MODIFIED KALMAN FILTER EXPLOITING INTERFRAME CORRELATION OF SPEECH AND NOISE MAGNITUDES

Thomas Esch and Peter Vary

Institute of Communication Systems and Data Processing (ind)
RWTH Aachen University, Germany
{esch|vary}@ind.rwth-aachen.de

ABSTRACT

A new method for single channel speech enhancement is presented which relies on a Kalman filter structure. The proposed scheme uses a two step approach. In the first step, temporal correlation of successive speech and noise magnitudes is exploited. Therefore, the current samples are propagated in time based on information taken from previous, enhanced samples. The resulting prediction errors are estimated in a second step by utilizing different statistical estimators. The performance of the proposed method is shown to be considerably better than purely statistical estimators which do not take into account temporal correlation.

***Index Terms*—** Speech enhancement, Kalman filter, linear prediction, noise suppression

1. INTRODUCTION

Speech quality and intelligibility may significantly deteriorate in the presence of background noise, e.g., engine noise or street noise. The problem of enhancing speech that is degraded by additive noise has been widely studied in the past and is still an active field of research. Speech enhancement has many applications in voice communications, speech recognition and hearing aids.

The design of many speech enhancement systems is based on modeling the noisy input coefficients in the short-time Fourier transform (STFT) domain to derive individual adaptive gains for each frequency bin. Most of the rules proposed in literature have been derived under certain assumptions about the statistics of the speech and noise signal. Considering a Gaussian speech and noise model, this enables to deduce *minimum mean-squared error* (MMSE) estimators, such as the well-known Wiener filter [1] or the *short-term spectral amplitude* (STSA) estimator [2]. Martin [3] proposed the use of a Gamma speech model and derived an MMSE estimator for the complex speech coefficients under the assumption of Gaussian and Laplacian noise models. Lotter [4] derived a *maximum a posteriori* (MAP) estimator using a super-Gaussian speech and Gaussian noise model. All of these estimators only utilize the statistical characteristics of speech and noise, correlation in time is explicitly not taken into account (except for smoothing purposes in [2]).

Paliwal and Basu [5] were the first to propose the use of a Kalman filter for the purpose of speech enhancement. In order to reduce complexity, Wu and Chen [6] derived a Kalman filtering system in the sub-band domain. Puder [7] further investigated the application of a Kalman filter in sub-bands and increased the performance compared to the full-band time domain approach. In addition to the exploitation of intra-frame correlation, model-based approaches that consider the correlation of successive speech frames can be found, e.g., in [8] and [9].

In this paper, the Kalman filter approach of [9] is improved. Instead of using a complex predictor to exploit the temporal correlation of successive spectral coefficients, only the real-valued magnitudes are propagated in time. Furthermore, the propagation step is not only applied to the speech signal, but also extended to the noise signal. The resulting prediction errors are estimated in a second step by utilizing different statistical estimators. The remainder of this paper is organized as follows: In Sec. 2, a brief overview about the proposed system is given. Secs. 3 and 4 comprise the procedure of propagation and update step in detail. Experimental results are shown in Sec. 5 and conclusions are drawn in Sec. 6.

2. SYSTEM OVERVIEW

A simplified block diagram of the proposed system is depicted in Fig. 1. It is assumed that the noisy input signal $y(k)$ consists of the clean speech signal $s(k)$ which is degraded by an additive noise signal $n(k)$ according to:

$$y(k) = s(k) + n(k), \tag{1}$$

where k is the discrete time index. To decompose the speech and noise signal, the noisy signal is transformed into the frequency domain. Therefore, $y(k)$ is segmented into overlapping frames of length L_{F}. After windowing and zero-padding, the fast Fourier transform (FFT) is applied to these frames. Hence, the spectral coefficient of the noisy input signal at frequency bin μ and frame λ is given by:

$$\begin{aligned} Y(\lambda,\mu) &= S(\lambda,\mu) + N(\lambda,\mu) && (2)\\ &= R(\lambda,\mu)\mathrm{e}^{j\vartheta(\lambda,\mu)} && (3)\\ &= A(\lambda,\mu)\mathrm{e}^{j\alpha(\lambda,\mu)} + B(\lambda,\mu)\mathrm{e}^{j\beta(\lambda,\mu)}, && (4) \end{aligned}$$

where $S(\lambda,\mu)$ and $N(\lambda,\mu)$ represent the spectral coefficients of speech and noise. Moreover, $R(\lambda,\mu)$, $A(\lambda,\mu)$ and $B(\lambda,\mu)$ denote the magnitudes of the noisy, the speech, and the noise signal and $\vartheta(\lambda,\mu)$, $\alpha(\lambda,\mu)$, $\beta(\lambda,\mu)$ are the corresponding phases.

