Einkanalige Sprachsignalverbesserung mit Hilfe eines marginalisierten Partikelfilters

S. Windmann, R. Häb-Umbach
FG Nachrichtentechnik / Universität Paderborn, D-33098 Paderborn
{windmann, haeb}@nt.uni-paderborn.de

Zusammenfassung

Es wird ein marginalisiertes Partikelfilter beschrieben, das zur einkanaligen Sprachsignalverbesserung mit einem nichtlinearen dynamischen Zustandsmodell eingesetzt werden soll. Das System besteht aus einem Partikelfilter zum Tracking von LSP-Parametern und einem Kalman-Filter für jedes Partikel, das zur Sprachsignalverbesserung verwendet wird. In unserem Ansatz wird angenommen, dass die Parameter in kurzen Sprachsignalblöcken konstant sind, während das Sprachsignal sich mit jedem Abtastwert ändert. Bei weißem Rauschen werden ähnliche SNR-Gewinne wie mit einem Kalman-EM-iterative Algorithmus erzielt, während das Hintergrundrauschen und die Log-spektrale Distanz etwas geringer sind. Mit einem erweiterten Zustandsmodell wurden auch Untersuchungen für farbiges Rauschen durchgeführt.

1. Einleitung

Einkanalige Sprachsignalverbesserung ist seit langer Zeit Gegenstand intensiver Forschung, der in den vergangenen Jahren wegen neuer herausfordernder Anwendungen im Mobilfunkbereich sogar ein zunehmendes Interesse erfahren hat. Während Algorithmen, die die Kurzzeit-Fouriertransformation des verrauschten Signals anwenden, zu großen SNR-Gewinnen führen, sind Verfahren im Zeitbereich dafür bekannt, exzellente Sprachqualität zu bewirken [1]. Der Zeitbereichsansatz basiert üblicherweise auf dem autoregressiven (AR-) Modell der Spracherzeugung von dem ein Zustandsmodell des Sprachsignals einfach abgeleitet werden und in einem Kalmanfilter zur Sprachsignalverbesserung verwendet werden kann [2][3]. Da die Parameter des AR-Modells nicht a-priori bekannt sind und sich abhängig von der Zeit ändern, müssen sie parallel zur Filterung geschätzt werden. Gannot hat zu diesem Zweck den Kalman-EM-iterative (KEMI) Algorithmus entwikkelt, einen Batch-Algorithmus, mit dem iterativ die Parameter des Zustandsmodells geschätzt und das verrauschte Sprachsignal verbessert werden können [3]. Kürzlich sind Ansätze erschienen, die die Änderung der AR-Parameter über der Zeit explizit mit einem Zustandsmodell modellieren. Eine approximative Lösung des resultierenden nichtlinearen Schätzproblems ist mit Hilfe des Unscented Kalman Filters [4] möglich, wodurch bislang jedoch keine überzeugenden Resultate erzielt wurden. Wir verwenden stattdessen eine marginalisierte Partikelfilterstruktur, die in [5] zur Verbesserung individueller Phonemübergänge mit einem Random-Walk-Modell für die zeitveränderlichen AR-Parameter ausgenutzt wird. In [6] wird dieser Ansatz in einem Modell eingesetzt, das auf PARCOR-Koeffizienten basiert und sowohl für kurze Sprach- als auch Musikdaten angewendet wurde. Unser Partikelfilter operiert auf Blöcken des verrauschten Sprachsignals, woduch neben einer reduzierten Rechenzeit eine Verbesserung der Sprachqualität erreicht wird.

Im nächsten Abschnitt beschreiben wir das marginalisierte Partikelfilter, das wir in [10] entwickelt haben. Eine Modellerweiterung für farbiges Rauschen wird in Abschnitt 3 beschrieben. In Abschnitt 4 präsentieren wir Testergebnisse für weißes und farbiges Rauschen.

