = A(z) - z^{-(p+1)}A(z^{-1})$, where $A(z)$ is the z-transform of the AR coefficients. The LSPs are selected to be the phases ϕ_i, $i = 1, \ldots, p$, where $e^{j\phi_i}$, $i = 1, \ldots, p$ are roots of $P(z)$ or $Q(z)$ [13].

The Frobenius norm $\|\mathbf{G}_a\|_F$ of the input matrix, i.e. the trace of $\mathbf{G}_a\mathbf{G}_a^T$, is a measure for the predictive power of the state space model, see eq. (7) further below. After learning state space models for AR and LSP coefficients from the same training data, we found that $\|\mathbf{G}_a^{(ar)}\|_F = 0.21$ and $\|\mathbf{G}_a^{(lsp)}\|_F = 0.08$, which demonstrates the higher predictive power of the LSP state space model.

5. STATE MODEL ESTIMATION

The matrices $\mathbf{F}_a^{(lsp)}$ and $\mathbf{G}_a^{(lsp)}$ of the state space model of LSP coefficients are estimated prior to speech enhancement as follows. First, sequences $\mathbf{a}^{(ar)}(m)$, $m = 1, 2, \ldots$ of AR coefficients are calculated from blocks of clean speech data using the autocorrelation method. From these, sequences $\mathbf{a}^{(lsp)}(m)$, $m = 1, 2, \ldots$ of LSP coefficients are computed [14]. The entries $f_{ij}^{(lsp)}$ of the state transition matrix $\mathbf{F}_a^{(lsp)}$ are determined by minimizing the expected values $E[(a_i^{(lsp)}(m) - \sum_{j=1}^{p} f_{ij}^{(lsp)} a_j^{(lsp)}(m-1))^2]$, $i = 1, \ldots, p$. Taking partial derivatives w.r.t. the unknowns the

following system of equations results:

$$\mathbf{P}_a = \mathbf{R}_a \mathbf{F}_a$$

with

$$\mathbf{P}_a = \begin{bmatrix} E[a_1(m-1)a_1(m)] & \dots & E[a_1(m-1)a_p(m)] \\ \vdots & \ddots & \vdots \\ E[a_p(m-1)a_1(m)] & \dots & E[a_p(m-1)a_p(m)] \end{bmatrix}$$

and

$$\mathbf{R}_a = \begin{bmatrix} E[a_1^2(m-1)] & \dots & E[a_1 a_p(m-1)] \\ \vdots & \ddots & \vdots \\ E[a_p a_1(m-1)] & \dots & E[a_p^2(m-1)] \end{bmatrix}, \tag{6}$$

where we have omitted the superscript for ease of notation. From this $F_{\mathbf{a}}$ can be obtained.

The matrix $\mathbf{G}_a$ can be determined with the relation

$$\mathbf{G}_a \mathbf{G}_a^T = E[(\mathbf{a}_k - F_{\mathbf{a}}\mathbf{a}_{k-1})(\mathbf{a}_k - F_{\mathbf{a}}\mathbf{a}_{k-1})^T] \tag{7}$$

and subsequent Cholesky factorisation.

6. ITERATIVE SPEECH ENHANCEMENT

A Sampling-Importance-Resampling (SIR) particle filter [10] was used for estimating the trajectory of the LSP coefficients.

Note that the N_p Kalman filters operate on a state transition matrix containing the AR coefficients as described in Section 2. Therefore the LSP coefficients have to be converted to AR coefficients, which is done as follows: the components of the p-dimensional particle $(\mathbf{a}^{(lsp)})^{(i)}$ are the phases $\phi_1, \dots, \phi_p$. From these the roots of $P(z)$ and $Q(z)$ are obtained, using the unique properties of the roots of $P(z)$ and $Q(z)$. From these in turn the AR coefficients are computed.

In the following we sketch the operations to be conducted on one block of noisy input speech data.

1. Draw N_p samples from the Gaussian $p(\mathbf{a}^{(lsp)}(m)|\mathbf{a}^{(lsp)}(m-1))$.
2. For each sample $(\mathbf{a}^{(lsp)})^{(i)}(m)$:
   a) Compute AR coeff. $(\mathbf{a}^{(ar)})^{(i)}(m)$ and plug them into the system matrices of the i-th Kalman filter.
   b) Perform Kalman filtering of the input data $y(mM+l)$, $l = 0, \dots, M-1$ to produce estimates $\hat{s}^{(i)}(mM+l), l = 0, \dots, M-1$ of the clean speech and correspondig prediction error covariances.
3. Compute weights $w^{(i)}(m)$, $i = 1, \dots, N_p$, which are proportional to the Gaussian whose exponent is given in (5).
4. Draw N_p samples with replacement from the set $\{(\mathbf{a}^{(lsp)})^{(i)}(m),\ i = 1, \dots, N_p\}$ to ensure particle diversity.
5. Set $m := m + 1$ and go to 1.