The investigated system is based on a *Kalman filter* structure that consists of two steps, namely *propagation* and *update step*. In the propagation step, temporal correlation (a priori information of higher order) of successive frames is exploited. In contrast to [9], the Kalman filter is not applied to the complex signal $Y(\lambda,\mu)$ but only to the noisy magnitude $R(\lambda,\mu)$. This is motivated by the fact that most part of the temporal correlation of the spectral coefficients can be found in successive magnitudes and only marginally in the phase samples. In addition, the propagation step is extended to the noise signal in order to additionally take into account correlated noise signals. Hence, the current speech and noise magnitudes are predicted

This work was supported by Nokia, Tampere, Finland.

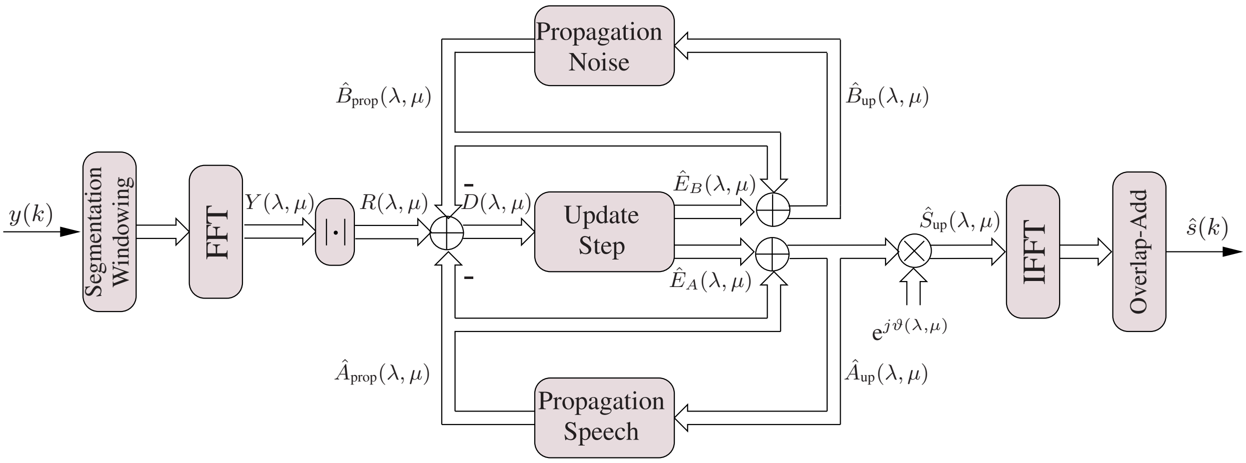

Fig. 1. System block diagram

based on information taken from previous, enhanced samples. Based on the *spectral subtraction method* that is used here (cf. Sec. 3.2), the resulting estimates $\hat{A}_{\text{prop}}(\lambda,\mu)$ and $\hat{B}_{\text{prop}}(\lambda,\mu)$ are combined to get an estimation of the current noisy magnitude:

$$\hat{R}_{\text{prop}}(\lambda,\mu) = \hat{A}_{\text{prop}}(\lambda,\mu) + \hat{B}_{\text{prop}}(\lambda,\mu). \tag{5}$$