2. Marginalisiertes Partikelfilter

Ziel ist es, das unverrauschte Sprachsignal $s(k)$ aus den verrauschten Beobachtungen

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

zu schätzen, wobei $v(k)$ normalisiertes (Mittelwert Null, Varianz Eins) additives, weißes Gauss'sches Rauschen und $\sqrt{g_v(k)}$ der Verstärkungsfaktor des Rauschens sind. Das unverrauschte Sprachsignal kann unter Verwendung des Quelle-Filter-Modells der Spracherzeugung in der Form

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

geschrieben werden. Dabei bezeichnen $a_j(k), j=1\ldots p,$ die zeitvarianten autoregressiven (AR-) Parameter, $w_s(k)$ die normalisierte, weiße gaußverteilte Anregung und $\sqrt{g_s(k)}$ die Verstärkung der Anregung. Unter der Annahme, dass die AR-Koeffizienten und der Rauschpegel bekannt sind, kann ein lineares Zustandsmodell mit dem Zustandsvektor $\boldsymbol{s}(k)=(s(k)\ldots s(k-p+1))^T$ und einer Zustandsübergangsmatrix, die die AR-Koeffizienten enthält, abgeleitet werden [3]. In [5] wurde dieses Zustandsmodell erweitert, so dass die zeitliche Änderung der AR-Parameter berücksichtigt wurde. Eine weitere Verallgemeinerung dieses Modells haben wir in [10] vorgenommen, indem wir ein stückweise konstantes Modell für die AR-Koeffizienten angenommen haben. Es wird also davon ausgegangen, dass der Koeffizientenvektor $\boldsymbol{a}(mM+l)$, der sich aus den Koeffizienten $a_j(mM+l), j=1\ldots p$ zusammensetzt, für Blöcke der Länge M Samples konstant ist, d.h. für $l=0,1,\ldots M-1$. Für die Änderung zwischen aufeinander folgenden Blöcken wird ein lineares Zustandsmodell angenommen:

$$\boldsymbol{a}(m+1)=F_a\boldsymbol{a}(m)+G_a w_a(m),$$

wobei m den Blockzähler und $w_a(m)$ weißes, gaußverteiltes Rauschen mit der Kovarianzmatrix $E[w_a w_a^T]=I_p$, einer $p\times p$-Einheitsmatrix, sind. Der Zustandsvektor

$$\boldsymbol{x}(m,k)=(\boldsymbol{s}^T(k),\boldsymbol{\Theta}^T(m))^T$$

besteht jetzt aus dem Zustand des Signals, d.h. den zuletzt abgetasteten Werten des Sprachsignals $\boldsymbol{s}(k)$, und dem Zustand des Parametervektors $\boldsymbol{\Theta}(m)=\boldsymbol{a}(m)$. Anstelle der AR-Koeffizienten verwenden wir dabei die äquivalenten Line Spectral Pair (LSP-) Koeffizienten [11], die einen geringeren Prädiktionsfehler aufweisen [10]. Die Matrizen F_a und G_a sind die Zustandsübergangs- bzw. Eingangsmatrix, die wir aus unverrauschten Sprachdaten trainiert haben. Dieses Modell ist nichtlinear, und sequentielle Monte Carlo Methoden können dazu verwendet werden, die a-posteriori Wahrscheinlichkeit des Zustandsvektors bei gegebenen Beobachtungen zu schätzen. Unter Verwendung der bedingten Linearität des Modells von $\boldsymbol{s}(k)$ bei gegebenem $\boldsymbol{\Theta}(m)$ erhält man durch Marginalisierung ein kombiniertes Partikel- und Kalmanfilter [10], dessen Funktion an dieser Stelle skizziert wird. Die Verteilungsdichte des Koeffizientenvektors $\boldsymbol{\Theta}(m)$ wird durch Partikel $\boldsymbol{\Theta}^{(i)}(m), i=1\ldots N_p,$ approximiert, die durch Sampling aus der Gaußverteilung $p(\boldsymbol{\Theta}(m)|\boldsymbol{\Theta}(m-1))$ prädiziert werden. Mit den prädizierten Koeffizienten $\boldsymbol{\Theta}^{(i)}(m)$ werden die Systemmatrizen von i Kalmanfiltern berechnet, mit denen die Zustandsschätzungen $\hat{s}^{(i)}(mM+l)$ und Kovarianzen $\Sigma_{y|Y}^{(i)}(mM+l)$ aus den Eingangsdaten $y(mM+l)$ bestimmt werden. Den Partikeln werden Gewichte $w^{(i)}(m)$ zugeordnet, die proportional zu dem Wert einer Gaußverteilung mit dem Exponenten