An overall estimate of clean speech can be obtained as

$$\hat{s}(mM+l) = \sum_{i=1}^{N_p} w^{(i)}(m)\hat{s}^{(i)}(mM+l); l = 0, \dots, M-1. \tag{8}$$

No voice activity detection (VAD) is needed in this proposed speech enhancement system. The particles are initialized with the overall mean vector of the LSP coefficients obtained from the clean speech training data. Unlike [7] the gain values g_v and g_s are not considered as being part of the state vector of the particle filter but as parameters which are estimated separately from the noisy speech signal $y(k)$. Since we assumed that the noise is stationary, g_v is actually only estimated once at the beginning of a data file before the onset of speech and this value is then kept constant from that on. An estimate $\hat{g}_s^{(init)}$ for the gain value g_s is calculated independently for blocks of length 128 using the approach proposed by [3]. In [3] this initial estimate is improved using the covariance matrix of the filtered state vector $\hat{\mathbf{s}}(k)$. However this approach only works for large blocks and is therefore not applicable for the small block sizes considered here, see Section 7. We therefore keep these initial estimates from enclosing frames of length 128, but obtain improved speech enhancement by the following iteration performed on blocks of size M: The Kalman filtering operation step 2.b) and the weighting computation step 3) in the proposed algorithm are repeated, where the noisy input speech is replaced by the overall estimates of the clean speech (8) obtained in the previous iteration.

7. EXPERIMENTAL RESULTS

We conducted experiments on sentences of the Wallstreet Journal corpus (WSJ) to which white Gaussian noise was artificially added at different SNRs.

In a first set of experiments parameters of the proposed Particle-Kalman speech enhancement system were determined. The number of modelled coefficients was fixed to the value $p = 3$. In informal experiments we observed that for higher model orders the SNR gain was smaller due to the difficulties in reliably tracking all model parameters. Tab. 1 shows the SNR gain as a function of the block length M for three different input SNR values. The value $M = 16$ was chosen for the subsequent experiments as a compromise between large blocks where the particle filter is unable to track fast changes of the parameter vector, e.g. at phoneme transitions, and small blocks where the measurements are less stable. Compared with $M = 1$ an SNR gain of at least $0.5dB$ is

M	**1**	**2**	**4**	**8**	**16**	**32**	**64**	**128**	**256**
3dB	3.2	3.4	3.8	3.9	4.1	4.1	4.1	4.0	3.9
5dB	3.3	3.6	3.8	3.9	4.0	4.0	4.0	3.7	3.6
10dB	2.5	2.7	2.9	3.0	3.0	3.1	3.1	3.0	2.8

Table 1. SNR gains with distinct input SNRs and block lengths M

achieved with $M = 16$. Also the runtime is reduced with larger blocks, because the conversion between LSP and LPC coefficients, which costs about half the time required to process a sample, must only be performed once per block.

In the next set of experiments the dependence of the SNR gain on the number of particles was determined, see Tab. 2. The value $N_p = 100$ was selected which led to almost the same SNR gain as $N_p = 500$.

Next the proposed Particle-Kalman algorithm (denoted PK2 in the following) with the afore determined parameters and one iteration was compared with Gannot's famous Kalman-EM-iterative

N_p	10	25	50	100	200	500
3dB	2.9	3.7	3.9	4.0	4.0	4.1
5dB	2.4	3.5	3.6	3.7	3.7	3.8
10dB	2.0	2.8	2.8	3.0	3.0	3.0

Table 2. SNR gains with distinct input SNRs and particles N_p

algorithm (KEMI) [3] and an algorithm very similar to Vermaak's approach (PK) [7], which we had used as a starting point for our research. The KEMI algorithm was applied with five iterations, non-overlapping blocks of size $M = 128$, model order $p = 10$ and noise modelled to be additive, white and gaussian. In PK the block length is selected to be $M = 1$ and the model order is $p = 3$. We used $N_p = 100$ particles, a random walk state model of AR coefficients and no iteration. Unlike [7] in PK the gain factor g_s is separated from the state vector and calculated as described in [3]. For the evaluation we used four different quality measures proposed by Gannot [3] [15] and averaged the results over 20 sentences (Fig. 1).