In general, the prediction in the propagation step is erroneous and the prediction errors

$$\hat{E}_{\text{A}}(\lambda,\mu) = A(\lambda,\mu) - \hat{A}_{\text{prop}}(\lambda,\mu) \quad \text{and} \tag{6}$$

$$\hat{E}_{\text{B}}(\lambda,\mu) = B(\lambda,\mu) - \hat{B}_{\text{prop}}(\lambda,\mu) \tag{7}$$

occur for the speech and noise magnitudes. Considering the differential signal

$$D(\lambda,\mu) = R(\lambda,\mu) - \hat{R}_{\text{prop}}(\lambda,\mu), \tag{8}$$

the update step estimates these prediction errors based on a conventional statistical estimator, utilizing a priori information of zeroth order. This estimator is adapted to the statistics of speech and noise and performs a spectral weighting of the differential signal by multiplying the Kalman gain $K(\lambda,\mu)$:

$$\hat{E}_{\text{A}}(\lambda,\mu) = K(\lambda,\mu)D(\lambda,\mu) \tag{9}$$

$$\hat{E}_{\text{B}}(\lambda,\mu) = \left(1 - K(\lambda,\mu)\right)D(\lambda,\mu). \tag{10}$$

To obtain the enhanced speech and noise magnitudes $\hat{A}_{\text{up}}(\lambda,\mu)$ and $\hat{B}_{\text{up}}(\lambda,\mu)$, the initial predictions of the propagation step are updated:

$$\hat{A}_{\text{up}}(\lambda,\mu) = \hat{A}_{\text{prop}}(\lambda,\mu) + \hat{E}_{\text{A}}(\lambda,\mu) \tag{11}$$

$$\hat{B}_{\text{up}}(\lambda,\mu) = \hat{B}_{\text{prop}}(\lambda,\mu) + \hat{E}_{\text{B}}(\lambda,\mu). \tag{12}$$

The estimated clean speech magnitude $\hat{A}_{\text{up}}(\lambda,\mu)$ is recombined with the noisy input phase:

$$\hat{S}_{\text{up}}(\lambda,\mu) = \hat{A}_{\text{up}}(\lambda,\mu)\mathrm{e}^{j\vartheta(\lambda,\mu)}, \tag{13}$$

before an inverse fast Fourier transform (IFFT) and the overlap-add method are applied.

3. PROPAGATION STEP

In this section, further details about the propagation step are given. The magnitudes $A(\lambda,\mu)$ and $B(\lambda,\mu)$ of the speech and the noise signal are both modeled as two independent autoregressive (AR) processes. While in [9], the speech phase was implicitly estimated within the complex linear prediction, an additional phase estimation term is not necessary here.

3.1. Magnitude Estimation

Within the modified Kalman filter, the AR model is used to exploit temporal correlation of the speech and noise magnitudes. In [9], a complex AR model was used to directly predict the spectral coefficient $\hat{S}_{\text{prop}}(\lambda,\mu)$. As there is almost no correlation in successive phase samples, linear prediction is explicitly applied to the speech and noise magnitudes in order to exploit the maximum temporal correlation within adjacent magnitudes.

The magnitude estimates $\hat{A}_{\text{prop}}(\lambda,\mu)$ and $\hat{B}_{\text{prop}}(\lambda,\mu)$ for speech and noise can be stated as:

$$\hat{A}_{\text{prop}}(\lambda,\mu) = \sum_{i=1}^{N_{\text{K}}} \hat{a}_i(\lambda,\mu)\hat{A}_{\text{up}}(\lambda-i,\mu) \quad \text{and} \tag{14}$$

$$\hat{B}_{\text{prop}}(\lambda,\mu) = \sum_{i=1}^{M_{\text{K}}} \hat{b}_i(\lambda,\mu)\hat{B}_{\text{up}}(\lambda-i,\mu), \tag{15}$$

where N_{K} and M_{K} represent the orders of the speech and the noise model respectively. The AR coefficients $\hat{a}_i(\lambda,\mu)$ and $\hat{b}_i(\lambda,\mu)$ are estimated in advance by minimizing the energies of the prediction errors. Therefore, the well-known Levinson-Durbin algorithm is used [10]. The required autocorrelation vector and matrix are calculated from the previous L_{AC} enhanced magnitudes of either speech or noise.