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

sind. Entsprechend der Gewichte wird ein Resampling durchgeführt. Die Schätzung des unverrauschten Sprachsignals ergibt sich als

$$\hat{s}(mM+l)=\sum_{i=1}^{N_p} w^{(i)}(m)\hat{s}^{(i)}(mM+l).$$

3. Erweiterung auf farbiges Rauschen

Zur Erweiterung auf farbiges Rauschen wird $y(k)$ in der Form $y(k)=s(k)+n(k)$ geschrieben, wobei $n(k)$ als AR-Prozeß der Form

$$n(k)=\sum_{j=1}^{q}\beta_j(m)n(k-j)+\sqrt{g_v(m)}v(k)$$

modelliert wird und $\beta_j(m), j=1...q,$ die zeitvarianten AR-Parameter des Rauschens sind. Der Zustandsvektor wird zur Erweiterung auf farbiges Rauschen um Komponenten für die q letzten Rauschsignalabtastwerte erweitert. Man erhält somit das erweiterte Zustandsmodell

$$\boldsymbol{x}(m)=(\boldsymbol{s}^T(k),\boldsymbol{n}^T(k),\boldsymbol{\Theta}^T(m))^T.$$

Die zusätzlichen Parameter $\beta_i(m)$ werden mit Hilfe eines parallel laufenden Kalman-EM-iterative-Algorithmus geschätzt. Zur Anpassung der Kalmanfilterung an das erweiterte Zustandsmodell werden für jedes Partikel die in [3] angegeben Kalmanfiltergleichungen verwendet.

4. Testergebnisse

Als Datenbasis für die Experimente dienten Sätze des Wallstreet Journal Corpus (WSJ), denen weißes, gaußverteiltes Rauschen bei verschiedenen SNR künstlich überlagert wurde. Die Testdurchläufe für den vorgeschlagenen Partikel-Kalman-Algorithmus (im folgenden als PK2 bezeichnet) erfolgten mit den in [10] ermittelten Parametern Blocklänge $M=16$, Modellordnung $p=3$ und Partikelanzahl $N_p=100$. Zum Vergleich der Ergebnisse wurden Gannot's bekannter Kalman-EM-iterative (KEMI) Algorithmus sowie ein Ansatz, der dem Verfahren von Vermaak ähnelt (PK) und als Ausgangspunkt für unsere Untersuchungen diente, herangezogen. Der KEMI-Algorithmus wurde mit fünf Iterationen, nichtüberlappenden Blöcken der Länge $M=128$, der Modellordnung $p=10$ sowie additivem, gaußverteilten weißen Rauschen parametrisiert. PK wurde mit $M=1$ und $p=3$ getestet. Wir haben $N_p=100$ Partikel sowie ein Random-Walk-Modell für das zeitliche Verhalten der AR-Koeffizienten verwendet. Anders als in [5] wurde der Gainfaktor g_s nicht in den Zustandsvektor integriert, sondern wie in [3] beschrieben berechnet. Außerdem diente ein zweistufiges Wiener-Filter [13] als Referenzverfahren. Für die Auswertung kamen vier verschiedene Qualitätsmaße zum Einsatz, die von Gannot in [3] und [12] vorgeschlagen wurden. Die über 20 Sätze gemittelten Ergebnisse sind in Bild 1 dargestellt.