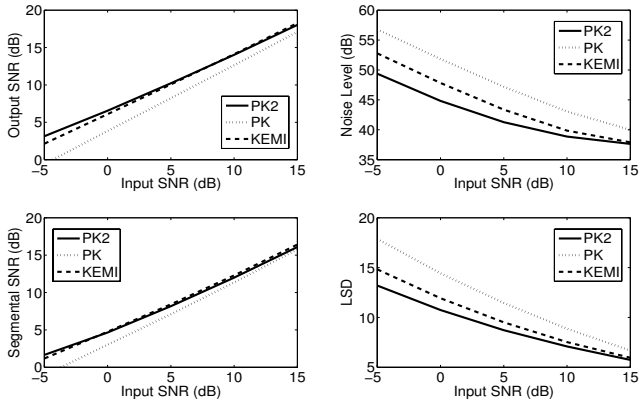

Fig. 1. *Figures-of-merit for 20 WSJ sentences with Gaussian noise*

First the output SNRs [15] of the considered algorithms were calculated at several input SNRs. For small input SNRs PK2 achieves slightly higher SNR gains than KEMI while for high input SNRs KEMI leads to slightly higher gains. This result may be explained by the fact that the AR parameters can be determined very well from an almost clean speech signal, while for low input SNRs the detection of the AR parameters is not very reliable and a particle filter tracking can be helpful. In doing so it must be taken into account that for KEMI the results can be improved for SNRs less than 5dB by using Higher Order Statistics for the initial parameter estimation, which we, however, did not consider here.
The Segmental SNR [3], which is calculated during active speaker periods, leads to similar results, while the noise level in speech pauses [15] is lower with PK2. Since the SNR gain is not very correlative with speech quality, in [15] Gannot proposed to use the log spectral distance (LSD), which is better correlated with the mean opinion score. In terms of the LSD distance, which is calculated during speech periods, PK2 yields better results than KEMI. The algorithm PK, which has been used as starting point for the development of PK2, yields for the quality measures and input SNRs considered here worse results than PK2 and KEMI.

8. CONCLUSIONS

In this paper we have extended the marginalized particle filter presented in [7] to iterative block processing and employed a state space model of LSP parameters which has been learnt from training data. The approach was applied to the enhancement of noisy sentences rather than single phoneme transitions. The SNR gains achieved were comparable to those obtained by the Kalman-EM-iterative algorithm, while better values of the noise level and the log-spectral distortion are achieved where the latter measure is known to be well correlated with perceptual speech quality. This improvement, however, comes at significantly larger computational complexity.

9. ACKNOWLEDGEMENTS

The research is partly supported by the DFG Research Training Group GK-693 of the Paderborn Institute for Scientific Computation (PaSCo).

10. REFERENCES

[1] J. Benesty, S. Makino, and J. Chen (Eds.), *Speech Enhancement*, Springer, 2005.

[2] K.K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering", in *Proc. ICASSP*, Dallas, Texas, 1987.

[3] S. Gannot, "Iterative and sequential Kalman filter-based speech enhancement algorithms", *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 4, pp. 373-385, July 1998.

[4] N. Ma, M. Bouchard, and R.A. Goubran, "Perceptual Kalman filtering for speech enhancement in colored noise", in *Proc. ICASSP*, Montreal, 2004.

[5] M. Gabrea, "An adaptive Kalman filter for the enhancement of speech signals", in *Proc. Int'l Conference on Spoken Language Proc.*, Jeju, Korea, 2004.

[6] S. Gannot and M. Moonen, "On the application of the unscented Kalman filter to speech processing", in *Proc. Int'l Workshop on Acoustic Echo and Noise Control*, Kyoto, Japan, 2003.

[7] J. Vermaak, C. Andrieu, A. Doucet, and S. Godsill, "Particle methods for bayesian modeling and enhancement of speech signals", *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 3, pp. 173-185, March 2002.

[8] W. Fong, S. Godsill, A. Doucet, and M. West, "Monte Carlo smooting with application to audio signal enhancement", *IEEE Trans. Signal Processing*, vol. 50, no. 2, pp. 438-449, Feb. 2002.

[9] A. Doucet, J.F.G. de Freitas and N.J. Gordon (Eds), *Sequential Monte Carlo Methods in Practice*, Springer, 2001.

[10] B. Ristic, S. Arulampalam, N. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications*, Artech House Publishers, 2004.

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

[12] R. Haeb-Umbach and J. Schmalenstroeer, "A comparision of particle filter variants for speech feature enhancement", in *Proc. Interspeech*, Lisboa, 2005.

[13] K. Soong, B.H. Juang, "Line spectrum pair (LSP) and speech data compression", in *Proc. ICASSP*, San Diego, 1984.

[14] J.R. Deller, J.H.L. Hansen, and J.G. Proakis *Discrete-Time Processing of Speech Signals*, IEEE Press, 2000.

[15] S. Gannot, "Speech Enhancement: Application of the Kalman Filter in the Estimate-Maximize (EM) Framework", in *Speech Enhancement*, J. Benesty, S. Makino, J. Chen (Eds.), Springer, 2005.