3.2. Phase Estimation

For the calculation of the differential signal $D(\lambda,\mu)$ in Eq. 8, an estimate of the noisy magnitude $R(\lambda,\mu)$ is needed. In general, $R(\lambda,\mu)$ is given by:

$$R(\lambda,\mu) = \sqrt{A^2(\lambda,\mu) + B^2(\lambda,\mu) \ldots + 2A(\lambda,\mu)B(\lambda,\mu)\cos\left(\alpha(\lambda,\mu) - \beta(\lambda,\mu)\right)}. \tag{16}$$

As the magnitudes $A(\lambda,\mu)$ and $B(\lambda,\mu)$ have already been predicted, estimates of the phases $\alpha(\lambda,\mu)$ and $\beta(\lambda,\mu)$ are still required in order to apply Eq. 16. If the predictions $\hat{A}_{\text{prop}}$ and $\hat{B}_{\text{prop}}$ are of sufficient accuracy, the range of possible values for the estimation of $\alpha(\lambda,\mu)$ is limited if $|S| > |N|$. This can be seen from the example in Fig. 2. The limit of the maximum phase deviation $\vartheta(\lambda,\mu) - \alpha(\lambda,\mu)$ is defined by the case when $Y(\lambda,\mu)$ is perpendicular to $N(\lambda,\mu)$, i.e.,

$$-\text{arsin}\frac{B(\lambda,\mu)}{A(\lambda,\mu)} \leq \vartheta(\lambda,\mu) - \alpha(\lambda,\mu) \leq \text{arsin}\frac{B(\lambda,\mu)}{A(\lambda,\mu)}. \tag{17}$$

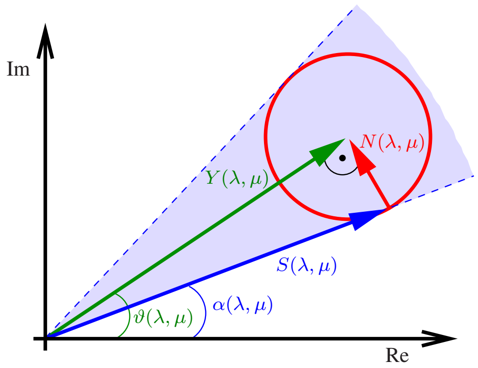

Fig. 2. Limitation of $\alpha(\lambda,\mu)$ for the case $|S| > |N|$

It is known that if this phase deviation is below a certain threshold (between $\pi/4$ and $\pi/8$), nothing is recognized due to psychoacoustical properties of the human ear. In [11], it was shown that no speech degradation is perceived as long as the signal-to-noise ratio $\frac{|S|^2}{|N|^2}$ is at least 6 dB. Therefore, the noisy input phase $\vartheta(\lambda,\mu)$ is utilized as estimate for $\alpha(\lambda,\mu)$. For the case $|N| > |S|$, a similar expression as in Eq. 17 can be derived for the maximum phase deviation $\vartheta(\lambda,\mu) - \beta(\lambda,\mu)$. Thus, the noisy phase $\vartheta(\lambda,\mu)$ is also applied as estimate for $\beta(\lambda,\mu)$. Using these phase estimates, the maximum phase deviation of $\alpha(\lambda,\mu)$ becomes smaller with an increasing SNR and that of $\beta(\lambda,\mu)$ becomes smaller with a decreasing SNR. Eq. 16 reduces to the equation of the well-known spectral subtraction method [12]. Here, the noisy magnitude is given as the sum of the STFT of speech and noise magnitude such that $R(\lambda,\mu)$ can be estimated according to Eq. 5. Hence, the noisy magnitude can directly be estimated from the estimates $\hat{A}_{\text{prop}}(\lambda,\mu)$ and $\hat{B}_{\text{prop}}(\lambda,\mu)$ and no additional phase estimation term is necessary.