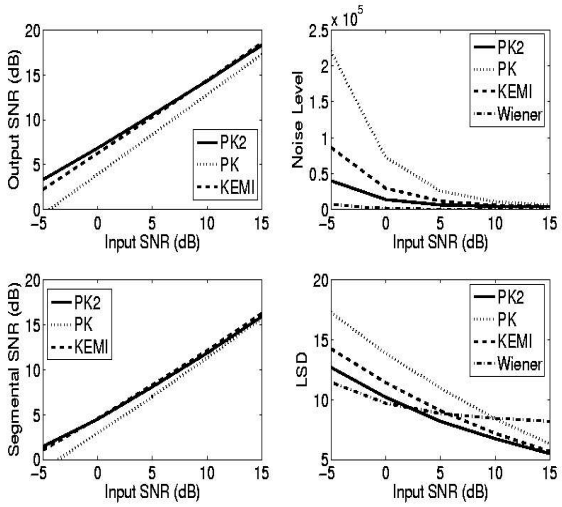

Bild 1: Qualitätsmaße für 20 WSJ-Sätze mit gaußverteiltem Rauschen

Das SNR des Ausgangssignals [12] ist für kleine Eingangs-SNR bei PK2 etwas höher als bei KEMI, während bei großen Eingangs-SNR die Ergebnisse von KEMI etwas besser sind. Für das segmentielle SNR [3] ergeben sich ähnliche Resultate, während der Rauschpegel [12] mit PK2 geringer ist. Da der SNR-Gewinn nicht stark mit der Qualität des Sprachsignals korreliert ist, hat Gannot in [12] vorgeschlagen, die Log-spektrale Distanz (LSD), die besser mit dem Mean-Opinion-Score korreliert ist, als Qualitätsmaß zu verwenden. Mit dem LSD-Maß ergeben sich mit PK2 etwas bessere Resultate als mit KEMI. Der Algorithmus PK, der als Ausgangspunkt für die Entwicklung von PK2 diente, erzielte für die betrachteten Qualitätsmaße und Eingangs-SNR-Werte schlechtere Resultate als PK2 und KEMI. Mit dem Wiener-Filter ergeben sich bei Eingangs-SNR-Werten von mehr als 0dB schlechtere Resultate bzgl. des LSD-Maßes als mit dem Verfahren PK2. Der Rauschpegel wird mit dem Wiener-Filter jedoch deutlich stärker reduziert. Die Qualitätsmaße für das Ausgangs-SNR und das segmentielle SNR wurden für Zeitbereichsverfahren entwickelt und liefern für das Wiener-Filter keine aussagekräftigen Resulate. Die Algorithmen PK2 und KEMI sowie die Erweiterung von PK2 auf farbiges Rauschen mit $q=2$ (siehe Abschnitt 3; im Folgenden als PK2 col

bezeichnet) sowie das Verfahren KEMI mit $q=2$ [3] (im Folgenden als KEMI col bezeichnet) wurden außerdem bei farbigen Rauschen untersucht (Bild 2).

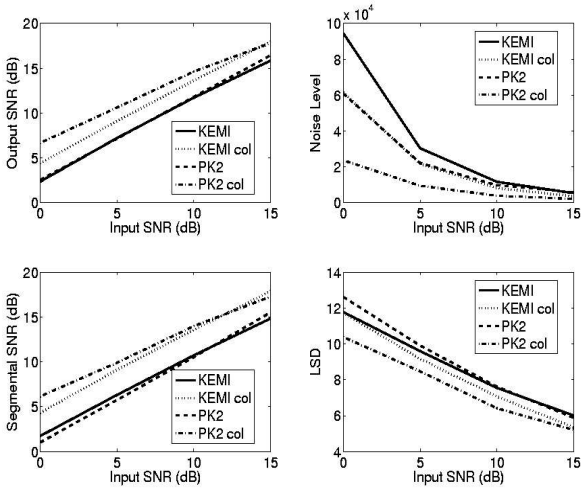

Bild 2: Qualitätsmaße für 20 WSJ-Sätze mit farbigem Rauschen

Durch die Modellerweiterungen ergeben sich bei farbigen Rauschen sowohl mit KEMI als auch mit PK2 deutliche Verbesserungen. Für die Verfahren mit erweitertem Modell sind das Ausgangs-SNR und das segmentielle SNR bei kleinen Eingangs-SNR-Werten für PK2 etwas besser als für KEMI, während sich bei großen Eingangs-SNR-Werten etwas bessere Resultate für KEMI ergeben. Der Rauschlevel und das LSD-Maß sind mit KEMI etwas geringer. Die Laufzeit von PK2 beträgt bei der angegebenen Parametrisierung sowohl für weißes als auch für farbiges Rauschen unter Matlab 7.0.0, Release 14 auf einem Intel Pentium III Prozessor mit 1 GHz Takt-frequenz ungefähr das hundertfache der Laufzeit von KEMI.

5. Schlußbemerkungen

In diesem Paper wurde ein marginalisiertes Partikelfilter vorgestellt, das auf iterativer Blockverarbeitung basiert und ein Zustandsmodell mit LSP-Koeffizienten verwendet, welches aus Trainingsdaten gelernt wird. Die SNR-Gewinne, die bei dem Test mit Sätzen aus der WSJ Sprachdatenbank erzielt wurden, sind vergleichbar mit denen, die sich durch Anwendung des Kalman-EM-iterative Algorithmus ergeben, während etwas bessere Werte des Rauschpegels und der Log-spektralen Distanz erzielt werden. Allerdings ist die Komplexität wesentlich höher. Durch ein erweitertes Zustandsmodell konnten die Resultate bei farbigen Rauschen verbessert werden.

6. Anmerkungen

Die Forschung wird unterstützt durch das DFG Graduiertenkolleg GK-693 des Paderborn Institute for Scientific Computation (PaSCo).

Literatur

[1] Benesty, J., Makino, S., Chen, J.: Speech Enhancement. Springer, 2005.
[2] Paliwal, K.K., Basu, A.: A speech enhancement method based on Kalman filtering. In Proc. ICASSP, Dallas, Texas, 1987.
[3] Gannot, S.: Iterative and sequential Kalman filter-based speech enhancement algorithms. IEEE Trans. Speech and Audio Processing, vol. 6, no. 4, pp. 373-385, Juli 1998.
[4] Gannot, S., Moonen, M.: On the application of the unscented Kalman filter too speech processing. In Proc. Int. Workshop on Acoustic Echo and Noise Control, Kyoto, Japan, 2003.
[5] Vermaak, J., Andrieu, C., Doucet, A. 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, März 2002.
[6] Fong, W., Godsill, S., Doucet, A., West, M.: Monte Carlo smoothing with application to audio signal enhancement. IEEE Trans. Signal Processing, vol. 50, no. 2, pp. 438-449.
[7] Doucet, A., de Freitas, J.F.G., Gordon, N.J. (Eds.): Sequential Monte Carlo Methods in Practice, Springer, 2001.
[8] Ristic, B. Arulampalam, S. Gordon, N.: Beyond the Kalman Filter: Particle Filters for Tracking Applications, Artech House Publishers, 2004.
[9] Raj, B., Singh, R., Stern, R.: On tracking noise with linear dynamical system models. In Proc. ICASSP, Montreal, 2004.
[10] Windmann, S., Haeb-Umbach, R.: Iterative speech enhancement using a non-linear dynamic state model of speech and its parameters. Accepted for ICASSP, 2006.
[11] Deller, J.R., Hansen, J.H.L., Proakis, J.G.: Discrete-Time Processing of Speech Signals, IEEE Press, 2000.
[12] Gannot, S.: Speech Enhancement: Application of the Kalman Filter in the Estimate-Maximize (EM) Framework. In: Speech Enhancement, Benesty, J., Makino, S., Chen, J. (Eds.), Springer, 2005.
[13] ETSI, „ES 202 212 V1.1.1, Advanced front-end feature extraction algorithm“, Techn. rep., Nov. 2003.