4. UPDATE STEP

While in the propagation step, the temporal correlation of successive speech and noise magnitudes is exploited, the update step makes use of the statistical characteristics of both signals. The objective in this step is to estimate the prediction errors $E_{\text{A}}(\lambda,\mu)$ and $E_{\text{B}}(\lambda,\mu)$, caused in the propagation step. By inserting Eqs. 5, 11 and 12 in Eq. 8 and using the spectral subtraction method also for $R(\lambda,\mu)$, it can be shown that the differential signal $D(\lambda,\mu)$ consists of the speech prediction error $E_{\text{A}}(\lambda,\mu)$ that is degraded by the noise prediction error $E_{\text{B}}(\lambda,\mu)$:

$$\begin{aligned} D(\lambda,\mu) &= R(\lambda,\mu) - \hat{R}_{\text{prop}}(\lambda,\mu) \\ &= A(\lambda,\mu) - \hat{A}_{\text{prop}}(\lambda,\mu) + B(\lambda,\mu) - \hat{B}_{\text{prop}}(\lambda,\mu) \\ &= E_{\text{A}}(\lambda,\mu) + E_{\text{B}}(\lambda,\mu). \end{aligned} \tag{18}$$

The estimation problem in the update step reduces to a 'classical' noise reduction problem: The target coefficient $E_{\text{A}}(\lambda,\mu)$ is assumed to be degraded by the additive 'noise' coefficient $E_{\text{B}}(\lambda,\mu)$ to produce the noisy coefficient $D(\lambda,\mu)$. Thus, a conventional statistical estimator can be applied which is adapted to the statistics of the prediction errors.

Based on the assumption that the coefficients $E_{\text{A}}(\lambda,\mu)$ and $E_{\text{B}}(\lambda,\mu)$ are statistically independent, two estimators are considered in the following for the update step, namely an MMSE estimator [1] and a recently published super-Gaussian joint MAP estimator [4]. Both estimators rely on a Gaussian model for the noise signal. Indeed, even if the initial speech signal $s(k)$ is degraded by a colored noise $n(k)$, the propagation step has the effect of a prewhitening filter as it reduces possible temporal correlation. In addition, both estimators require the *a posteriori* SNR $\gamma(\lambda,\mu)$ and the *a priori* SNR $\xi(\lambda,\mu)$ which are defined as follows:

$$\gamma(\lambda,\mu) = \frac{|D(\lambda,\mu)|^2}{\mathcal{E}\{|E_{\text{B}}(\lambda,\mu)|^2\}} \text{ and } \xi(\lambda,\mu) = \frac{\mathcal{E}\{|E_{\text{A}}(\lambda,\mu)|^2\}}{\mathcal{E}\{|E_{\text{B}}(\lambda,\mu)|^2\}}, \tag{19}$$

where $\mathcal{E}\{\cdot\}$ represents the expectation operator. The two estimators are briefly described in the following:

1. *Gaussian MMSE Estimator (Wiener Filter)*
   This Gaussian MMSE estimator corresponds to the well-known Wiener filter solution and is derived from the optimal filter theory [1]. This linear estimator minimizes the mean square error between clean and enhanced coefficient. Applied to the update step, the enhanced coefficient $\hat{E}_{\text{A}}(\lambda,\mu)$ can be stated as:
   $$\hat{E}_{\text{A}}(\lambda,\mu) = \underbrace{\frac{\xi(\lambda,\mu)}{\xi(\lambda,\mu)+1}}_{K_{\text{G}}(\lambda,\mu)} D(\lambda,\mu), \tag{20}$$
   Note that this MMSE estimator in the update step equals the conventional Kalman filter gain as it arises from the same assumption that the prediction errors for speech and noise are Gaussian distributed [6].
2. *Super-Gaussian Joint MAP Estimator*
   Applied to the update step, this generalized super-Gaussian estimator [4] uses the following parametric function to approximate the *probability density function* (PDF) of the amplitude $|E_{\text{A}}|$:
   $$p(|E_{\text{A}}|) = \frac{\delta^{\eta+1}}{\Gamma(\eta+1)} \frac{|E_{\text{A}}|^\eta}{\sigma_{E_{\text{A}}}^{\eta+1}} \exp\left\{-\delta\frac{|E_{\text{A}}|}{\sigma_{E_{\text{A}}}}\right\}, \tag{21}$$
   where $\Gamma(\cdot)$ states the Gamma function and $\sigma_{E_{\text{A}}}$ the standard deviation of the speech prediction error. The parameters δ and η can be selected in order to obtain the optimal approximation. Therefore, the Kullback-Leibler distance between measured and modeled PDF is used [4]. The resulting weighting rule of this MAP estimator is given by:
   $$\hat{E}_{\text{A}}(\lambda,\mu) = \underbrace{\left(u(\lambda,\mu) + \sqrt{u^2(\lambda,\mu) + \frac{\eta}{2\gamma(\lambda,\mu)}}\right)}_{K_{\text{S}}(\lambda,\mu)} D(\lambda,\mu),$$
   $$\text{where } u(\lambda,\mu) = \frac{1}{2} - \frac{\delta}{4\sqrt{\gamma(\lambda,\mu)\xi(\lambda,\mu)}}. \tag{22}$$

Based on the calculation of either K_{G} or K_{S}, the noise prediction error can be estimated according to Eq. 10.

5. RESULTS

For the evaluation of the proposed noise reduction scheme, five speech signals from the NTT speech database were degraded by six different noise types (f16, buccaneer, car, factory1, factory2, white), taken from the NOISEX-92 database. Among the speech signals, there were three male and two female speech sequences, each with a length of 8 seconds. The input SNR was varied between -10 dB and 35 dB (step size: 5 dB). For the analysis and synthesis structure, 75% overlapping Hann windows with a length of 20 ms and a 256-FFT (including zero-padding) were used. It turned out that good results were achieved by the following parameters applied to the modified Kalman filter: $L_{\text{AC}} = 6$, $N_{\text{K}} = 3$ and $M_{\text{K}} = 2$ (sampling frequency f_s=8 kHz). While the power of the noise prediction error $\mathcal{E}\{|E_{\text{B}}(\lambda,\mu)|^2\}$ was estimated by using [13], the *decision-directed* approach [2] was utilized for the estimation of the a priori SNR.

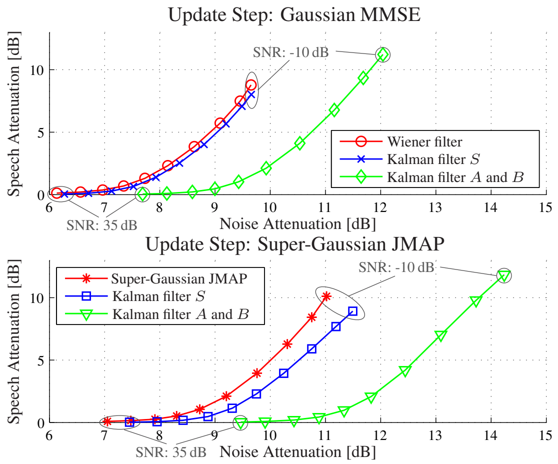

Fig. 3. Speech attenuation vs. noise attenuation

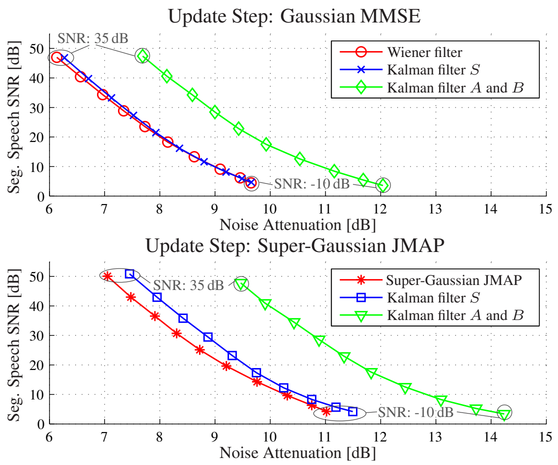

Fig. 4. Segmental speech SNR vs. noise attenuation

A total of six different noise suppression techniques were investigated. Among them were the purely statistical weighting rules Wiener filter [1] and super-Gaussian joint MAP (JMAP) estimator [4]. They were compared with the modified Kalman filter in [9] (Kalman filter S) and the new approach (Kalman filter A and B) that is proposed in this paper. For each Kalman filter, the above mentioned weighting rules (cf. Sec. 4) were applied in the update step respectively. For the evaluation, three different kinds of instrumental measurements were used, namely the segmental noise attenuation (NA), the segmental speech attenuation (SA) and the segmental speech SNR (SegSNR) (e.g., [14]).

Figs. 3 and 4 illustrate the averaged results for SA and SegSNR, respectively, both plotted over NA with the input SNR as control variable. This procedure makes a fair comparison between noise attenuation and speech distortion possible. In Fig. 3, a low SA and a high NA is desirable, in Fig. 4 a high SegSNR and a high NA. In the upper plots of Figs. 3 and 4, the Gaussian MMSE estimator was used in the update step of the Kalman filters, in the lower plots the super-Gaussian JMAP estimator respectively.

The results show that both types of Kalman filters achieve better results than the corresponding purely statistical estimator. Even though the benefits in noise suppression are at the expense of an increase in speech attenuation/distortion at very low input SNR values, the proposed Kalman filter outperforms the approach in [9]. The results show a considerable enhancement by the new estimator. Furthermore, it can be seen that the utilization of the super-Gaussian JMAP estimator, i.e., the adaptation to the PDF of the prediction error signal, leads to better results than the application of the Gaussian MMSE estimator. The instrumental measurements were confirmed by informal listening tests.

6. CONCLUSIONS

This paper presents a new method for single channel speech enhancement that relies on a Kalman filter structure. In the first step, this model-based approach exploits the temporal correlation of successive speech and noise magnitudes using two independent AR processes. In the second step, the statistics of the differential signal are utilized to estimate the prediction errors by applying two different statistical estimators. Although the complexity is moderately increased by the proposed technique, the instrumental measurements in terms of segmental speech SNR, speech and noise attenuation clearly show the better performance compared to the Wiener filter, the super-Gaussian JMAP estimator and another recently published Kalman filter approach.

7. REFERENCES

[1] J. S. Lim and A. V. Oppenheim, “Enhancement and Bandwidth Compression of Noisy Speech,” *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[2] Y. Ephraim and D. Malah, “Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator,” *IEEE Transaction on Acoustic, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[3] R. Martin, “Speech Enhancement Using MMSE Short Time Spectral Estimation with Gamma Distributed Speech Priors,” in *Proc. of ICASSP*, Orlando, USA, 2002.

[4] T. Lotter and P. Vary, “Speech Enhancement by MAP Spectral Amplitude Estimation using a Super-Gaussian Speech Model,” *EURASIP Journal on Applied Signal Processing*, pp. 1110–1126, 2005.

[5] K. Paliwal and A. Basu, “A Speech Enhancement Method Based on Kalman Filtering,” in *Proc. of ICASSP*, Dallas, USA, 1987.

[6] W.-R. Wu and P.-C. Chen, “Subband Kalman Filtering for Speech Enhancement,” *IEEE Transactions on Circuits and Systems II*, vol. 45, no. 8, pp. 1072–1083, Aug. 1998.

[7] H. Puder, “Kalman-Filters in Subbands for Noise Reduction with Enhanced Pitch-Adaptive Speech Model Estimation,” *European Transactions on Telecommunications*, vol. 13, no. 2, pp. 139–148, 2002.

[8] E. Zavarehei and S. Vaseghi, “Speech Enhancement in Temporal DFT Trajectories Using Kalman Filters,” in *Proc. of INTERSPEECH*, Lisbon, Portugal, 2005.

[9] T. Esch and P. Vary, “Speech Enhancement Using a Modified Kalman Filter Based on Complex Linear Prediction and Supergaussian Priors,” in *Proc. of ICASSP*, Las Vegas, USA, 2008.

[10] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*. Upper Saddle River, New Jersey: Prentice Hall, 1996.

[11] P. Vary, “Noise suppression by spectral magnitude estimation-mechanism and theoretical limits,” *Signal Processing*, vol. 8, no. 4, pp. 387–400, 1985.

[12] S. F. Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction,” *IEEE Transactions on Speech and Audio Processing*, vol. 27, pp. 113–120, 1979.

[13] R. Martin, “Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 501–512, 2001.

[14] S. Gustafsson, R. Martin, P. Jax, and P. Vary, “A Psychoacoustic Approach to Combined Acoustic Echo Cancellation and Noise Reduction,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 245–256, July 2002.